entropy

MDPI

Article

Improved Treatment of the Independent Variables for the Deployment of Model Selection Criteria in the Analysis of Complex Systems

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

[1] Department of Industrial Engineering, University of Rome "Tor Vergata", Via Del Politecnico 1, 00133 Roma, Italy; michela.gelfusa@uniroma2.it (M.G.); r.rossi@ing.uniroma2.it (R.R.)
[2] Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy; andrea.murari@istp.cnr.it
* Correspondence: luca.spolladore@uniroma2.it

Abstract: Model selection criteria are widely used to identify the model that best represents the data among a set of potential candidates. Amidst the different model selection criteria, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) are the most popular and better understood. In the derivation of these indicators, it was assumed that the model's dependent variables have already been properly identified and that the entries are not affected by significant uncertainties. These are issues that can become quite serious when investigating complex systems, especially when variables are highly correlated and the measurement uncertainties associated with them are not negligible. More sophisticated versions of this criteria, capable of better detecting spurious relations between variables when non-negligible noise is present, are proposed in this paper. Their derivation is obtained starting from a Bayesian statistics framework and adding an a priori Chi-squared probability distribution function of the model, dependent on a specifically defined information theoretic quantity that takes into account the redundancy between the dependent variables. The performances of the proposed versions of these criteria are assessed through a series of systematic simulations, using synthetic data for various classes of functions and noise levels. The results show that the upgraded formulation of the criteria clearly outperforms the traditional ones in most of the cases reported.

Keywords: model selection criteria; Akaike information criterion; Bayesian information criterion; overfitting; redundancy; variable selection; complexity; information theory; relevance

check for updates

Citation: Spolladore, L.; Gelfusa, M.; Rossi, R.; Murari, A. Improved Treatment of the Independent Variables for the Deployment of Model Selection Criteria in the Analysis of Complex Systems. *Entropy* **2021**, *23*, 1202. https://doi.org/10.3390/e23091202

Academic Editor: Song-Nam Hong

Received: 9 August 2021
Accepted: 8 September 2021
Published: 11 September 2021

1. Introduction to Model Selection Criteria Based on Bayesian Statistics and Information Theory

Model Selection (MS) can be defined as the task of identifying the model best supported by the data, among a set of potential candidates [1]. In many fields, model selection is an essential part of scientific enquiry [2]. It can also be argued that this step is often among the most delicate in statistical inference.

The exact definition of what is meant by the best model is controversial and is probably application dependent [3]. Indeed, the requirements of models are not the same if the goal of the study is prediction, explanation, or control. In any case, basically all approaches to model selection try to find a compromise between goodness of fit and complexity. At the same level of goodness of fit, simpler models, implementing some form of Occam's razor, are preferred. The goodness of fit is assessed with the likelihood or, when this is not possible, with some metric quantifying the residuals, the distance between the model predictions, and the data. The complexity of the models is identified with the number of the model parameters. In the following, attention will be focussed on MSC derived with the help of Bayesian statistics and information theory, since these are the ones explicitly designed to find a trade-off between goodness of fit and complexity. In any case, similar considerations apply also to frequentist types of techniques. A remark about the nomenclature is in place at this point. Since the application covered in the present work is regression, with the term database in the following, it is indicated as a finite ordered list of entries. Each entry consists of a sequence of elements formed by a dependent variable, y, and a series of p regressors or predictors, x_i.

The most widely accepted and best understood model selection criteria, based on information theory and Bayesian statistics, are the Akaike Information Criterion (AIC) [4] and the Bayesian Information Criterion (BIC) [5].

The theoretical derivations of these metrics result in the following unbiased forms of the criteria:

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

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

where L is the likelihood of the model given the data, k the number of parameters in the model, and n the number of entries in the database (also called the sample size). Both AIC and BIC metrics are basically cost functions, which have to be minimized; they favour models with a high likelihood but implement a penalty for complexity (the term proportional to k).

Since in most applications, such as the ones discussed in this work, it is impossible to calculate the likelihood of the models, the metric adopted for the goodness of fit is the Euclidean distance of the residuals. Under the traditional assumption, that the data are identically distributed and independently sampled from a normal distribution, it can be demonstrated that the AIC can be written (up to an additive constant, which depends only on the number of entries in the database and not on the model) as follows:

$$AIC = n \cdot \ln(MSE) + 2k \tag{3}$$

where MSE is the mean-squared error of the residuals, n the number of entries in the database, and k the number of parameters in the model. Similar assumptions allow expressing the BIC criterion as follows:

$$BIC = n \cdot \ln\left({\sigma_{(\epsilon)}}^2\right) + k \cdot \ln(n) \tag{4}$$

where ${\sigma_{(\epsilon)}}^2$ is the variance of the residuals, n is again the number of entries in the database, and k the number of parameters in the model. The derivation of these two criteria in the various approximation is fully covered in [6].

These two indicators, and all the others belonging to the same families, are cost functions to be minimised, in the sense that the better the model the lower their value. This can be intuitively appreciated by a simple inspection of their structure. The first term favours models that are closer to the data. The second addend is the penalty term for complexity.

In the last years, various upgrades of these criteria have been proposed. They are mainly meant at improving the goodness of fit, by utilising more sophisticated statistics than the simple MSE, and at devising more accurate estimates of the penalisation for complexity [7,8]. All these improvements have proved to be quite significant, but they do not consider explicitly the problems related to the choice of the regressors and the effects of the measurement uncertainties. They basically assume that the independent variables have already been properly identified without any specific provision for this aspect. Some of them deploy quite sophisticated statistical indicators of the distribution of the residuals, but they all take the measurements as given without any error bar. These are all issues which can be quite relevant when investigating complex systems. Typically, in the field of complexity, various quantities can be spuriously correlated with the dependent one, and measurements can be affected by significant uncertainties due to the poor accessibility of many systems. In this situation, as will be shown in the following, the performance of the traditional versions of the AIC and BIC are unsatisfactory, both being prone to include redundant variables in the selected models.

This work aims to provide an upgraded version of the traditional AIC and BIC criteria to alleviate the problems posed by quantities spuriously correlated with the actual predictors. These quantities tend to mislead the available versions of the indicators, inducing them to converge on models with an excessive number of non-relevant regressors. The situation is significantly worsened by the presence of significant levels of noise, which tend to blur the relations between the dependent quantities and the predictors, as shown in Section 4, which is devoted to the numerical tests. It should be mentioned that the vast majority, if not all, of the applications of model selection criteria involve experimental measurements, which are always affected by some form of noise. The capability of the proposed improvements of dealing with uncertainties is therefore an important aspect that needs to be assessed.

The paper is organized as follows. In the next section, the main information theoretic indicators used in the rest of the paper are reviewed. In Section 3, the derivation of the upgraded version of the AIC and BIC is covered. In Section 4, the performances of the upgraded criteria are evaluated through a series of systematic tests. In Section 5, an application of the derived criteria to a real-life database is reported. The conclusions of the paper are presented in the final section.

