Artificial Intelligence Review (2023) 56:S2825–S2859
https://doi.org/10.1007/s10462-023-10591-4

A practical utility-based but objective approach to model selection for regression in scientific applications

Andrea Murari[1,2] · Riccardo Rossi[3] · Luca Spolladore[3] · Michele Lungaroni[3] · Pasquale Gaudio[3] · Michela Gelfusa[3]

Published online: 5 October 2023

Abstract
In many fields of science, various types of models are available to describe phenomena, observations and the results of experiments. In the last decades, given the enormous advances of information gathering technologies, also machine learning techniques have been systematically deployed to extract models from the large available databases. However, regardless of their origins, no universal criterion has been found so far to select the most appropriate model given the data. A unique solution is probably a chimera, particularly in applications involving complex systems. Consequently, in this work a utility-based approach is advocated. However, the solutions proposed are not purely subjective but all based on "objective" criteria, rooted in the properties of the data, to preserve generality and to allow comparative assessments of the results. Several methods have been developed and tested, to improve the discrimination capability of basic Bayesian and information theoretic criteria, with particular attention to the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) indicators. Both the quality of the fits and the evaluation of model complexity are aspects addressed by the advances proposed. The competitive advantages of the individual alternatives, for both cross sectional data and time series, are clearly identified, together with their most appropriate fields of application. The proposed improvements of the criteria allow selecting the right models more reliably, more efficiently in terms of data requirements and can be adjusted to very different circumstances and applications. Particular attention has been paid to ensure that the developed versions of the indicators are easy to implement in practice, in both confirmatory and exploratory settings. Extensive numerical tests have been performed to support the conceptual and theoretical considerations.

Keywords Model selection criteria · Bayesian Information Criterion (BIC) · Akaike Information Criterion (AIC) · Shannon entropy · Goodness of fit tests · Mutual information · Feedback loops

Extended author information available on the last page of the article

1 Introduction: a short overview of model selection criteria and machine learning to motivate a utility-based approach

The summit of any scientific endeavour is the formulation of an appropriate theory to describe the phenomena under study (Bailly and Longo 2011; D'Espargnat 2002) A fundamental ingredient in any theory, at least in the so-called exact disciplines, is the availability of a satisfactory mathematical model. Therefore, the selection of the most appropriate model, to interpret the evidence and to make predictions, is a major task of modern research. In most of the history of science, the models to be validated with experimental data were hypothesis driven, i.e., derived from previous knowledge or theories. Nowadays, various aspects of most research projects are addressed with artificial intelligence tools. Consequently, the objectives of model selection have become more numerous and various, and they are not limited simply to choosing the analytic formula best fitting the data. Some examples of these new aspects of models selection are the identification of: the basis terms for polynomial or wavelet functions, the structure of machine learning tools (number of neurons in networks or kernel in SVM), the order of autoregressive models, the most appropriate parametric family, the number of components in a mixture model, to name just a few (Ding 2018; Stoica and Selen 2004). Moreover, modern sensor and storage technologies allow acquiring enormous amounts of data about the most diverse phenomena. Machine learning tools are therefore often deployed to produce models themselves. Recently evolutionary computational methodologies, such as Genetic Programming based Symbolic Regression (GPSR) (Schmid and Lipson 2009; Murari et al. 2019, 2020), have been developed to extract mathematical models directly from data with a minimum of a priori hypotheses and constraints.

Whatever the source of the models, scientists and statisticians have naturally been devoting a lot of attention to their selection. Not surprisingly, many approaches have been proposed over the years, but no general consensus has emerged on a single technique or procedure (Ding 2018). Such an unsatisfactory position is certainly not the consequence of lack of efforts. In the last decades, the issue of model selection has been addressed with Bayesian methods (Key et al. 1999) (Mark 2018), frequentist techniques (Miller 2002) and even information theoretic indicators (Claeskens 2016; Kenneth and Anderson 2002). All these approaches have their strong points and weaknesses, which render them more suited to certain applications than others. The present work is motivated by the observation that the ultimate goal of devising a best technique for all possible situations and tasks is probably a chimera. The approach, which informs the present work, is based on the assumption that the lack of generality of model selection criteria (MSC) is not their intrinsic limitation but a consequence of the fact that any optimal solution is contingent on the specific application.

The context dependence of model selection can be appreciated from the following considerations. MSC need to be deployed in a great variety of situations and with completely different objectives. The measurements can have various nature, from time series to cross sectional data or probability distribution functions (pdfs). The goals of the analysis can range from purely observational science to direct deployment in real life, with potential vital consequences, for example in medicine and engineering. The interest of the researchers can be exploratory or confirmatory. It is therefore not unreasonable to expect that, given the context dependence of the input data and the objectives of the investigations, a single fit all solution is probably neither realistic nor desirable. It should also be considered that qualitative knowledge and many forms of prior information are impossible to

accommodate in too rigid universal criteria. On the other hand, completely subjective criteria are not satisfactory either. Being typically too arbitrary they are prone to various errors, such as conscious or unconscious bias and data dredging. Moreover, too subjective criteria do not to allow for meaningful comparisons and the growth of knowledge, essential elements of any scientific endeavour.

In the present work the proposed approach is therefore utility based but "objective"; the mathematical framework is general and rooted in specific properties of the data and the models, but the user is granted the possibility to fine tune the developed tools in such a way to take into account the details and objectives of the specific application. Particular attention has been paid to formulate versions of the indicators easy to implement in practice. Indeed, often the MSC reported in the literature are well grounded theoretically but, to preserve their mathematical properties, they can be very difficult if not impossible to deploy in real life situations.

The treatment followed is based on information theoretic and Bayesian approaches, according to which the MSC are cost functions depending on two terms, one quantifying the goodness of fit and the other penalising the complexity of the models (see Sect. 2). The frequentist types of solutions are not neglected, though, but used to improve some of the versions of the criteria (see Sect. 3).

As representative of the information theoretic family of metrics, the Akaike Information Criterion (AIC) (Akaike 1974) (Cavanaugh 2019) is discussed in detail but the proposed solutions can be easily applied to the other indicators of the same mathematical background such as the Deviance Information Criterion (DIC), the Takeuchi Information Criterion (TIC), the Focussed Information Criterion (FIC) and the Kashyap Information Criterion (KIC) (Claeskens 2016; Zhou and Herath 2016).

With regard to the Bayesian framework, the popular Bayesian Information Criterion (BIC) (Schwarz 1978), (Lofti 2022) is detailed as an example of this entire set of criteria, which include the Extended Bayesian Information Criterion (EBIC) and the Extended Fisher Information Criterion (EFIC) (Ando 2010).

In Section 3, attention is devoted to improving the first term of the model selection criteria, the one qualifying the goodness of fit. The rationale for the proposed modifications derives from the consideration that the residuals (the differences between the data and the model predictions) contain much more information than the simple mean square error, the metric used in the traditional version of the indicators. Leveraging this additional information with frequentist and information theoretic methods increases appreciably the discriminatory powers of the criteria. Section 4 discusses the refinements of the second term, the quantification of complexity, which is too rudimentary in the original AIC and BIC since it is reduced to the simple number of parameters in the models. Again deploying more sophisticated quantifiers of the models' complexity has a very good impact in many applications requiring analysis of complex systems. A different set of criteria, based on weighting the pdf of the data, is the subject of Sect. 5; the modifications suggested in this section are meant to obviate the simplistic assumption that the available data are perfect and not affected by any form of uncertainties or distortions. The techniques developed to refine the databases have always a very positive effect on the quality of the final classifications. The family of functions and the noise statistics, implemented to investigate the various refinements, are summarised in Sect. 6, whose main part is devoted to an overview of the results obtained with a series of systematic numerical tests, for both cross sectional data and time series. The application to exploratory techniques, such as GPSR, which are becoming increasingly important in this era of data overload, is covered in Sect. 7. Conclusions are the subject of the last section of the paper.

2 The main aspects of model selection and their representation in terms of negative utility

Learning from data has two main objectives. The first one, typically pursued in the sciences, consists of understanding the data generation process, which means shedding light on the actual reality of the processes generating the data. The second approach is more concerned with prediction and therefore is mainly focussed on the accuracy of forecasting, independently from the fidelity to the mechanisms at play in the phenomena under study. In line with these two different, even if not necessarily incompatible goals, model selection can also have two diverse directions: *model selection for inference* and *model selection for prediction*. The aspiration of the first approach is to converge on models that reflect the underlying reality, providing insight for interpretation and eventually interventions. The studies belonging to the second framework are concerned with maximising performance in terms of predicative accuracy, without any additional requirement about the form of the models. To appreciate the distinctive specificities of the two priorities, one can consider the dependence on the number of examples. In the case of model selection for inference, the results should be independent from the size of the available databases. However, in the case of model selection for prediction, it would be perfectly legitimate to converge on different models depending on the amount of data to fit (typically in this context the larger the sample size the larger the dimensionality of the model).

The present work is mainly concerned with model selection for inference, even if the proposed improvements could be very useful also for model selection for prediction. In this perspective, the various criteria will be qualified in terms of their capability to identify the "best model" or the "right model", the actual deterministic equation or probability distribution function generating the data.

More formally, a model is a deterministic equation or a probability distribution function (pdf) used to describe a set of n samples. Even in the deterministic context, very common in many fields of the exact sciences, the data are typically affected by not negligible uncertainties, requiring a probabilistic treatment. Indeed, the noise is typically considered randomly generated by a given pdf (see later).

In the framework of model selection for inference, one very important property of MSC is consistency or asymptotic convergence, which means that they select the best model in the limit of infinite samples according to the following definition.

Definition A model selection procedure is consistent or asymptotically convergent if it selects the best model with probability converging to one for n tending to ∞.

The goodness of fit and the rate of convergence of a model are typically determined in terms of a loss function, normally called cost function in utility-based treatments. In the field of model selection, the loss functions typically depend on the residuals, the differences between the estimates and the true values of the data. It will be shown in more detail later that the most widely used cost functions are proportional to the logarithm of the Euclidian norm of this difference.