2. Brief Review of the Information Theoretic Indicators Relevant to the Upgrades of the Model Selection Criteria

The first information theoretic quantity [9], required to understand the improvements of the MSC proposed in this work, is the Mutual Information (MI) between two random variables, X and Y [9]:

$$MI\ (X_i, Y) = -\sum_X \sum_Y P_{XY} \ln\left(\frac{P_{XY}}{P_X P_Y}\right) \tag{5}$$

where P_{XY} is the joint probability distribution function (pdf) of the random variables X and Y. Being fully nonlinear, contrary to the Pearson correlation coefficient, the MI is well suited to extract, from a given database, the best features, i.e., the best regressors, X_i, to reproduce the desired dependent variable Y.

The second important information theoretic indicator, used in the rest of the paper, is the concept of redundancy, RD, between a variable X_i and a set, S, of other variables, X_j:

$$RD(X_i, S) = \sum_{X_j \in S} MI\left(X_i, X_j\right) \tag{6}$$

Mutual information and redundancy allow defining a quantity, called relevance RL, which quantifies the net contribution of a variable to reducing the uncertainty in a different one, Y, above what is already contributed by another set of quantities. Relevance is defined as

$$RL(X_i, Y) = MI(X_i, Y) - RD(X_i, S_{PS}) = MI(X_i, Y) - \sum_{X_j \in S_{PS}} MI\left(X_i, X_j\right) \tag{7}$$

3. Derivation of the Upgraded Version of the BIC and AIC

In this section, the original versions of the BIC and AIC criteria are reviewed, and this provides an introduction to the derivation of the upgraded versions of the criteria. The BIC criterion is discussed first because it allows a more natural introduction of the proposed improvements.

3.1. Upgraded Version of the BIC

The Bayesian approach to model selection is based on the maximization of the posterior probability of a model M_i given the data $Y = y_1, \ldots, y_n$. From the Bayes theorem, this posterior probability can be written as follows:

$$p(M_i|Y) = \frac{p(Y|M_i) \cdot p(M_i)}{p(Y)} \tag{8}$$

where $p(Y|M_i)$ is the marginal likelihood of the Model M_i and can be evaluated as follows:

$$p(Y|M_i) = \int L(Y|M_i, \boldsymbol{\theta_i}) \cdot f(\boldsymbol{\theta_i}|M_i) d\boldsymbol{\theta_i} \tag{9}$$

where $\boldsymbol{\theta_i}$ is the vector of the parameters of the model M_i and $f(\boldsymbol{\theta_i}|M_i)$ is the probability distribution of the parameters.

It can be demonstrated that for high n, and setting $f(\boldsymbol{\theta_i}|M_i) = 1$ (uninformative prior), Equation (9) can be approximated with

$$p(Y|M_i) \approx L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big) \cdot e^{-\frac{|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}|}{2}\log n} \tag{10}$$

With $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} = argmax_{\boldsymbol{\theta_i} \in \Theta}(L(Y|M_i, \boldsymbol{\theta_i}))$.

Substituting (10) in (8), we obtain the following:

$$p(M_i|Y) \propto L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big) \cdot e^{-\frac{|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}|}{2}\log n} \cdot p(M_i) \tag{11}$$

If we set $p(M_i) = 1$, which implies considering all the models equally probable, (11) leads to the traditional definition of the BIC. Indeed, after taking the logarithm and simple mathematical manipulations, (11) becomes the following:

$$2 \cdot \log(p(M_i|Y)) \approx 2 \cdot \log\big(L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big)\big) - \big|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big| \cdot \log n \tag{12}$$

The right-hand side of Equation (12) can be recognized as the BIC criterion estimate for the model M_i with an inverted sign. Indeed maximizing (12) is equivalent to minimizing:

$$BIC \approx -2 \cdot \log\big(L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big)\big) + \big|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big| \cdot \log n \tag{13}$$

In situations for which the relevant assumptions are valid, the likelihood can be replaced with the standard deviation of the residuals, with $\big|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big|$ as the number k of parameters in the model, allowing to recover Equation (4).

As shown in the following sections, when the redundancy between the regressors is not negligible, the traditional BIC criterion can fail to identify the right model, showing a tendency to include redundant variables in the selected solutions. To address this problem, a modified version of the BIC criterion is proposed, which, instead of assuming that the models have all the same probability, includes a penalty term for models with high redundancy in the predictor variables.

The proposed a priori probability distribution of the models depends on an overall quantity that we will indicate as $WMRR$ (Weighted Mutual Regressor Relevance). Given a set of regressors $X_1, X_2, \ldots X_N$ and a dependent variable Y, MRP_{XY} is defined as

$$WMRR = n \cdot \sum_{i=1}^{N}\sum_{j=1}^{N} MI\big(X_i, X_j\big) \cdot (1 - RL_N(X_i, Y)), \quad for\ i \neq j; \tag{14}$$

where $MI\big(X_i, X_j\big)$ is the mutual information estimate between the i-th and j-th predictor variables and $RL_N(X_i, Y) = \frac{RL_N(X_i,Y)}{\max\limits_j(RL_N(X_i,Y))}$ is the relevance between the i-th predictor and the predicted variable normalized to the maximum value.

This quantity is higher for models which make use of predictors highly correlated between them and that at the same time have low relevance to the dependent variable.

Note that since $MI(X_i, X_j) \geq 0 \ \forall \ X_i, X_j$, $WMRR$ is also positively define.

The proposed a priori models' probability density function is a Chi-squared distribution function that can then be written as

$$p(M_i) = \frac{MI_{XY}^{\frac{k}{2}-1} \cdot e^{-\frac{WMRR}{2}}}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)}, \ with \ k = 2 \tag{15}$$

where $\Gamma\left(\frac{k}{2}\right) = \left(1 - \frac{k}{2}\right)!$. In this way, Models with $WMRR = 0$ have the highest probability of being chosen, while models with greater $WMRR$ are penalised.

Plugging (15) into (11) one obtains the following:

$$p(Y|M_i) \approx L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big) \cdot e^{-\frac{|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}|}{2}\log n} \ . \ \frac{e^{-\frac{WMRR}{2}}}{2} \tag{16}$$

Which can be rewritten as

$$2 \cdot \log(p(Y|M_i)) \approx 2 \cdot \log\big(L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big)\big) - \big|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big| \cdot \log n - WMRR \tag{17}$$

Maximizing (17) is equivalent to minimizing

$$MIBIC = -2 \cdot \log\big(L\big(Y\big|M_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big)\big) + \big|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big| \cdot \log n + WMRR \tag{18}$$

If, as it is often the case, the likelihood is difficult or impossible to calculate, and the variables are identically distributed and independently sampled from a normal distribution, the MIBIC can be written in the practical form:

$$MIBIC = n \cdot \ln\left({\sigma_{(\epsilon)}}^2\right) + k \cdot \log n + WMRR \tag{19}$$

where as usual k indicates the number of the model's parameters.

The choice of the prior, which is a delicate point in any Bayesian statistical treatment, deserves a comment. Since $WMRR$ is positive definite, its probability distribution function should also be supported on semi-infinite intervals [0, ∞). Moreover, since the main idea behind the proposed improvement of the criterion hinges on penalizing models with strongly correlated variables, this pdf should reach its maximum value when $WMRR = 0$ and decrease as $WMRR$ increases. There are several pdf that satisfy these conditions, but the Chi-squared distribution with k = 2 is the most uninformative in the exponential family. Indeed, its implementation implies the simplest $WMRR$ linear correction term in the upgraded BIC.

3.2. Upgraded Version of the AIC

The derivation of the AIC criteria is based on the concept of minimizing the Kullback–Leibler divergence between the model generating the data and the fitted candidate model. Given the different derivation approach compared to the BIC, the formal addition of an a priori probability distribution function of the models is not possible. Nevertheless, since the AIC is also based on the assumption that the independent variables have already been properly identified and that the effects of the measurement uncertainties are negligible, it is reasonable to include a correction term also in the AIC, which can help in the model selection process when these assumptions are not met. As a consequence, in analogy with the already described MIBIC, the following upgraded version of the AIC, called *MIAIC*, is proposed:

$$MIAIC = n \cdot \ln\left({\sigma_{(\epsilon)}}^2\right) + 2\big|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\big| + WMRR \tag{20}$$

It is worth mentioning that the same argument, leading to the same upgrade, is equally valid for the other indicators belonging to the AIC family, such as the c-AIC and the QAIC. Indeed, for the types of applications that are the subject of this work, these indicators can be expressed as the original AIC plus an additive term [6]. Consequently, perfectly analogue versions including the $WMRR$ can be easily calculated and have proved to be at least equally effective.

4. Results of Systematic Tests with Synthetic Data

To evaluate the performance of the upgraded versions of the indicators developed in this work, a series of systematic tests have been performed. The main families of functions have been investigated: power laws, polynomials, exponentials, and combinations thereof.

Given the importance of the functional dependence and of the fact that the experimental case studied in the following belongs to this family, power laws are discussed first, which illustrates the methodology of the test in detail.

A synthetic dependent variable is generated from a set of predictor variables in the power-law form reported below:

$$Y = \alpha_0 \cdot X_1^{\alpha_1} \cdot X_2^{\alpha_2} \ldots \cdot X_N^{\alpha_N} \tag{21}$$

The predicted variable Y is generated with Equation (21) using 3 uncorrelated random predictor variables, $X_1,\ X_2,\ X_3$, from the $N(\mu = 10, \sigma_N = 1)$ distribution. The coefficients in (21) are all set equal to 1, and the number of data points generated is $n = 5000$.

A fourth correlated predictor variable is added to the set of possible regressors of y in the form reported below:

$$X_4 = X_1 + N((\mu = 0, \sigma_N = 0.3 \cdot std(X_1))) \tag{22}$$

Then, a normally distributed noise $N\left(\mu = 0, \sigma_N = \frac{noise\%}{100} \cdot std(X_1)\right);\ for\ i = 1, \ldots, 4$ is added to all predictors. The parameter $noise\%$ is the percentage of noise with respect to the standard deviation of the regressor and is varied between 1% and 30%. A noise of the same type $N(\mu = 0, \sigma_N = 0.1 \cdot std(Y))$ is added to the independent variable Y.

After generating the variables and adding the noise, two models of the predicted variable fitting (20) to the noised values of Y have been obtained: The first using all the four noised predictors available, $X_1,\ X_2,\ X_3,\ X_4$, and the second using only the noised predictors $X_1,\ X_2,\ X_3,$ used to build $\boldsymbol{y}$.

The two obtained models are compared using both the standard and the modified version of the AIC and BIC.

The results of the comparison varying the parameter $noise\%$ are reported in Figure 1. Each result reported in these plots is an average of over 5 repetitions of the calculations.

As can be noted from inspection of Figure 1, apart from the cases with very low noise, the model obtained, including the redundant variable, would always be chosen over the right model by the traditional AIC/BIC. Instead, the modified versions always succeed in identifying the right model.

The analysis has then been performed for other two types of correlation functions for the redundant variables:

$$X_4 = X_1 + X_2 + N\ (\mu = 0, \sigma = \sigma_N = 0.3 \cdot std(X_1 + X_2)) \tag{23}$$

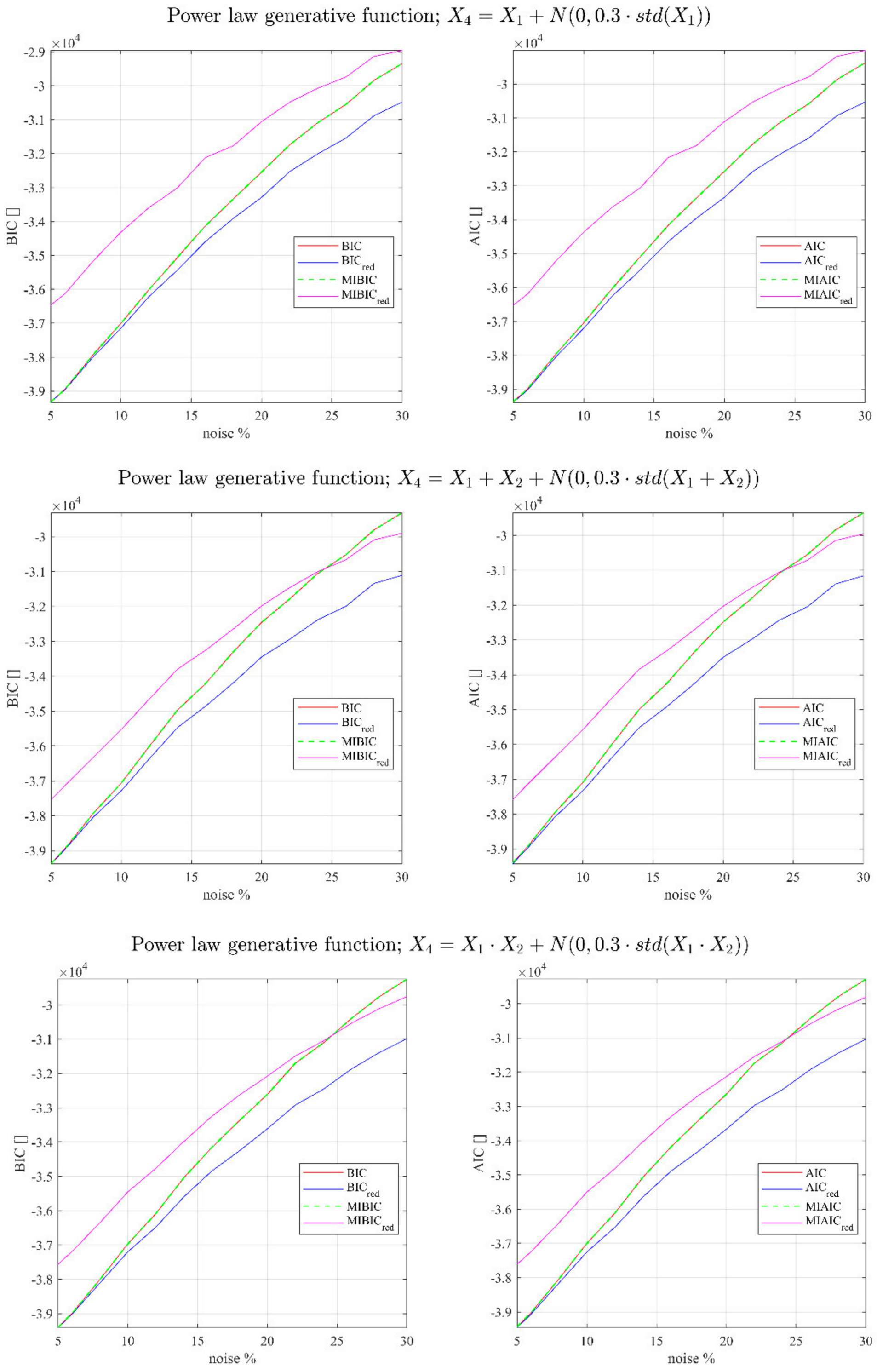
$$X_4 = X_1 \cdot X_2 + N\ (\mu = 0, \sigma = \sigma_N = 0.3 \cdot std(X_1 \cdot X_2)) \tag{24}$$