As mentioned in the introduction, many tools for identifying the "best model", among a set of candidates, have been reported in the literature (Breiman 2001). The Akaike Information Criterion AIC is an information theoretic indicator, derived from the Kullback–Leibler divergence, which is basically designed to quantify the information lost by a given model when representing the data (Kenneth and Anderson 2002). The basic principle underlying the AIC

criterion is indeed the consideration that the less information a model loses, the higher its quality. The theoretical derivation of the AIC provides the following unbiased form of the criterion:

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

where L is the likelihood of the model given the data and k the number of estimated parameters in the model.

Bayesian theory informs the Bayesian Information Criterion, which is designed to maximize the posterior probability of a model given the data (Ando 2010). The most general form of BIC is:

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

where again 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.

Both AIC and BIC metrics are negative utility indicators, which have to be minimized; the best models are the ones with the lowest values of the criteria. They also have the same conceptual structure. The first term favours models with a high likelihood, the second implements a penalty for complexity (the term proportional to k). Therefore, these two MSC, and the many variations belonging to their families, try to find an optimal and universal compromise between goodness of fit and complexity, without leaving any margins of manoeuvring to the user.

In this work the potential of a more general interpretation of Eqs. (1) and (2) (and of their respective families) is investigated. The main framework of AIC and BIC is retained; the solution is expected to be a trade-off between goodness of fit and complexity. On the other hand, the form of each of the two terms can be modified, depending on the nature of the problem to be solved and the objectives of the analysis. Therefore, in general the proposed enhancements of the model selection criteria (CRIT), discussed in the rest of the paper, can be cast in the context of utility-based indicators of the form:

$$CRIT = NU_{GoF} + NU_{Compl} \tag{3}$$

where NU_{GoF} is the negative utility to be associated to the goodness of fit and NU_{Compl} the negative utility to be attributed to complexity. As for the traditional AIC and BIC, the utilities are negative and therefore can be interpreted as cost functions. Consequently, also all the proposed criteria of the form (3) are indicators to be minimised, the better the model the lower their value.

To improve their potential for many applications in the sciences, the first consideration to be kept in mind is that the original formulation of the AIC and BIC criteria is not necessarily easy to implement in practice. The most delicate part is the likelihood of the models, which can be virtually impossible to calculate. This difficulty can be due to various causes: the type of noise affecting the data, the nature of the models to be tested, lack of the "a priori "information about the properties of the systems under study etc. The traditional assumption, that the data are identically distributed and independently sampled from a normal distribution, is the most common step taken to bypass the practical difficulties of calculating the likelihood. If this hypothesis is valid, it is possible to demonstrate that the AIC can be expressed (up to an additive constant, which depends only on the number of entries in the database and not on the model) as:

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

where MSE is the mean-squared error of the residuals, the differences between the data and the predictions of the models.

Similar assumptions allow expressing the BIC criterion as:

$$BIC = n \cdot \ln\left(\sigma_{(\epsilon)}{}^2\right) + k \cdot \ln(n) \tag{5}$$

where $\sigma_{(\epsilon)}{}^2$ is the variance of the residuals.

Versions (4) and (5), formally derived in (Kenneth and Anderson 2002), constitute the most widely used expressions of AIC and BIC. As can be easily appreciated by inspection of these equations, the statistical information, originally provided by the likelihood, is reduced to the mere MSE and variance of the residuals (Wang and Bovik 2009). A natural question is whether additional statistical information, about the distribution of the residuals, could be taken into account and improve the performance of the two criteria. The practical importance of this question is not to be underestimated also because, in many applications, the assumptions behind Eqs. (4) and (5) are clearly violated. In real life, indeed, the statistics of the noise can be different from a Gaussian, memory effects can be important, correlations between noise and measurements can be present etc. How to improve the model selection criteria in this respect, with objective indicators based on the properties of the data, is the subject of Sect. 3.

Another important line of investigation involves the second term in the AIC and BIC. Indeed, it is well known that the simple number of parameters is a very poor quantifier of the complexity of a model (Vapnik 2000). Very sound and quite sophisticated criteria to quantify the complexity of functions do exist; probably two of the most solid and advanced are the VC dimension (Vapnik 2000) and the Rademacher dimension (Bartlett and Mendelson 2002). In addition to the obvious difficulty that an exact estimate of their value is not available for the vast majority of functions, their practical and approximate evaluations are also very problematic and computationally intensive (McDonald, Shalizi, and Schervish, 2011; Chen et al. 2020). More importantly in the context of the present work, both criteria, which by the way are related, are not really suited to the objectives of model selection, in which the complexity measure is meant at reducing overfitting. Indeed, the VC and the Rademacher dimensions have been conceived to determine the complexity of entire families of functions, for example the polynomials of order n. However, depending on their parameters, functions belonging to the same class can have completely different overfitting capability (for example, a high order polynomial can be very smooth or very flexible depending on the value of its coefficients). Consequently, the VC and the Rademacher dimensions have limited discriminatory power in this sense. In many applications, different requirements could therefore be better satisfied by alternative estimates of complexity, possibly less universal and theoretically sound but of more direct practical value. The improvements of model selection criteria, based on alternative but again "objective" definitions of complexity, are discussed in Sect. 4.

A different approach to the definition of utility relates to the probability distribution function of the data. For a variety of reasons, the experience and knowledge of the experts can strongly suggest attributing more or less importance to various parts of the data probability distribution function (pdf). A typical example from economics is the relevance of rare events in influencing quantities, normally poorly correlated. These events should be assigned strong weights because they can have important consequences. In other applications, rare events have to be considered outliers and should therefore be eliminated or granted low importance. These issues can be addressed either by weighting differently various parts of the pdf or by implementing some form of robust statistics. These approaches

again fall in the framework of utility-based techniques but involve both terms in Eq. (3) and are therefore discussed in detail in Sect. 5.

The three main families of improvements are summarised graphically in Fig. 1 and are presented in the text in order of increasing generality of application (from the bottom up in Fig. 1). The techniques dealing with the goodness of fit tests are mainly aimed at improving the MSC when the likelihood is not computable for whatever reason. The complexity term is a delicate issue that would benefit from refinements in practically all the existing criteria. The interventions on the pdf are absolutely general and can be a preliminary step to the calculation of any estimator, including frequentist indicators. Even if the likelihood can be calculated, the other proposed improvements (complexity estimate and weighting the pdf) remain valuable. The version to be preferred depends on the situation, on the objectives of the analysis and on the available a priori knowledge about the systems under investigation.

For all the families of developed modifications, careful attention has been paid to their practical implementation. The indicators proposed do not require excessive computational resources. Moreover, the assumptions they are based on are typically more realistic than the ones of the traditional criteria. In addition, the demands in terms of "a priori" knowledge are more than reasonable for the typical applications in most fields. Excessive arbitrary and subjective solutions are avoided. For example, the improved tools proposed do not require estimating the prior probability of the various models, a very delicate issue, which will be discussed in more detail in Sect. 8.

3 Improved cost functions for the goodness of fit term

The main consideration, behind the improvements proposed in this section, is that the residuals of a perfect model should simply contain the noise affecting the measurements. Therefore, the better the model, the lower the structure of the residuals and the closer their pdf to the one of the noise. On this basis, various improvements of the NU_{GoF} cost function

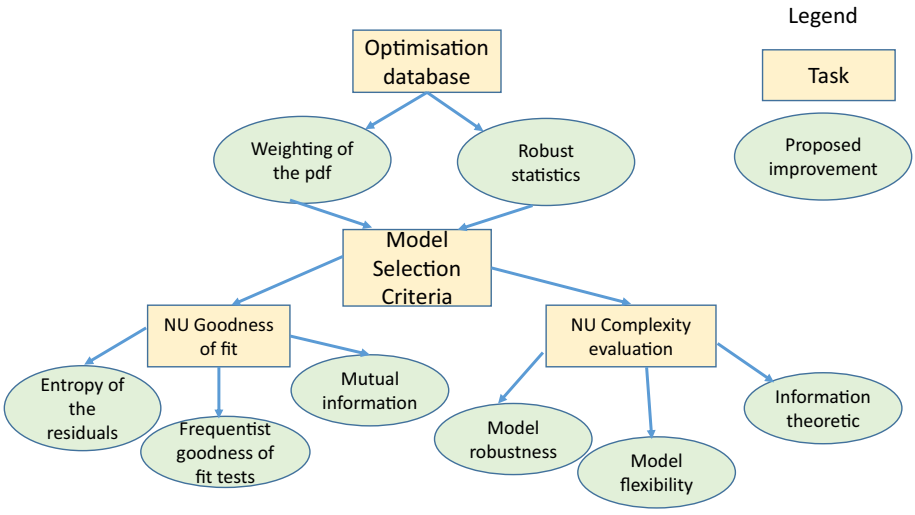

Fig. 1 Block diagram summarising the various improvements proposed in the paper. NU is the acronym for Negative Utility (see text)

can be implemented to improve the discriminatory capability of the criteria. It should be emphasised one more time that all the proposed solutions are objective, in the sense that the indicators are all based on well-defined properties of the available data.