The results of the analysis are also reported in Figure 1.

The same analysis has been repeated for polynomial and exponential types of functions. The functions used to generate the data are the following, respectively:

$$Y = X_1 + X_2 + X_3 + X_1^2 + X_2^2 + X_3^2 \tag{25}$$

$$Y = X_1 \cdot X_2 \cdot X_3 \cdot e^{X_1 + X_2 + X_3} \tag{26}$$



Power law generative function; $X_4 = X_1 + N(0, 0.3 \cdot std(X_1))$

Power law generative function; $X_4 = X_1 + X_2 + N(0, 0.3 \cdot std(X_1 + X_2))$

Power law generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$

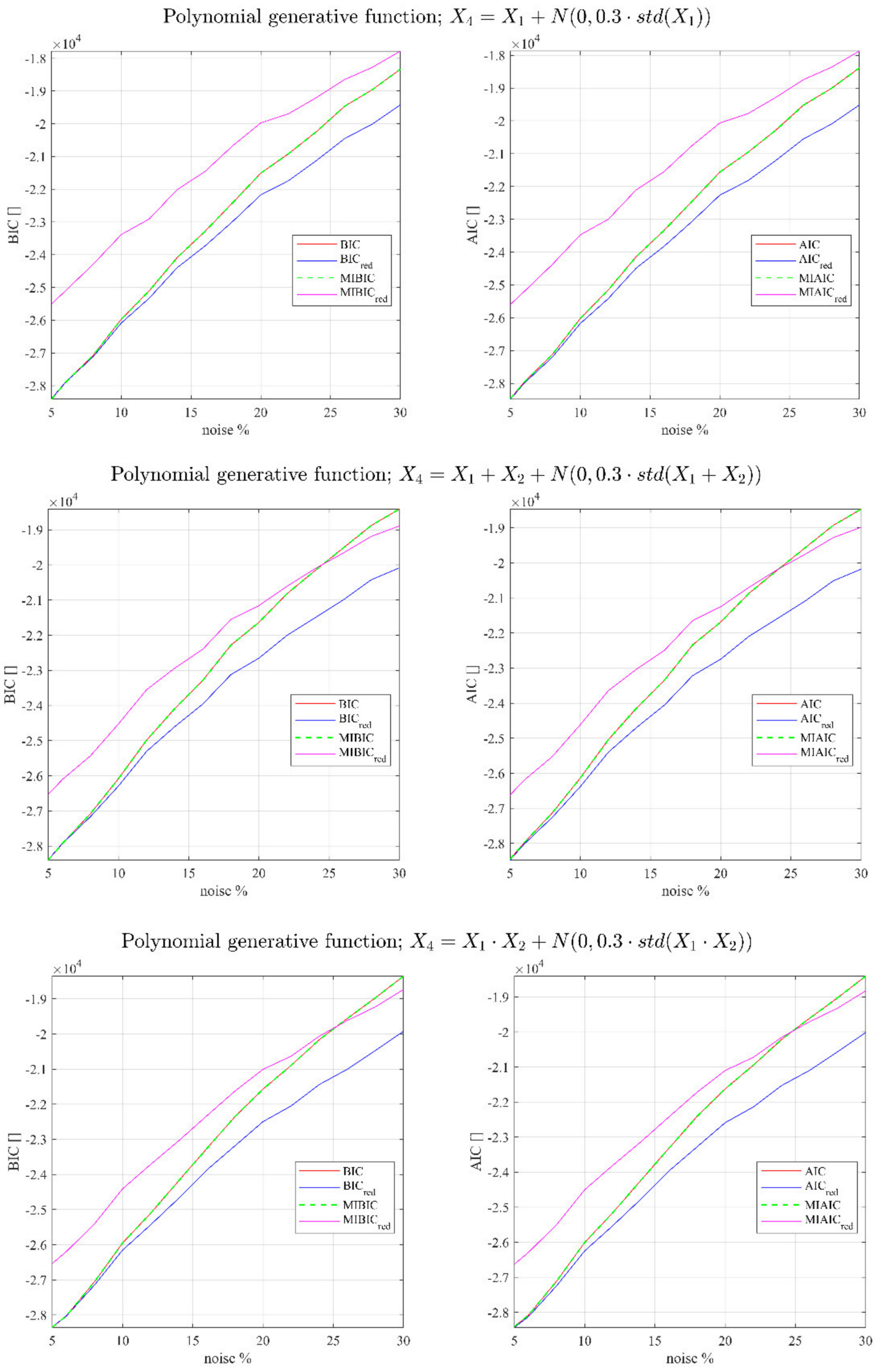
Figure 1. BIC, BICred, MIBIC, and MIBICred for the power-law generative function and different correlations as a function of the percentage of noise in the predictors. The subscript red indicates the models using the redundant regressors.

While the functions used to fit the data are respectively of the form

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_N \boldsymbol{x}_N + \alpha_1 \boldsymbol{x}_1^2 + \alpha_2 \boldsymbol{x}_2^2 + \cdots + \alpha_N \boldsymbol{x}_N^2 \tag{27}$$

$$Y = \alpha_0 \cdot X_1 \cdot X_2 \ldots \cdot X_N \cdot e^{\alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_N X_N} \tag{28}$$

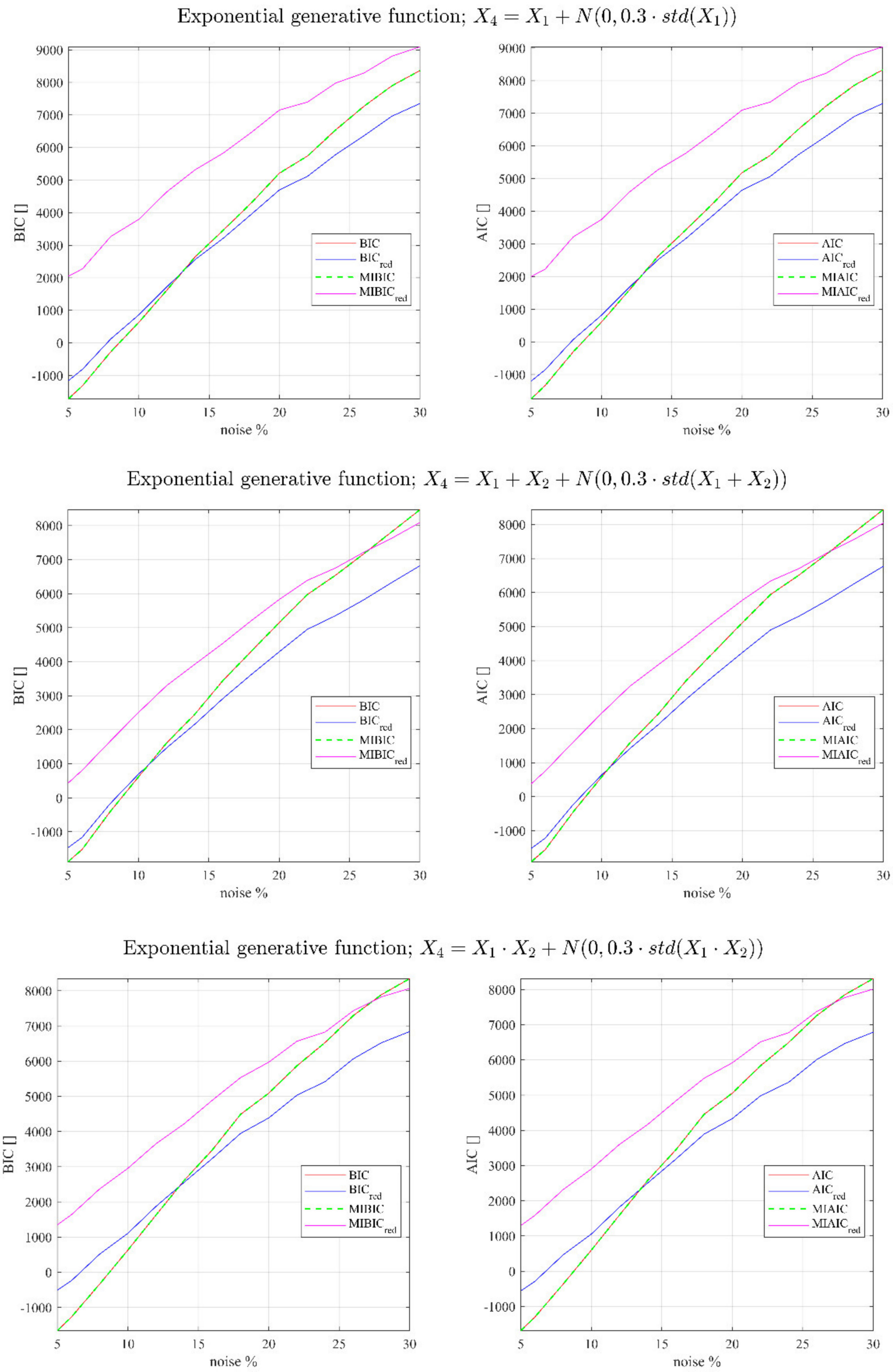
Fitting (27) and (28) to the noised values of Y with and without the redundant predictor and evaluating the traditional and the modified version of AIC and BIC, provides the results reported in Figures 2 and 3. All the results shown have been obtained using $n = 5000$ data points.


Polynomial generative function; $X_4 = X_1 + N(0, 0.3 \cdot std(X_1))$

Polynomial generative function; $X_4 = X_1 + X_2 + N(0, 0.3 \cdot std(X_1 + X_2))$

Polynomial generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$

Figure 2. BIC, BICred, MIBIC, and MIBICred for the polynomial generative function and different correlations as a function of the percentage of noise in the predictors. The subscript red indicates the models using the redundant regressors.



Exponential generative function; $X_4 = X_1 + N(0, 0.3 \cdot std(X_1))$

Exponential generative function; $X_4 = X_1 + X_2 + N(0, 0.3 \cdot std(X_1 + X_2))$

Exponential generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$

Figure 3. BIC, BICred, MIBIC, and MIBICred for the exponential generative function and different correlations as a function of the percentage of noise in the predictors. The subscript red indicates the models using the redundant regressors.

One important thing to notice in order to interpret the next figures is that without the redundant regressors, MIBIC and MIAIC provide exactly the same results as the traditional AIC and BIC, proving the consistency of the devised new indicators. On the other hand, if the redundant variables are added to the inputs, the traditional versions of the indicators would always select the wrong model (they assume a lower value), whereas the upgraded versions are not misled (the new indicators assume always higher values than when the redundant quantities are not considered).

In all the reported cases, except for small percentages of noise in the predictors, the traditional version of AIC and BIC are not able to identify the correct model, showing a tendency to select the models including redundant regressors. On the contrary, a significant improvement in the ability to detect the right model is achieved by the MIBIC as well as by MIAIC, which fails only in some cases when the noise percentage is significant.

The effect of the noise in the dependent variable has also been evaluated, but the results of the analysis are not significantly different and the conclusions are the same as for the examples reported.

5. Application to a Real-Life Database

In this section, to prove the generality of the results obtained with the upgraded criteria described in the previous sections, an application to a real-life database has been considered. The analysed database is called the ITPA database, which is the most advanced Multi-machine database built to support studies of plasma confinement in Tokamaks [10]. A description of this database is given in the next subsection. The results obtained applying the criteria developed in the present work to this database are reported in the following subsection.