3.1 The entropy of the residuals

A quite general way to improve the practical implementation of AIC and BIC is based on the Shannon entropy of the residuals. Indeed, the main idea behind this way to ameliorate the treatment of the residuals is the observation that, if a model were perfect, the residuals should reflect the statistics of the noise contaminating the data. In many fields of science and engineering, it is reasonable to assume that the noise affecting the measurements is additive and random. Therefore, other things being equal, models, whose residuals present a more uniform pdf, should be considered of better quality. It is well known that the Shannon entropy H can be interpreted as an indicator of how uniform a distribution is. The entropy of the residuals can therefore be included in the AIC and BIC, to favour models with a more uniform pdf of the residuals and consequently a higher value of H. In this perspective, the following versions of the BIC and AIC criteria are proposed:

$$AIC_H = n \cdot \ln\left(\frac{MSE}{H}\right) + 2k \tag{6}$$

$$BIC_H = n \cdot \ln\left(\frac{{\sigma_{(\epsilon)}}^2}{H}\right) + k \cdot \ln(n) \tag{7}$$

where $H = -\sum_i p_i ln p_i$ indicates the Shannon entropy of the residuals and p_i is the probability of the i-th residual. A formal justification of the expressions (6) and (7) requires the demonstration that the new forms of the indicators are asymptotically unbiased:

Lemma 1 *In the hypotheses used to derive the practical versions of the BIC and AIC, relations* (6) *and* (7) *possess the property of asymptotic convergence.*

Indeed, under the assumptions that the residuals are normally distributed, homoscedastic (constant variance σ) and with vanishing expectation value, the Shannon entropy can be written as:

$$H = \sum_{i=1}^{n} p_i(-\ln p_i) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}} \left[\frac{\left(y_i-\hat{y}_i\right)^2}{2\sigma^2} + \ln\left(\sqrt{2\pi}\sigma\right)\right] \tag{8}$$

where y_i denote the measured values and $\hat{y}_i$ are the predictions, which are specific of the adopted models. In the limit $n \to \infty$, the summation can be replaced by the integral across the entire probability distribution. The Shannon entropy can be then explicitly computed finding:

$$H = \frac{1}{2} + \ln\left(\sqrt{2\pi}\sigma\right) \tag{9}$$

Equation (9) does not contain the predictions $\hat{y}_i$ and therefore it is independent from the chosen model. Consequently, asymptotically the Shannon entropy contributes the same numerical factor to all models, implying that the new BIC_H and AIC_H criteria coincide

with the standard ones in the limit $n \to \infty$. The new BIC_H and AIC_H indicators therefore inherit the asymptotic converge properties of their traditional counterparts (q.e.d.).

It is worth mentioning that very similar lemmas, which are not reported for brevity's sake (the interested reader is referred to the literature), are valid also for the other improvements proposed in SubSect.s 3.2 and 3.3.

3.2 Frequentist goodness of fit tests

As already reported in (Murari 2019), the versions of the criteria including the entropy of the residuals, Eqs. (6) and (7), clearly outperform the traditional version of the AIC and BIC. They also possess nice properties of convergence. On the other hand, the entropy is not a completely satisfactory metric. First, in many cases the simple assumption of zero-sum Gaussian noise is not valid. Moreover, entropy is a quite blunt indicator of the residual distributions. It is therefore not unreasonable to question the general applicability of AIC_H and BIC_H. If the statistics of the uncertainties is known, more refined cost functions can further improve the discriminatory power of the criteria and provide more flexibility to the user, in agreement with the adopted philosophy of utility-based criteria.

In this perspective, it has proved useful to make recourse to frequentist techniques and in particular to various goodness of fit tests (for example Chi-squared, Anderson Darling and Kolmogorov–Smirnov) (Corder and Foreman 2014). For the null hypothesis, it is assumed that the residuals present the same pdf as the noise or uncertainties in the data. The outputs of the goodness of fit tests can be expressed in terms of their Z score; the lower its value, the closer the residuals to the pdf of the null hypothesis. Since the AIC and BIC criteria are cost functions, i.e., they are indicators to be minimised, the Z scores of the goodness of fit tests can be naturally included in their mathematical expressions as follows:

$$AIC_{GF} = n \cdot \ln\left(\frac{MSE}{H}\left(1 + Z^2_{score}\right)\right) + 2k \tag{10}$$

$$BIC_{GF} = n \cdot \ln\left(\frac{\sigma^2_{\epsilon}}{H}\left(1 + Z^2_{score}\right)\right) + kln(n) \tag{11}$$

where the subscript GF stands for Goodness of Fit. This new version of the AIC and BIC criteria is quite intuitive to interpret. The better the model, the closer the residuals to the pdf of the noise and therefore the lower the Z_{score} of the residuals, which tends to reduce the numerical value of the criteria. Equations (10) and (11) also grant the asymptotic convergence of AIC_{GF} and BIC_{GF}; if the model is perfect, the Z_{score} will tend to zero with increasing the number of points and the AIC_{GF} and BIC_{GF} will converge to AIC_H and BIC_H (Rossi 2020). This formulation gives the practitioner the opportunity to exploit any prior knowledge about the nature of the uncertainties affecting the data, a fact that can have a major impact on the quality of the results in many real-life applications.

3.3 Information theoretic estimate of the goodness of fit

A complementary analysis of the goodness of fit can be performed using the Mutual Information (MI) (Arndt 2004) (Baudot 2019). This alternative is particularly useful when the statistics of the uncertainties cannot be determined and therefore the null hypothesis cannot be formalised. In such a situation, the approach described in the previous subsection is

not viable. In any case, again, the basic observation is that, in the hypothesis of a perfect model, the residuals should contain only noise. Under the additional assumption, verified in many applications, that the sources of noise are not correlated with the useful signal, the mutual information between the perfect model predictions and the residuals should be zero. Moreover, the better the model the lower the MI between the model predictions and the residuals. These considerations suggest rewriting the AIC and BIC in the form:

$$AIC_{MI} = n\ln\left(MSE\left(1 + MI_{MR}\right)\right) + 2k \tag{12}$$

$$BIC_{MI} = n\ln\left(\sigma_{res}^2\left(1 + MI_{MR}\right)\right) + k\ln(n) \tag{13}$$

where MI_{MR} indicates the mutual information between the model predictions and the residuals and is defined as $MI(X;Y) = H(X) - H(X|Y)$, with again H indicating the Shannon entropy. The interest of this formulation of the MSC is manifold. In addition to good performance, it provides a completely alternative approach to the evaluation to the goodness of fit compared to the traditional frequentist criteria. Consequently, the versions of the model selection criteria given by Eqs. (12) and (13) are good complement to the formulations provided in the previous subsection.

4 Improving the measures of complexity

The subject of Sect. 3 are the improvements of the model selection criteria terms, which quantify the goodness of fit, thanks to a more sophisticated analysis of the residuals. In this section, the inadequacies of the addends addressing model complexity are tackled. The most widely used versions of AIC and BIC enforce parsimony by penalising the models with a higher number of parameters. This form of penalty, equating complexity with the number of parameters, does not have any sound theoretical justification. Indeed, the number of parameters k in an equation is a rudimentary indicator of its complexity and therefore cannot serve as a very satisfactory measure to avoid overfitting (Vapnik 2000). Two much more conceptually sound quantifiers of complexity have been developed: the Rademacher dimension and the VC dimension. Unfortunately, despite their theoretical pedigree, exact formulas have not been found for most types of equations; moreover approximate estimates are quite complicated to implement and are very heavy computationally (Bousquet 2016; Karpinsky 1997). These are the main reasons why the Rademacher and VC dimensions have not been adopted in many practical applications. A second subject of criticism relates to the implicit assumption that always simpler models are preferable, even when modelling complex systems. There is indeed no general theoretical justification why a complex model, with more parameters, should always be "a priori" less adequate to interpret the data generated by a complex system. A better conceptualisation of model complexity would be very advantageous.

The nature of the first issue, that the number of parameters can be a misleading quantifier of complexity, can be appreciated by the following example, in which synthetic data has been generated with a polynomial of order five, as shown in Fig. 2 (Vapnik 2000):

$$y = -10 + x + 0.03x^2 - 810^{-3}x^3 + 110^{-6}x^5 \tag{14}$$

The important observation is that it is possible to fit the same data with a high frequency sinusoid, as illustrate graphically in the second and third plots in Fig. 2. In presence of

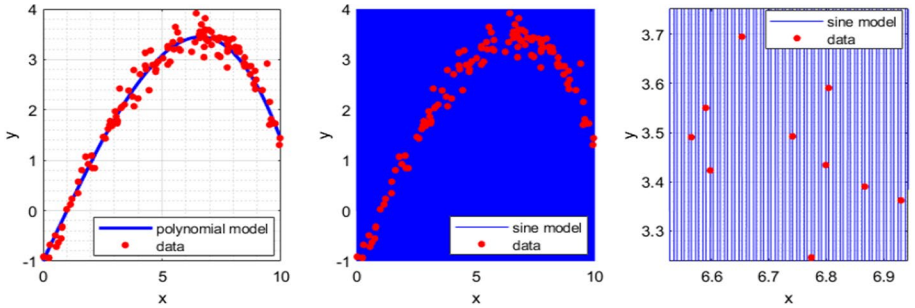

Fig. 2 Left: data generated by a 5-order polynomial, namely Eq. (14). Centre: fit with a high frequency sinusoidal function. Right: zoom of the plot in the centre to show the quality of the fit of the sinusoid

even a very small amount of additive noise, by tuning the three parameters of the sinusoid (amplitude, frequency and phase) it is possible to achieve a better fit to the data, even if it is generated by Eq. (14). Since the sinusoid can be fine-tuned by acting on only half the number of parameters required to adjust the polynomial, the original versions of AIC and BIC would always wrongly select the sinusoid; indeed both terms of the criteria would be lower for the sinusoidal model.

The second comment is motivated by a critical appraisal of the Occam Razor. Indeed, a dogmatic acceptance of the principle that simpler models are always to be selected becomes doubtful when dealing with systems of high complexity. Of course, there are perfectly valid reasons, to be careful with adopting models with an excessive number of parameters. However also oversimplification can become a significant issue, as expressed elegantly by the quote attribute to Einstein that everything should be made as simple as possible, but not simpler. A quite striking example of oversimplification is the modelling of all scaling laws with power law monomials, based on the uncritical assumption of scale invariance of the underlying phenomena. Even if power laws have proved to be very useful in many fields, the inadequacies of models relying on this type of equations when not justified, have been extensively documented (Murari 2012). Moreover, it is worth remembering that simplicity and its dual concept of complexity are subjective to a large extent, to the point that it has proved impossible to converge on a fully general definition of either. In agreement with the line of reasoning informing this work, this lack of consensus is considered a consequence of the fact that these concepts are so vague and universal that must be adjusted to the specific application and its objectives.

The previous considerations are qualitative and to a certain extent debatable but should be sufficient to substantiate the stance that a unique fit-all definition of complexity is not realistic, and that different interpretations of the complexity are fully legitimate and should take into account the nature and the objectives of the specific data analysis task. In the following subsections, three alternative approaches to interpreting and quantifying complexity are proposed, which attempt to address the issues just discussed. First a falsification criterion, which in a certain sense equates simplicity with stability, is introduced; it favours models, whose predictions are less affected by errors in their parameters. The second criterion is explicitly designed to reduce the probability of overfitting, by penalising models' flexibility. The third one is a purely information theoretic concept, which renders the AIC criterion more coherent. Again, these improvements are all objective, in the sense that they are rooted in unique properties of the models to be assessed.

4.1 A falsification approach to model selection

The improvements introduced in this section are motivated by the consideration that model robustness is an important aspect to favour. According to this position, other things being equal, a model is to be preferred when small unavoidable errors in its parameters have limited consequences on its estimates. Consequently, the desideratum of parsimony should translate into criteria, which penalise the repercussions on a model's final estimates of small errors in its parameters. This approach is particularly suited to a very important application of model selection criteria: the design of new plants and experiments. In that perspective, an essential property of the candidate models is their out of sample validity. They are indeed developed to understand the scaling properties of systems and phenomena (Murari 2015; Murari et al. 2019; Murari et al. 2020). Robust criteria, whose predictions do not change substantially as a consequence of small errors in their parameters, are particularly desirable for this task.

The proposed approach can be implemented by exploiting the available knowledge about the uncertainties in the candidate models' parameters. The procedure can be summarised as follows:

- Generate a sufficiently large number of parameter combinations for each model, sampling randomly the probability distribution function of their uncertainties.
- Calculate the model predictions for all these parameter combinations.
- Devise and compute a suitable estimator quantifying how much the model predictions vary with the uncertainties in its parameters.

To estimate the crucial aspect of each model, its prediction stability, several indicators are equally valid: mean, standard deviation, max value etc. The analyst is free to choose the most suited to the application. For example, if the worst-case scenario can have particularly negative consequences, an appropriate choice could be the maximum variation in the predictions. With regard to the uncertainties in the model parameters and their pdf, the choice must of course be guided by the objectives of the application and by the knowledge of the type of errors affecting the model estimates. In any case, indicating the estimator of the model's parameter stability with the acronym PS, the proposed version of the criteria can be written as:

$$AIC_{PS} = n\ln(MSE) + 2k + n\ln(PS) \tag{15}$$

$$BIC_{PS} = n\ln(MSE) + kln(n) + n\ln(PS) \tag{16}$$

For the delicate case of the sinusoid discussed in Sect. 4, assuming that the parameters are affected by zero mean Gaussian noise, an appropriate metric is certainly the MSE. In these hypotheses, the proposed improved versions of the indicators (15) and (16) achieve noticeable better performances, as can be seen by simple inspection of Table1.

4.2 Quantifying model flexibility (MF)

This subsection is based on an interpretation of complexity, which is meant to address a major objective in machine learning and model building: avoiding overfitting. To his end, a simple practical solution, to counteract the tendency of a model to overfitting, consists of quantifying its flexibility in the region covered by the independent

Table 1 Estimates of the various versions of the indicator for the problematic example of the comparison between a polynomial and a sinusoid introduced in Sect. 4

	BIC	AIC	BIC_{PS}	AIC_{PS}	BIC_{MF}	AIC_{MF}
Model 1	− 925	− 945	− 1132	− 1164	− 925	− 945
Model 2	− 939	− 949	− 913	− 949	− 721	− 730
Model selected	2	2	1	1	1	1
Result	Wrong	Wrong	Correct	Correct	Correct	Correct

Model 1 is the polynomial, model 2 the sinusoid

variables in the available database. A suitable indicator of a function flexibility is the moving average of its standard deviation (Murari 2021). Such an indicator, which is called Model Flexibility (MF) in the following, can be easily estimated by calculating the moving average of the model and then summing the squares, therefore implementing the equations:

$$MovSTD_{y,x}(x_i) = \sqrt{\frac{\sum_{j=i-\Delta}^{i+\Delta}\left(\frac{df}{dx} - \overline{\frac{df}{dx}}\right)^2}{2\Delta}} \tag{17}$$

$$MF = \frac{\sum_{i=1}^{N} MovSTD_{y,x}^2(x_i)}{N} \tag{18}$$

where N is a set of synthetic points, to be chosen as discussed later. In the interpretation of complexity, quantified by Eqs. (17) and (18), model A is to be considered more complex than model B if its derivatives present a higher variation in the interval covered by the independent variables $\boldsymbol{X}$. A quite straightforward and effective version of the AIC and BIC, to take into account the metric expressed by the MF, is:

$$AIC_{MF} = \ln(MSE) + 2k + \ln(1 + MF) \tag{19}$$

$$BIC_{MF} = \ln\left(\sigma_{(\epsilon)}{}^2\right) + kln(n) + \ln(1 + MF) \tag{20}$$

The interpretation of the last two equations becomes even more intuitive if they are rewritten as:

$$AIC_{MF} = \ln(MSE) + 2k + \ln(1 + MF) = \ln(MSE(1 + MF)) + 2k \tag{21}$$

$$BIC_{MF} = \ln\left(\sigma_{(\epsilon)}{}^2\right) + kln(n) + \ln(1 + MF) = \ln\left(\sigma_{(\epsilon)}{}^2(1 + MF)\right) + kln(n) \tag{22}$$

An important point to appreciate is that the MF indicator cannot be computed for the n entries of the database, otherwise it would simply reinforce the classification of the first term of the criteria, instead of counterbalancing it. Consequently, it must be calculated for synthetic points, albeit in the interval of the independent variables covered in the database. The algorithm for evaluating the new version of the indicators AIC_{MF} and BIC_{MF} can therefore be summarised as follows:

- Generate a suitably high number N of independent variable points in the domain covered by the independent variables (different from those in the original DB),
- Calculate the predictions of the models for these additional points
- Compute the MF indicator to be included in the AIC_{MF} and BIC_{MF}

The routine contains only two free parameters: the number N of generated synthetic points and the interval Δ over which to calculate the moving average. A simple approach to determine N consists of progressively increasing its value, until the indicators stabilise to a constant output. To achieve sound convergence, typically a multiple of the entries in the database (between 3 and 10 depending on the problem) must be generated. Also the parameter Δ has to be optimised empirically depending on the application. In the authors' experience, a quite safe value is $\Delta = \sqrt[3]{N}$. In any case, typically the results of AIC_{MF} and BIC_{MF} do not vary substantially with the choice of Δ and they provide the same classification for a quite wide range of this parameter value. Again a direct application to the delicate example of the sinusoid reported in Sect. 4 shows that the new proposed versions of the indicators have not negligible better performances than the traditional AIC and BIC, as can be deduced again by inspecting Table 1.

4.3 An information theoretic quantifier of complexity for AIC

Information theoretic estimates of complexity have been developed and are covered quite well in the literature (Meyers 2009; Mitchell 2009). In the context of the present work, they are particularly relevant because they can be formulated as described in the present subsection and then deployed, to obtain a version of the AIC criterion expressed completely in terms of information theoretic quantities. The proposed fully information theoretic version of AIC relies on considering complexity a sort of middle ground between randomness and determinism. This interpretation of complexity is not new. It has a long pedigree and can be traced back to the founding concept of information theory, i.e. the understanding of information as reduction of uncertainty (Piqueira 2018). The complexity measure *C[X]*, discussed in the following, is a natural translation of this idea in mathematical terms:

$$C[X] = H[X]\, D[X] \tag{23}$$

Where H is the usual Shannon entropy [31] and D the distance from a uniform distribution:

$$D[X] = \sum_{1}^{n} \left(p_i - \frac{1}{n} \right) \tag{24}$$

The implicit assumption behind the last two relations is that not only a pure deterministic system but also a uniform distribution is indicative of a lack of complexity. Indeed, in Eq. (23) the increases of entropy H for more uniform distributions is compensated by the D term, which tends to zero when the probability of all the elements becomes more similar. The models can be evaluated over a suitable interval and with enough resolution to produce a sufficient number of points, to reliably calculate probability distribution functions. The pdf of the models is all that is required to compute (23) and to obtain a simple indicator of complexity, according to the aforementioned information theoretic interpretation. Plugged into Eq. (12), this formulation of the complexity term provides a coherent version

of AIC, in the sense that the criterion depends only on information theoretic quantities (except the MSE of course).

It is possible to generalise the expression of complexity implemented by Eq. (23) (Piqueira 2018):

$$\mathrm{COMP}\,\alpha,\beta \;=\; \Delta^{\alpha}(1-\Delta)^{\beta} \tag{25}$$

where

$$\Delta = \frac{H[X]}{H_{max}[X]} \tag{26}$$

where again H is the Shannon entropy and $H_{max}[X]$ its maximum value log n. With this more flexible formulation, the analyst has much more freedom to give different weights to the two terms in the definition of complexity. However, the main drawback of Eqs. (25) and (26) resides in the fact that there is no principled procedure to select the α and β exponents in an objective way. This version of the criterion is therefore excessively subjective and therefore, unless very sound prior information is available to inform the choice of α and β, the use of Eq. (25) is not recommended.

5 The utility-based approach applied to the probability distribution of the data

The ameliorations of the MSC discussed so far have concentrated on two main aspects; on a better quantification of the residuals' statistical properties and on alternative views of the complexity term. This section is devoted to a different series of techniques, which act directly on the pdf of the data. The motivation behind an objective utility-based approach at this level is at least twofold. On the one hand, certain parts of the pdf could be inherently of more relevance than others for the investigation of the problem to be studied (see Sect. 5.1); on the other hand, the data could be affected by outliers, which should be eliminated even if no detailed information is available about their characteristics (see Sect. 5.2). The solutions proposed in this section are not in contrast with the ones described previously; on the contrary, they can be deployed preliminary to the application of the versions of the criteria presented in the Sect.s 3 and 4. They can help cleaning and optimising the databases, to which the MSC criteria can then be more profitably applied. Indeed, no matter the power of the criteria, the results of the analysis cannot exceed the quality of the input data and refining the databases can have a strong impact on the final selection.