5.1. The ITPA Database of the Energy Confinement Time for the H Mode

One of the most crucial quantities to assess the relevance of a nuclear fusion reactor is the so-called energy confinement time τE, which quantifies how fast the internal energy of the plasma is lost [11–13]. Unfortunately, the transport mechanisms affecting the energy confinement in high-temperature plasmas are very complex and nonlinear, including effects at many scales. So, even if the understanding of the instabilities and turbulence effects influencing transport has progressed a lot in the last years, a theoretical or numerical solution for the proper estimation of the energy confinement time, τ_E, remains unfeasible. As a consequence, this problem has been approached empirically with the extraction of robust scaling laws for τE from experimental data. This led to the construction of several multi-machine databases for the plasma confinement time, including the ITPA database analysed in this paper. In particular, the DB3v13f version of the ITPA with the same selection rules reported in [10] is the one used in the following analysis.

The variables that are known to be relevant for the estimation of the confinement time and that will be taken into consideration in this work are Ip, BT, $PLTH$, nel, $Meff$, $RGEO$, ϵ, and ka, where Ip is the plasma current, BT is the toroidal magnetic field, $PLTH$ is the power loss across the last closed surface, nel is the line average electron density, $Meff$ is the plasma isotopic composition, $RGEO$ is the plasma major radius, $\epsilon = \frac{a}{R_{GEO}}$ where a is the plasma minor radius, and ka is the volume measure of elongation [10]. Indeed, these variables are the ones used in the most widely accepted scaling law for the Tokamak energy confinement time in H mode, called the IPB98(y,2):

$$\tau_E = 5.62 \cdot 10^{-2} \cdot I_p^{0.93} \cdot BT^{0.15} \cdot n_e^{0.41} \cdot P^{-0.69} \cdot R^{1.97} \cdot k_a^{0.78} \cdot \epsilon^{0.58} \cdot M_{eff}^{0.19} \tag{29}$$

Due to the physical constraints and the fact that each machine is optimized to work within specific parameter ranges, the degree of correlation of the mentioned regressors is quite high, as shown in Table 1.

Moreover, the regressors, as well as the confinement time, are affected by significant uncertainties, as shown in Table 2.

Table 1. Person correlation coefficient matrix for the eight regressors used to model the energy confinement time.

	ϵ	M_{eff}	R_{GEO}	k_a	B_T	I_P	n_e	P_{LTH}
ϵ	1.00	0.29	0.11	0.41	−0.02	0.41	0.10	0.26
M_{eff}	0.29	1.00	0.30	0.42	0.17	0.38	0.16	0.36
R_{GEO}	0.11	0.30	1.00	0.30	0.09	0.73	−0.36	0.67
k_a	0.41	0.42	0.30	1.00	−0.07	0.48	0.15	0.42
B_T	−0.02	0.17	0.09	−0.07	1.00	0.43	0.52	0.34
I_P	0.41	0.38	0.73	0.48	0.43	1.00	0.00	0.76
n_e	0.10	0.16	−0.36	0.15	0.52	0.00	1.00	0.03
P_{LTH}	0.26	0.36	0.67	0.42	0.34	0.76	0.03	1.00

Table 2. Lower bounds of the uncertainties for the entries of the ITPA database.

	ϵ	M_{eff}	R_{GEO}	k_a	B_T	I_P	n_e	P_{LTH}	τ_E
Rel. err.	1%	8%	1%	10%	1%	1%	5%	14%	10%

5.2. Results

Employing the upgraded model selection criteria proposed in this work, the main objective of the analysis consists of identifying, within all the possible power-law models obtained combining the predictor variables included in Equation (29), the one which best represents the τ_E data.

In order to do this, the following iterative procedure has been adopted:

The first step consists of evaluating the MIAIC and MIBIC for the power-law model with all the eight regressors included. Then, removing one variable at a time from the list of regressors, eight more models are obtained and their MIAIC and MIBIC evaluated.

The model with the lowest MIAIC/MIBIC is identified, and the regressors included in the model will form the new list of possible regressors. The process is then iterated eliminating one variable at the time, until removing any of the variables included in the list does not produce any benefit in terms of MIAIC and MIBIC. At this point, the algorithm is topped and the best model is retained

Applying this procedure, the model which shows itself to be the best in terms of both MIAIC and MIBIC is

$$\tau_E = \alpha_0 \cdot I_p^{\alpha_1} \cdot P^{\alpha_2} \cdot R^{\alpha_3} \cdot k_a^{\alpha_4} \cdot \epsilon^{\alpha_5} \cdot M_{eff}^{\alpha_6} \tag{30}$$

Instead, using the traditional AIC and BIC criteria, the resultant models are, respectively,

$$\tau_E = \alpha_0 \cdot I_p^{\alpha_1} \cdot BT^{\alpha_2} \cdot n_e^{\alpha_3} \cdot P^{\alpha_4} \cdot R^{\alpha_5} \cdot k_a^{\alpha_6} \cdot \epsilon^{\alpha_7} \cdot M_{eff}^{\alpha_8} \tag{31}$$

$$\tau_E = \alpha_0 \cdot I_p^{\alpha_1} \cdot BT^{\alpha_2} \cdot n_e^{\alpha_3} \cdot P^{\alpha_4} \cdot R^{\alpha_5} \cdot k_a^{\alpha_6} \cdot \epsilon^{\alpha_7} \tag{32}$$

The first obvious advantage of the upgraded versions of the criteria is that they provide coherent results, whereas the traditional versions of the indicators do not seem to agree on a single model, rendering the choice of the most appropriate scaling law very difficult. More importantly, the model obtained with MIAIC and MIBIC is more parsimonious, and indeed, it utilises fewer quantities than the ones derived by the AIC and IC. It retains the plasma current but considers redundant magnetic field and plasma density. This is coherent with the statistical analysis of the database, which presents very strong collinearities between these three quantities, as reported in Table 1. The obtained results are also in harmony with the everyday experience of the device operators, since the experiments are indeed typically designed with strong correlations between plasma parameters.

6. Conclusions

In applications to regression, the most widely used versions of the model selection criteria AIC and BIC are vulnerable to the presence of variables correlated to the actual predictors, particularly when the percentage of noise in the regressors is not negligible, as it is in most practical applications. To address this problem, an upgraded version of these criteria is proposed, adding an a priori Chi-squared probability distribution function of the models. This function depends on a quantity that penalizes the model with highly correlated predictors, which bring little new information about the dependent variable. The performance of the proposed criteria has been assessed with different types of generative functions, correlation functions and percentage of noise in the predictors. The results indicate that, in most cases, the newly defined criteria possess an improved capability of detecting redundancy in the predictors and thus of selecting the correct model. The improved performances are not substantially affected by the sample size as reported in Appendix A. To show the generality of the obtained results, an application to an international database built by the thermonuclear fusion community has also been reported in the final section.

With regard to future developments, from a methodological standpoint, it would be interesting to improve the treatment of the uncertainties in both the dependent and independent variables, implementing techniques inspired by the error in the variable approach [10]. Moreover, the introduction of metric alternatives to the Euclidean, such as the geodesic distance [14–16], has the potential to provide significant added value. An additional interesting activity would be the systematic analysis of possible prior alternatives to the one chosen for the present version of MIBIC and MIAIC. In terms of applications, the scaling laws of the more recent metallic Tokamaks, and particularly JET with the new ITER-like wall [17], are nowadays a topic of great interest in the fusion community. The new versions of the indicators could become quite useful in the investigation of scaling laws in non-power law monomial form [18–21].