5.1 Weighting the various parts of the pdf

All the enhancements of MSC, described in the previous sections, take the pdf of the data as given. The underlying assumption behind this position is that the user attributes equal value to all the parts of the data distribution function. This is another piece of conventional wisdom, which makes calculations convenient but is not realistic in most real-life tasks. In many branches of the physical sciences, for example, the noise sources affecting different measurements are assumed to be completely independent from the system under investigation (and from one another), which can be clearly unfounded. A paradigmatic example is constituted by thermonuclear fusion plasmas, whose diagnostics are affected by enormous

electromagnetic compatibility issues, since instruments measuring mT are all immersed in huge fields of the order of several Teslas (Wesson 2004). Therefore, assuming that the various measurements are affected by independent sources of noise is clearly a bit of a stretch. Another example, this time from economics, relates to the subprime financial crisis, not predicted by models, which did not take into account the important systemic effects of rare but highly correlated events. In the first case, thermonuclear fusion, it would be appropriate to get rid of the additional correlation between signals due to the noise; in the second, macroeconomics, the mutual correlation introduced by rare events should be strengthened.

The previous short discussion emphasises one more time the context specific nature of model selection and therefore the important of utility-based techniques. In reality, a context dependent but objective approach can be easily applied to the probability distribution of the original data, by weighting different parts according to the needs of the analysis and the characteristics of the problem. Consequently, the MSE and standard deviation in the original versions of the AIC and BIC criteria should be appropriately modified.

All the information theoretic elements, of the previously proposed versions of the criteria, can be easily adapted to take into account the different relevance of various parts of the pdfs. The entropy, for example, can easily be rewritten introducing utility-based weights w_i:

$$H^w(X) = H^w(w,p) = \sum_{i=1}^{n} w_{(i)} p_i \, log\left(\frac{1}{p}\right) \tag{27}$$

The weighted entropy H^w just defined preserves most of the useful properties of the original entropy. Particularly interesting and reassuring are the ones summarised in the following Lemma (Guiasu).

Lemma 2 *Let X be a stochastic random variable of n possible states with probability mass function* $p=(p_1,....p_n)$. *For a vector of weights* $w=(w_1,....w_n)$ *to be associated with these states and the condition* $w_i \geq 0$, $i=1....n$, *the weighted entropy* H^w *defined by (27) presents the following properties:*

1. $H.^w(X) \geq 0$
2. *If* $w_1=w_2 ....=w_n$ *then* $H.^w(X)=w\, H(X)$
3. *If* $p_i=1$ *for any* $i=1,....n$ *then* $H.^w(X)=0$
4. *For any non-negative real number l,* $H^w(l\, w)=l\, H.^w(w)$
5. $H^w(w_1,....w_n, w_{n+1}; p_1,....p_n, 0)=H^w(w_1,....w_n, w_{n+1}; p_1,....p_n)=H.^w(X)$ *for any* w_{n+1}

The properties of the weighted entropy apply only to the discrete case. In the continuous case there are difficulties discussed in (Kelbert, Stuhl, and Suhov, 2017). On the other hand, the continuous or differential version can also be implemented quite easily, provided adequate precautions are taken to double-check the consistency of the results.

The weighting can be extended also to other indicators, derived from the entropy. In the present context, the mutual information is a particularly useful quantity to refine taking into account the considerations of SubSect. 3.3. The weighted version of the mutual information MI^w can be written as:

$$MI^w = \sum \sum w_{ij} P_{xy} \frac{P_{xy}}{P_x P_y} \tag{28}$$

where P_{xy} is the joint probability distribution function of the two stochastic variables x and y. The main mathematical properties of the mutual information are preserved, rendering evident the benefits of implementing MI^w, provided of course an appropriate choice of the weights can be devised.

The just described weighting is intuitive and clear. The properties of the resulting information theoretic quantities are well understood in mathematical terms (Guiasu) but this is not always the most immediate nor even the most appropriate solution to adopt in practice. Indeed, in most applications, what is known is the quality of the individual measurements; how this information translates into the quality of the residuals pdf is not necessarily possible or simple to determine. Moreover, in practice it is not uncommon for data in the same parts of the pdf to present completely different quality. Applying the same weight to all of them would therefore not be justified and could lead to important distortions of the results. In this quite common situation, a more reliable alternative, also easier to implement, would therefore consist of weighting individual residuals not their pdf. Such as solution requires calculating the necessary pdfs on the basis of the weighted residuals; this can be achieved with the following relations:

$$p_x(x_b) = \frac{J_x(x_b)}{\sum w_i} where$$

$$J_x(x_b) = \sum w_i \, if \, x_b \le x_i < x_{b+1}$$

$$J_x(x_b) = 0 \, if \, x_i < x_b \ \& \ x_i \ge x_{b+1}$$

$$p_y(y_b) = \frac{J_y(y_b)}{\sum w_i} where$$

$$J_y(y_b) = \sum w_i \, if \, y_b \le y_i < y_{b+1}$$

$$J_y(y_b) = 0 \, if \, y_i < y_b \, \& \, y_i \ge y_{b+1}$$

$$p_{xy}(x_b, y_b) = \frac{J_{xy}(x_b, y_b)}{\sum w_i} where$$

$$\begin{aligned} &J_{xy}(x_b, y_b) = \sum w_i \, if \, x_b \le x_i < x_{b+1} \ \& \ y_b \le y_i < y_{b+1} \\ &J_{xy}(x_b, y_b) = 0 \, if \, x_i < x_b \, \& \ x_i \ge x_{b+1} \, \& \, y_i < y_b \, \& \\ &y_i \ge y_{b+1} \end{aligned} \tag{29}$$

In Eqs. (29), the subscript b is the index of the various bins; therefore the points within a bin are attributed weights w_i depending on their quality and those outside have weight zero. With this approach, the pdfs can be calculated on the basis of the knowledge about the quality of the individual entries in the DB, which is sometimes the only realistic alternative, as shown in Sect. 6. The versions of AIC and BIC, in which the individual entries have been weighted, are indicated by the subscript w. To exemplify the importance of weighting

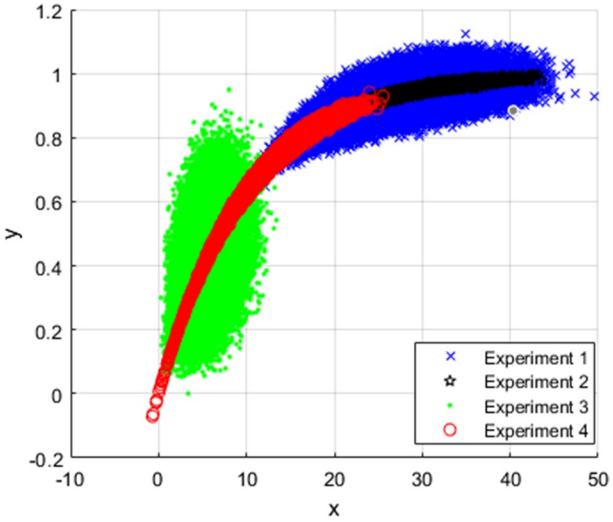

Fig. 3 Hypothetical example of data collected in different experiments and therefore of different quality

Table 2 The values of the indicator for the case of data generated with Eq. (29)

	BIC	AIC	BIC_{MI}	AIC_{MI}	$BIC_{MI,W}$	$AIC_{MI,W}$
Correct model	− 5.35e6	− 5.34e6	− 5.14e6	− 5.14e6	− 5.21e6	− 5.21e6
Alternative model	− 5.52e6	− 5.53e6	− 5.12e6	− 5.12e6	− 5.08e6	− 5.08e6
Decision	Wrong	Wrong	Correct but marginal	Correct but marginal	Correct	Correct

the individual points, let us suppose that the phenomenon under study is investigated with measurement systems producing data according to the function:

$$y = 1 - e^{-0.1x} \tag{30}$$

Let us suppose that the model alternative to (30) is a nine order polynomial obtained by fitting the DB entries with the MATLAB function polyfit. It is also assumed that the data, as not unusual in practice, have been collected in different experiments of different quality, as illustrated graphically in Fig. 3. Determining where the residuals of each experiments affect the various parts of the pdf is tricky; the results are too dependent on the choice of the weights and therefore the confidence in the conclusions would be very weak. The best way is to weight the individual points proportionally to the inverse of the noise standard deviation. The MI after weighting the points is a factor of two better than the one without weights. Consequently, the AIC_w and BIC_w manage to identify the right model, whereas the traditional versions fail and the ones using the MI without weights provide poorly discriminatory results, as shown in Table 2.

5.2 Robust statistics

In some applications, a fundamental objective of acting on the pdfs would consist of eliminating outliers and spurious cases, in order to increase the realism of the results. This is the realm of robust statistics (Huber 1981; Hettmansperger. and McKean 1998). Indeed, classic summary statistics and significance tests are based on certain

specific assumptions, which have to be reasonably satisfied. On the contrary, the probability density functions sampled in experiments are not necessarily Gaussians and can present heavy tails or be skewed. Homoscedasticity is even less frequently verified. If these hypotheses are violated, even slightly, the accuracy of the results can be seriously compromised. In the last decades, a lot of evidence has emerged, showing that a blind reliance on the previously mentioned assumptions of Gaussianity and homoscedasticity can produce rather inaccurate results (Wilcox 2012). Consequently, significant efforts have been recently devoted to developing robust tools, with a dual objective. On the one hand, they seek to provide methods that compare well with popular statistical techniques, when the classic hypotheses are satisfied. On the other hand, they are designed not to be unduly affected by departures from the model assumptions (Huber 1981) (Farcomeni 2013). These techniques can be very useful in improving the goodness of fit tests, when the general and typically unrealistic assumptions of equal mean and homoscedasticity are violated. Specific metrics can also be introduced to counteract the effects of outliers and noise of various statistics (Wilcox 2012). To provide a flavour of how to deal with outliers, in the rest of this subsection robust criteria are briefly discussed for both the measures of location and scale, the two most relevant for the present subject (since they are the two measures entering in the model selection criteria) (Rouseeuw 2011).

The most common robust statistical measures of central tendency are the trimmed mean and the winsorised mean. A trimmed mean or truncated mean is obtained by calculating the mean of the available data, once the high- and low-end parts of the samples have been discarded. The number of discarded entries is usually given as a percentage of the total number of samples and is applied symmetrically to the two ends of the range. For most statistical applications, 5 to 25 percent of the ends are discarded; the 25% trimmed mean (when the lowest 25% and the highest 25% of the data are discarded) is known as the interquartile mean. The winsorized mean is calculated replacing given parts of a probability distribution, at the high and low end, with the most extreme remaining values. In detail, the traditional mean, the trimmed and winsorized means are calculated according to the following formulas, in which f(xi) indicate the values sampled from the data pdfs.

The traditional mean is defined as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} f\left(x_i\right) \tag{31}$$

where n is the number of available samples. The trimmed mean is defined as:

$$\mu_t = \frac{1}{N}\sum_{i=g+1}^{n-g} f\left(x_i\right) \tag{32}$$

where g corresponds to the number of trimmed points. The winsorized mean is defined as:

$$\mu_w = \frac{1}{n}\sum_{i=1}^{n} f_w\left(x_i\right) \tag{33}$$

where

$$f_w(x_i) = \begin{cases} f(x_{g+1}) \, iff \, (x_i) \le f(x_{g+1}) \\ f(x_i) \, iff \, (x_{g+1}) < f(x_i) < f(x_{N-g}) \\ f(x_{N-g}) \, iff \, (x_i) \ge f(x_{N-g}) \end{cases}$$

The robust statistical methods developed in the last decades allow improving not only the estimates of location but also those of scale (see [37]). As a reference, the classic standard deviation is defined as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \left[f(x_i) - \mu \right]^2}{n-1}} \tag{34}$$

In the case the location is calculated with the trimmed mean, the appropriate version of the standard deviation σ_t to implement is:

$$\sigma_t = \sqrt{\frac{1}{n(n-1)(1-2\gamma)^2} \sum_{i=1}^{n} \left(f_t(x_i) - \mu_t \right)^2} \tag{35}$$

where n, γ, $f_t\,(x_i)$ and μ_t are the number of points sampled from the pdf, the percentage of trimming, the trimmed data and the trimmed mean respectively.

A similar definition applies to the standard deviation of the winsorized mean:

$$\sigma_w = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} \left(f_w(x_i) - \mu_w \right)^2} \tag{36}$$

Obviously, the robust versions of the location and scale measures are more resilient against the presence of outliers. Their adoption in the calculation of the model selection criteria can improve their discriminatory capability and reliability, if outliers are even a quite limited fraction of the data. The robust versions of AIC and BIC are indicated by $\mathrm{AIC_{Rob}}$ and $\mathrm{BIC_{Rob}}$.

6 Results of numerical tests

To investigate and quantify the performance of the alternative formulations of the model selection criteria, a series of systematic tests has been performed. Some didactic cases have already been shown. In this section, representative results of more general sets of examples are reported. As mentioned in Sect. 2, since the focus of the present work is model selection for inference, the analysis has been framed in the parametric context, in which the equation generating the data is included in the set of candidate models. Section 6.1 summarises the main classes of functions and noise statistics considered. Section 6.2 provides some of the most representative results obtained for cross sectional data. In Sect. 6.3, applications to time series, including systems with feedback loops, are exemplified.

6.1 The main model classes and noise statistics of relevance for scientific and engineering applications

Regarding cross sectional data, the analysis has been focused mainly on three classes of models. They are the exponential functions, polynomials and power laws. For clarity of exposition, in the following only the results for bidimensional functions ($z=f(x,\ y)$) are discussed, because they can be easily visualized. The extension to higher dimensionality is obvious even if in practice of course the requirements, in terms of quantity and quality of the data, can become severe.

The exponential functions investigated in this paper have the form:

$$z(x,y) = axe^{(bx+cy)} + dxe^{(ex+fy)} + g \tag{37}$$

Polynomials are mathematical expressions that contain two or more algebraic terms, which can be added, subtracted, or multiplied (no division allowed!). In general, polynomial expressions include at least one variable and typically also constants and positive exponents. Polynomial functions have the following form:

$$z(x,y) = p_{00} + p_{10}x + p_{20}x^2 + p_{01}y + p_{02}y^2 + p_{03}y^3 + p_{11}x + p_{21}x^2y + p_{12}xy^2 \tag{38}$$

The power laws considered in this paper are monomials of the form:

$$z(x,y) = cx^a y^b \tag{39}$$

where the exponents can be either positive or negative. With regard to the noise statistics, four of the most relevant distribution functions have been tested: Gaussian, uniform, Poisson and gamma distributed noise.

From a mathematical point of view, time series are sequences of data indexed (or listed) in time order. Also in the case of time series, the most widespread used types of functions have been considered, such as combinations of sines and cosines. A very important a delicate class of equations is the one of autoregressive models. A generalized auto-regressive model of order n can be represented as:

$$\begin{aligned}\hat{y}(t) = f(&\mathrm{y}(t-1), \mathrm{y}(t-2), \ldots\ldots, x_1(t-1), x_1(t-2)\\ &, \ldots, x_2(t-1), x_2(t-2) \ldots ..., x_n(t-1), x_n(t-2), ...)\end{aligned} \tag{40}$$

where $x_1, ..., x_n$ are the independent time series that are thought to influence y, t is time and $x_k(t-m)$ is the kth independent time series time-shifted of m temporal lags.

6.2 Overview of the main results of the numerical tests for cross sectional data

With regard to the improvement proposed to the first term of the AIC and BIC, the positive effects of introducing the entropy (SubSect. 3.1) have already been extensively documented (Murari 2019). As mentioned, the introduction of the Z_{score} (SubSect. 3.2) becomes very useful if the statistics of the noise is known and is not Gaussian. Various examples of the very relevant improvements, which can be achieved implementing this version of the indicators, are reported in Table 3 for a variety of functional dependencies and noise statistics. The models, used to generate the synthetic data, are reported in the second column. The alternative models have been obtained by fitting the data with the same type of function using the MATLAB routine nlinfit. This is typically an extremely severe type of test and the systematic improvements obtained with

Table 3 Results of introducing the score in the indicators as proposed in subSect. 3.2. The cells in light blue indicate the cases, in which the BICGF and AICGF outperform the traditional versions of the indicators

Case	Function	Noise type	Model	BIC	AIC	BIC_{GF}	AIC_{GF}
1	*Exponential* $z(x,y) = 100(xe^{-x} + ye^{-y})$	*Gaussian*	Correct model	0	0	-7.5e4	-7.2e4
			Alternative model	0	0	-7.5e4	-7.2e4
2		*Gamma*	Correct model	1.0e5	1.0e5	0	0
			Alternative model	1.0e5	1.0e5	3.0e5	3.5e5
3		*Poisson*	Correct model	3.0e5	3.0e5	1.0e5	1.0e5
			Alternative model	3.0e5	3.0e5	5.0e6	5.0e6
4	*Polynomial* $z(x,y) = 1 + x^2 + 3y^4$	*Gaussian*	Correct model	2.0e5	2.0e5	8e4	8.0e4
			Alternative model	2.0e5	2.0e5	8e4	8.0e4
5		*Gamma*	Correct model	1.0e5	1.0e5	0	0
			Alternative model	1.0e5	1.0e5	3.0e5	3.0e5
6		*Poisson*	Correct model	0	0	0	0
			Alternative model	0	0	4.0e5	4.0e5
7	*Power law* $z(x,y) = 3x^3y^{-1}$	*Gaussian*	Correct model	3.0e5	3.0e5	7.5e5	7.5e5
			Alternative model	3.0e5	3.0e5	7.5e5	7.5e5
8		*Gamma*	Correct model	1.0e5	1.0e5	0	0
			Alternative model	1.0e5	1.0e5	3.0e5	3.0e5
9		*Poisson*	Correct model	5.0e4	5.0e4	5.0e4	5.0e4
			Alternative model	5.0e4	5.0e4	6.0e5	6.0e5

the proposed refinements of the criteria are therefore very encouraging. To interpret the entries of Table 3, one should consider that for the functions analysed the dominant term in the criteria is the goodness of fit, which is the same in both AIC ad BIC. This is the reason why the indicators give the same numerical results. The improved AIC_{GF} and BIC_{GF} outperform the traditional versions of the indicators except for the case of Gaussian noise. Such a result is to be expected since for this type of noise also the alternative models converge on the right ones; this case has been indeed reported to show that the modified version of the criteria provide the same results as the AIC and BIC, when the hypotheses, under which the original versions have been derived, are satisfied.