Author Contributions: Conceptualization, L.S. and A.M.; methodology, L.S. and A.M.; software, L.S.; validation, A.M. and R.R.; formal analysis, L.S.; writing—original draft preparation, L.S. and A.M.; writing—review and editing, R.R. and M.G. All authors have read and agreed to the published version of the manuscript.

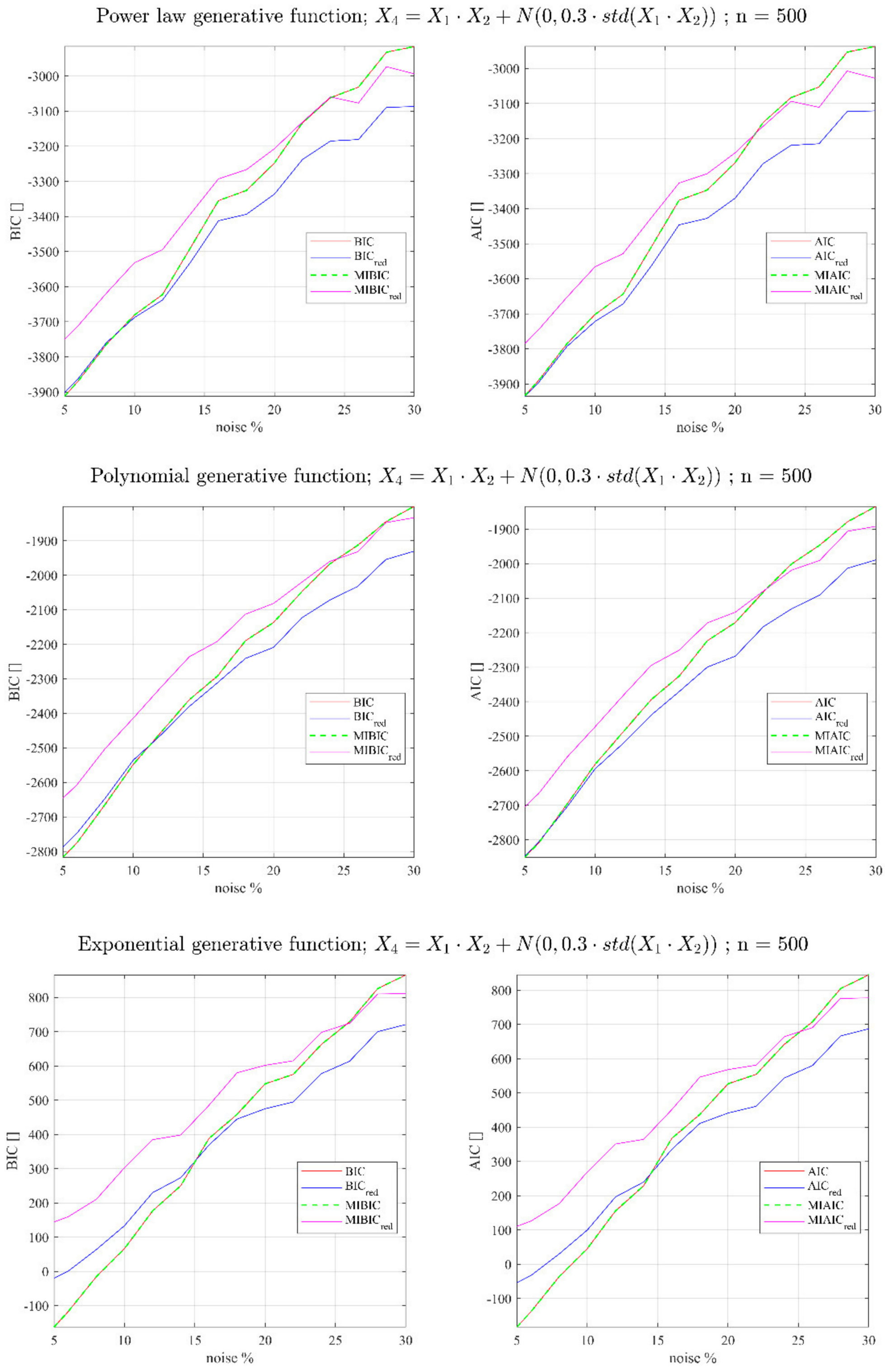
Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A. MIBIC and MIAC Performance with Sample Size

The Appendix A contains the analysis of the effect of the number of samples on the performance of the proposed upgraded criteria. In particular, the same tests reported in Section 4 have been repeated for two different sample sizes $n = 500$ and $n = 50,000$. The results of the analysis for the different generative functions and number of entries are reported in Figures A1 and A2. For simplicity, only the results for the correlation function reported in (24) have been included in the figures. As it can be noted by visual insèecion of the plots reported in this Appendix A, the performance's improvement associated with the proposed criteria is not affected by the sample size. This is a general result: the MIBIC and MIAIC perform better than the traditional versions independently from the number of entries in the datbases. These results can be also extended to the other correlation functions considered in Section 4, even though the related figures have not been explicitly reported in the paper.