In the case the statistics of the noise is unknown, the traditional versions of the AIC and BIC can in any case be improved by considering the mutual information between the model estimates and the residuals, as described in SubSect. 3.3. Table 4 reports a representative and quite challenging example to show the improvements, which can be achieved with this approach. The results, reported graphically in Fig. 4, show clearly how the discriminatory capability of the new version of the indicators is remarkably better than the one of the original AIC and BIC over the entire range of entries. Indeed,

Table 4 Examples to test the effects of introducing the MI in the indicators as proposed in SubSect. 3.3. Ref is the model used to generate the data

#	Model	k
1	$\frac{0.204x_2}{\sin\left(x_1\left(\frac{0.46}{x_1^{8.72}}+\frac{0.61}{x_2}\right)\right)}$	8
2	$0.258\left(x_3^{3.08} - x_3\right) - 0.03\sin\left(x_3^{-12.62}\right)$	6
3	$31.23\left(x_1^{2.21} - \sin\left(x_2\right)\right)$	5
4	$50 + 10.45x_1x_3\sin\left(1.07x_3\right)$	6
Reference	$2.x_1^{2.5}x_2^{-0.75}x_3^{2.5}$	4

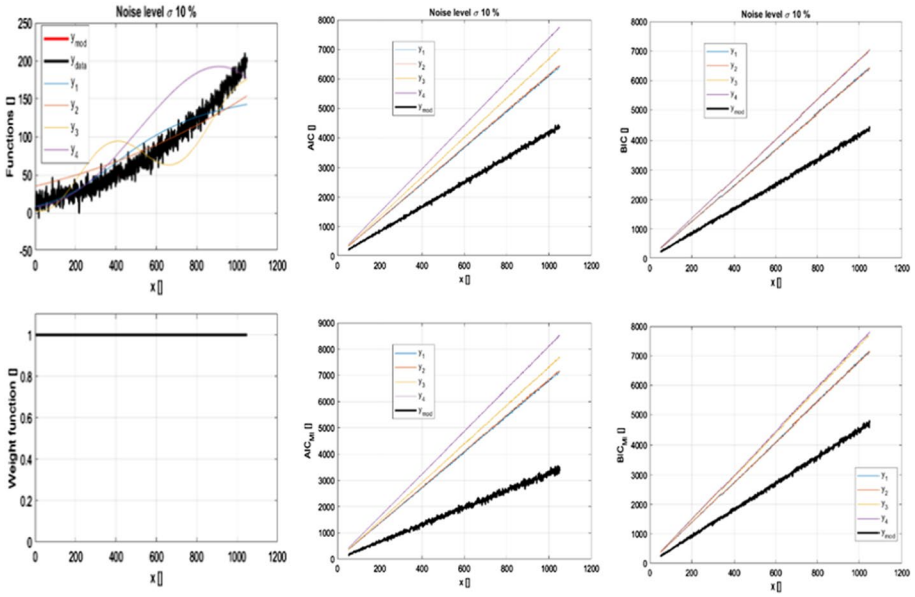

Fig. 4 Results for the case shown in Table 4. Left column: the tested functions, in black the one generating the data plus 10% of additive Gaussian noise. Central column: comparison between AIC and AIC_{MI} vs the number of entries. Right column: comparison between BIC and BIC_{MI} vs the number of entries

the gap between the right model and the others is much larger for AIC_{MI} and BIC_{MI} than for the traditional AIC and BIC.

Referring to the improvements of the complexity term in the AIC and BIC, the results reported in Table 1 for a very challenging case illustrate the potential of the proposed indicators. It is typically quite difficult to find a case, for which the alternatives described in Sect. 4 do not provide a very noticeable improvement in the discriminatory capability of the original criteria. The same can be said even more strongly for the methods acting directly on the pdfs of the data, which are very powerful and produce practically always an appreciable improvement in the results, when the traditional versions of the indicators are in difficulty for reasons due to issues related to the data distributions. It has not possible to devise situations, in which reducing the relevance of bad parts of the pdfs decreased the quality of the final classification. Of course, the weights have to be properly chosen, in an objective and sound way.

The advantage of weighting the individual points, and not parts of the pdf, is not only implementation convenience but can be substantial as in the case reported in SubSect. 5.1. With regard to the robust statistics versions, even if somehow less powerful than weighting the entries of the DB, they are much less delicate to apply, since they tend to reproduce the results of the AIC and BIC in absence of outliers. It is therefore good practice, if appropriate, to always implement the techniques suggested in Sect. 5 preliminary to the application of the other versions of the criteria. An important example, involving a system of equations with feedback, is reported in Sect. 7. To conclude, it should be mentioned that the results detailed for the AIC and BIC criteria are valid also for the other members of their families. It has indeed been checked that the modifications proposed in this work have the same effects also on the other information theoretic and Bayesian criteria derived from the original AIC and BIC, such as those mentioned in Sect. 1.

6.3 Overview of the main results of the numerical tests for time series including feedback loops

Time series are a very important class of signals. Indeed, they are often the natural output of experiments in physic and engineering. Nowadays, time indexed data have become available also in many other disciplines, ranging from economics to medicine and the earth sciences. Time series analysis comprises methods for extracting meaningful statistics from the data. Models are used in time series forecasting to predict the future evolution of time series based on previously observed values.

To exemplify the potential of the developed enhancements, a synthetic time-series database has been generated with the model:

$$y = 5 \cdot e^{-t} + 0.9 \cdot t \cdot \sin(t) + 1 + \epsilon(t) \tag{41}$$

where t is the time variable and ϵ represents a noise term of the following form:

$$\in (t) = \begin{cases} \sigma \cdot \gamma(a,b); t \in (1s, 5s) \\ -\sigma \cdot \gamma(a,b); t \in (5s, 10s) \end{cases}$$

where $\gamma(a,b)$ is a random gamma noise with shape parameter a, scale parameter b and amplitude σ. The analysis has been carried out for $t \in [1s, 10s]$, $dt = 0.001$, $a = b = 0.5$ and $\sigma = [0.01, 0.05, 0.2, 0.5, 1]$. In Fig. 5, a visual representation of the synthetic signals generated to carry out the analysis is reported. The competing model is a 9th-degree polynomial.

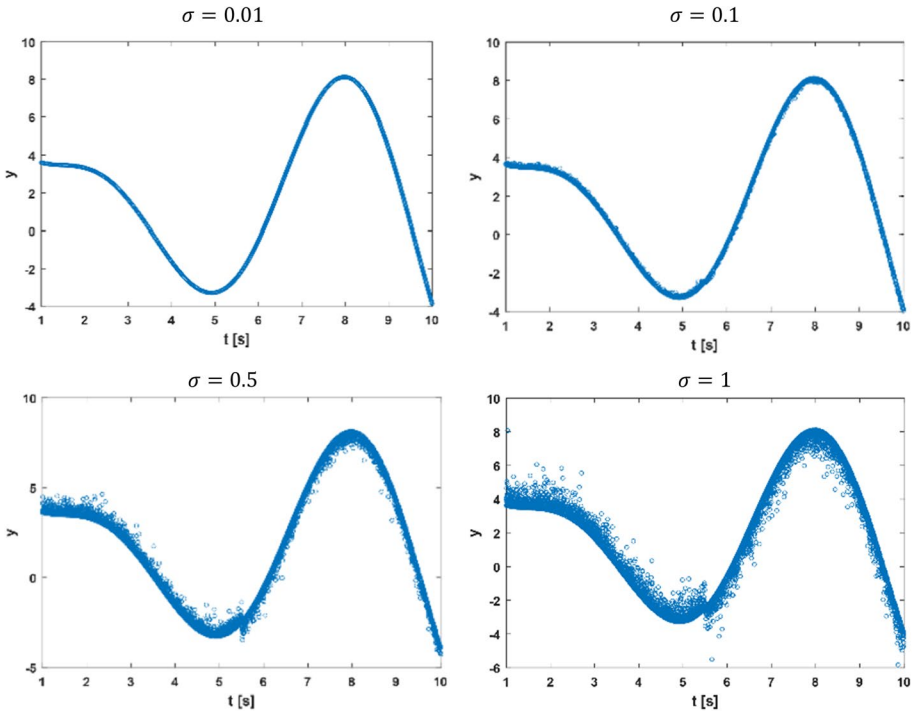

Fig. 5 Examples of signal realisations generated with Eq. (41) for different levels of noise

The model selection performance of the standard versions of AIC and BIC have then been compared with the proposed version AIC_{GF}, BIC_{GF}. The performance metric used is the number of times the model selection criteria succeed in finding the correct model over 100 noise instances. The statistical results are reported in Table 5. When the amplitude of the noise increases to realistic values, the superiority of the BIC_{GF}, AIC_{GF} over the standard version is clearly evident.

Among time series, a category of models very difficult to identify are typically those containing feedback loops. Feedback takes place when some outputs of a system are routed back as inputs. The resulting chain of cause-and-effect forms a circuit called a feedback loop. The concept of cause-and-effect is problematic when applied to systems containing feedback, because causality becomes circular. Indeed, even in the basic case of only two subsystems, simple causal reasoning is delicate because the first system influences the second and second system influences the first. Reductionist approaches to the analysis typically break down and it is necessary to investigate the system as a whole (Åström and Murray 2008).

Given the inherent difficulties of modelling systems with feedback, the tools proposed in the present work can present quite competitive advantages. A representative example of application is constituted by the Lotka-Volterra equations, which are also known as the predator–prey equations (Bomze 1983). The Lotka-Volterra system comprises a pair of first-order nonlinear differential equations (Turchin 2003). They are often utilised to model the dynamics of biological systems, in which two species interact, one as a predator and the other as prey. The two populations are evolved through time according to the equations:

$$\frac{dx}{dt} = \alpha x - \beta xy \tag{42}$$

$$\frac{dy}{dt} = \delta xy - \gamma y \tag{43}$$

where x is the number of prey individuals, y is the number of some predator species and t represents time. The interaction of the two species is governed by the parameters α, β, γ and δ. This set of equations has been chosen not only for its difficulty but also for its generality. Indeed, the system of Eqs. (42) and (43) has the same mathematical form as the law of mass action for two chemical elements of concentrations [A] and [B]:

$$\frac{d[A]}{dt} = k_1[A] - k_2[A][B] \tag{44}$$

$$\frac{d[B]}{dt} = k_3[B] - k_4[A][B] \tag{45}$$

Table 5 Results of the tests performed for the case of synthetic data generated with Eq. (41). the table reports the number of successful identifications out of 100 different noise realisations

σ	BIC	BIC_{GF}	AIC	AIC_{GF}
0.01	100	100	100	100
0.1	0	59	0	59
0.2	0	64	0	63
0.5	0	68	0	68
1	0	64	0	64

The Lotka-Volterra system of ordinary differential equations has been solved in MATLAB using the ode45 function based on an explicit Runge–Kutta (4,5) formula. The initial conditions for the numerical case reported are $x_0 = y_0 = 2$, and the solutions have been evaluated between 0 and 5 s for 1000 points. The chosen model's coefficient are $\alpha = 2/3, \beta = 4/3, \delta = \gamma = 1$. A normal random noise, with mean $\mu = 0$ and standard deviation $\sigma = \sigma_{perc} * mean(mean(x), mean(y))$, has been added to the solution obtaining two noised signal x_{noise}, y_{noise}. The analysis has been carried out for $\sigma_{perc} = 0.1, 0.2, 0.3$. A visual representation of an instance of the generated signals with $\sigma_{perc} = 0.2$ is reported in Fig. 6.