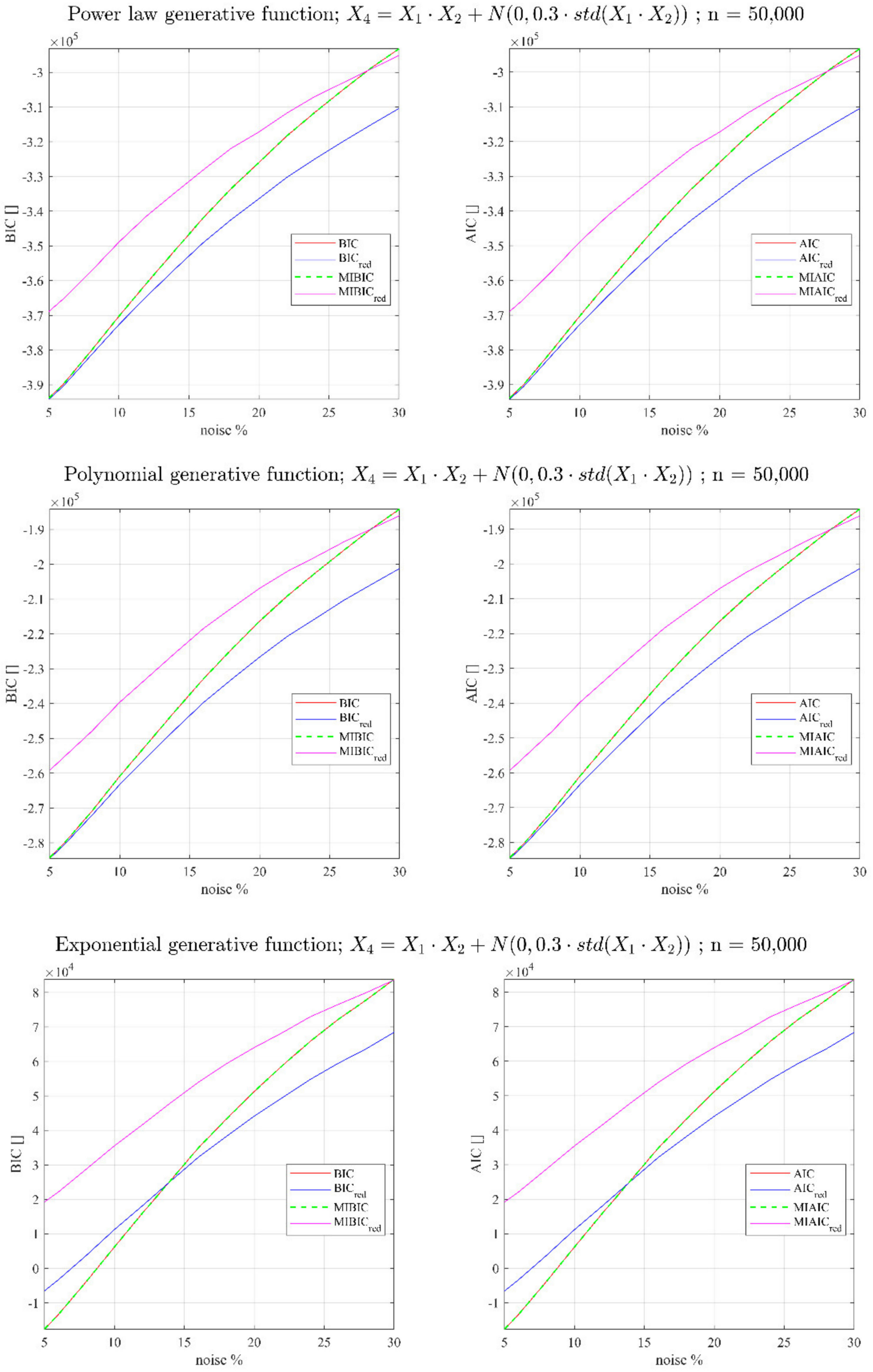


Power law generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$; n = 500

Polynomial generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$; n = 500

Exponential generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$; n = 500

Figure A1. BIC, BICred, MIBIC, MIBICred for n = 500 and different generative functions vs. the percentage of noise in the predictors. The subscript red indicates the models using the redundant regressors.



Power law generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$; n = 50,000

Polynomial generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$; n = 50,000

Exponential generative function; $X_4 = X_1 \cdot X_2 + N(0, 0.3 \cdot std(X_1 \cdot X_2))$; n = 50,000

Figure A2. BIC, BICred, MIBIC, MIBICred for *n* = 50000 and different generative functions vs. the percentage of noise in the predictors. The subscript red indicates the models using the redundant regressors.

References

1. Bailly, F.; Longo, G. *Mathematics and the Natural Sciences*; Imperial College Press: London, UK, 2011.
2. D'Espargnat, B. *On Physics and Philosophy*; Princeton University Press: Oxford, MS, USA, 2002.
3. Claeskens, G. Statistical model choice. *Annu. Rev. Stat. Its Appl.* **2016**, *3*, 233–256. [CrossRef]
4. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
5. Schwarz Gideon, E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]

6. Kenneth, P.B.; Anderson, D.R. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: Berlin, Germany, 2002.
7. Murari, A.; Peluso, E.; Cianfrani, F.; Gaudio, P.; Lungaroni, M. On the Use of Entropy to Improve Model Selection Criteria. *Entropy* **2019**, *21*, 394. [CrossRef] [PubMed]
8. Rossi, R.; Murari, A.; Gaudio, P.; Gelfusa, M. Upgrading Model Selection Criteria with Goodness of Fit Tests for Practical Applications. *Entropy* **2020**, *22*, 447. [CrossRef] [PubMed]
9. MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
10. McDonald, D.; Cordey, J.; Righi, E.; Ryter, F.; Saibene, G.; Sartori, R.; Alper, B.; Becoulet, M.; Brzozowski, J.; Coffey, I.; et al. ELMy H-modes in JET helium-4 plasmas. *Plasma Phys. Control Fusion* **2004**, *46*, 519–534. [CrossRef]
11. Wesson, J. *Tokamaks*, 3rd ed.; Clarendon Press: Oxford, UK, 2004.
12. Romanelli, F.; Laxåback, M. Overview of JET results. *Nucl. Fusion* **2009**, *49*, 104006. [CrossRef]
13. Ongena, J.; Monier-Garbet, P.; Suttrop, W.; Andrew, P.; Bécoulet, M.; Budny, R.; Corre, Y.; Cordey, G.; Dumortier, P.; Eich, T. Towards the realization on JET of an integrated H-mode scenario for ITER. *Nucl. Fusion* **2004**, *44*, 124–133. [CrossRef]
14. Craciunescu, T.; Murari, A. Geodesic distance on Gaussian manifolds for the robust identification of chaotic systems. *Nonlinear Dyn.* **2016**, *86*, 677–693. [CrossRef]
15. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
16. Murari, A.; Boutot, P.; Vega, J.; Gelfusa, M.; Moreno, R.; Verdoolaege, G.; de Vries, P.C.; JET-EFDA Contributors. Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruption. *Nuclear Fusion* **2013**, *53*, 033006. [CrossRef]
17. Pamela, J.; Romanelli, F.; Watkins, M.L.; Lioure, A.; Matthews, G.; Philipps, V.; Jones, T.; Murari, A.; Géraud, A.; Crisanti, F.; et al. The JET programme in support of ITER. *Fusion Eng. Des.* **2007**, *82*, 590–602. [CrossRef]
18. Murari, A.; Lupelli, I.; Gelfusa, M.; Gaudio, P. Non-power law scaling for access to the H-mode in tokamaks via symbolic regression. *Nucl. Fusion* **2013**, *53*, 043001. [CrossRef]
19. Murari, A.; Lupelli, I.; Gelfusa, M.; Peluso, E. Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form. *Plasma Phys. Control Fusion* **2015**, *57*, 014008. [CrossRef]
20. Murari, A.; Lupelli, I.; Gelfusa, M.; Peluso, E.; Lungaroni, M. Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities. *Nucl. Fusion* **2015**, *56*, 26005. [CrossRef]
21. Murari, A.; Lupelli, I.; Gelfusa, M.; Gaudio, P.; Vega, J. A statistical methodology to derive the scaling law for the H-mode power threshold using a large multi-machine database. *Nucl. Fusion* **2012**, *52*, 063016. [CrossRef]