One of the most severe tests has been the comparison of the correct model, the one we used to generate the data, with one slightly more complex but much more flexible. This second model is reported below:

$$\frac{dx}{dt} = \alpha x - \beta xy - \zeta x^2 \tag{46}$$

$$\frac{dy}{dt} = \delta xy - \gamma y \tag{47}$$

The number of parameters of the models has been used as a measure of complexity. The correct model has complexity equal to 5, while the alternative model has complexity equal to 6.

Given the well-known sensitivity of the Lotka-Volterra equations to noise (Bomze 1983), the refinements of Sect. 3 are obvious candidates for the analysis. Assuming that no information about the noise is available, the evaluation of the goodness of fit with the help of the mutual information seems particularly appropriate. The evaluation of the BIC, BIC_{MI}, AIC, AIC_{MI} of the competing models for 100 instances of the noise, has provided the results reported in Table 6.

As expected, the benefits of the proposed improvements of the criteria become more evident the higher the noise. The standard BIC is less affected for a high number of time slices but, if only a reduced number of points is available, it becomes very vulnerable as

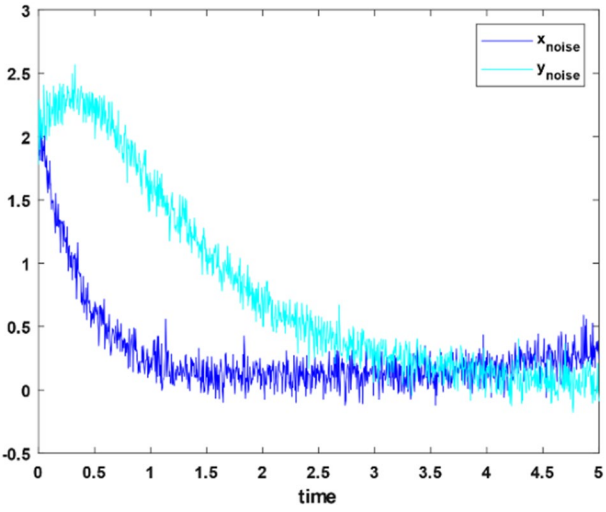

Fig. 6 Time evolution of the Lotka-Volterra equations for the choice of parameters reported in the text

Table 6 Results for the traditional version of the AIC and BIC criteria, compared with the accuracy of the improved version of the indicators described in SubSect. 3.3

σ_{perc}	BIC	BIC_{MI}	AIC	AIC_{MI}
0.1	100	100	90	99
0.2	99	100	87	97
0.3	98	100	83	96

The table reports the number of successful identifications out of 100 different noise realisations

well. In case of high disturbance levels and scarcity of examples, knowing the statistics of the noise can improve the situation, allowing to deploy the version of the criteria based on frequentist indicators, as formulated in SubSect. 3.2. Such an alternative, perfectly in line with the approach of developing utility-based context dependent indicators, can provide an important improvement with respect to the already more than satisfactory results reported in Table 6.

7 Examples of exploratory applications: genetic programming supported symbolic regression

In the last decades, complex optimization problems have been addressed with a series of techniques, grouped under the name of evolutionary computation or evolutionary algorithms (Sumathi et al. 2008). They are inspired by natural selection and have been very successful in many fields, ranging from bioinformatics and operations research to machine learning, automatic programming and even branches of the social sciences. From a technical perspective, a variety of methods are included in evolutionary computation, the main families being standard Genetic Algorithms (GA), Genetic Programming (GP), Evolution Strategies (ES) and Evolutionary Programming (EP) (Koza 1992). Motivated by the results in various disciplines, more recently the methods of evolutionary computation, and in particular genetic programming, have been applied to solve scientific and engineering problems. The so-called Genetic Programming supported Symbolic Regression (GPSR) has been deployed to determine the relationship between a dependent quantity and one or more independent regressors (also called 'predictors'). In particular, GPSR allows deriving model equations directly from the database available, without assuming a priori the mathematical form of the final equation (Murari 2019b, Murari 2020).

Genetic Programs (GPs) are designed to emulate the behaviour of living organisms and work with a population of individuals, e.g., mathematical expressions in our case, each representing a possible solution. The best individuals of each generation are selected as the parents for the application of the genetic tools (copy, mutation, cross over), to obtain the next set of candidate models, hopefully more performing than the previous ones. Contrary to traditional fitting routines, the task does not consist of identifying the parameters of a predefined class of equations and indeed no specific model is provided as a starting point. The objective is to determine the best mathematical form interpreting the data and this is achieved by combining mathematical building blocks such as constants, operators, basic analytic functions, state variables and even user defined elements (Koza 1992) (Hingston 2008).

Regarding the knowledge representation, the candidate mathematical formulas are codified as trees, which have a very high expressive capability, typically not limiting the

potential of the developed tools for scientific and engineering applications. On the other hand, representing formulas as trees is particularly suited to the implementation of the genetic programming operators.

The fundamental aspect of this type of evolutionary programmes consists of determining the quality of the candidate models. This is achieved with a specific metric called fitness function (FF). The FF is typically a cost function, i.e., an indicator to be minimised. On the basis of the FF, the best candidate models of each generation are selected and they are granted a higher probability to have descendants. Therefore, the better the FF of an individual (the lower the value of its FF), the higher is the probability that its genes can be passed on to the next generations.

Originally the most widely implemented form of the fitness function was the MSE. Nowadays, to find a better trade-off between goodness of fit and complexity, in practice model selection criteria of the AIC and BIC families are the most often adopted. Consequently, improving these indicators with the ameliorations proposed in this work is expected to increase the performance of GPSR as well. This has been checked with a series of systematic tests to identify tens of equations and models from the Feynman Lectures on Physics, adopting an approach very similar to the one described in (Udrescu and Tegmark 2020). Some representative cases are reported in Table 7. Again, the results have been very positive, in the sense that the appropriate choice of the cost function always tends to improve the selectivity of GPSR. This has been verified for most improvements described in Sect.s 3, 4 and 5.

To further exemplify the benefits of adopting ameliorated versions of the fitness function in GPSR, again a quite challenging case of a system with feedback is described in the following, namely the set of Lotka-Volterra ordinary differential Eqs. (42), (43). The initial conditions in this case are $x_0 = y_0 = 2$, and the solutions have been evaluated between 0 and 13 s in 500 points. The chosen model's coefficients are: $\alpha = 2/3, \beta = 4/3, \delta = \gamma = 1$. A normal random noise with mean $\mu = 0$ and standard deviation $\sigma = 0.15$ has been added to the solution. Also, the signals have been contaminated with outliers, obtained summing a quantity equals to $4 \cdot \sigma$ to 50 randomly selected points in the time series. The two obtained signals have been indicated with x_{out}, y_{out}. The two synthetic signals have been given as input to the GPSR algorithm.

The performances of the algorithm with two different fitness functions have been compared. The first fitness function used is the standard BIC version, while the second one is the $\mathrm{BIC_{Rob}}$, described in SubSect. 5.2 The most important configuration parameters for the GPSR runs are reported in Table 8. The k-step ahead prediction parameter represents the number of forward steps for evaluating the model residuals.

For the modelling of the x_{out}, the two models identified by the algorithm using the BIC and the $\mathrm{BIC_{Rob}}$ are reported in Table 9. From inspection of Table 9, it can be noted that the algorithm could not converge on the right solution with the standard version of the BIC as the fitness function. On the contrary, deploying the enhanced $\mathrm{BIC_{Rob}}$, the algorithm manages to find the correct model. Indeed, the corresponding Eq. (46.2) can be rewritten as:

$$\frac{x_{out}(t) - x_{out}(t-1)}{\Delta t} = \frac{0.007}{\Delta t} * x_{out}(t-1) - \frac{0.043}{\Delta t} y_{out}(t-1) \cdot x_{out}(t-1) \qquad (48)$$

Approximating the finite difference as the derivative and substituting $\Delta t = 0.0261$, Eq. (47) becomes:

Table 7 Some physical models selected for application of the improvements of Sect. 3 and 4

Feynman Eq	Equation	Correct model	Alternative model
I.6.20	$f = \frac{e^{-\frac{\theta^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$	$f = b(1)\frac{e^{-\frac{\theta^{b(2)}}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} + b(3)$	$f = b(1)\frac{e^{-\frac{\theta^{b(2)}}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} + b(3)\frac{e^{-\frac{\theta^{b(4)}}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} + b(5)$
I.13.4	$K = \frac{1}{2}m(v^2 + u^2 + w^2)$	$K = m\left(b(1)v^2 + b(2)u^2 + b(3)w^2\right) + b(4)$	$K = m\left(b(1)v^2 + b(2)u^2 + b(3)w^2 + b(4)v + b(5)u + b(6)w\right) + b(7)$
I.34.14	$\omega = \frac{1+v/c}{\sqrt{1-v^2/c^2}}\omega_0$	$\omega = \frac{b(1)+b(2)v/c}{\sqrt{1-v^2/c^2}}\omega_0 + b(3)$	$\omega = \frac{b(1)+\frac{b(2)v}{c}+\frac{b(3)v^2}{c}}{\sqrt{1-v^2/c^2}}\omega_0 + b(4)$
I.37.4	$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta$	$I = b(1)I_1 + b(2)I_2 + b(3)\sqrt{I_1 I_2}\cos\delta + b(4)$	$I = b(1)I_1 + b(2)I_2 + b(3)\sqrt{I_1 I_2}\cos\delta + b(4)I_1 I_2\cos\delta + b(5)$
I.44.4	$E = nk_bTln\left(\frac{V_2}{V_1}\right)$	$E = b(1)nk_bTln\left(\frac{V_2}{V_1}\right) + b(2)$	$E = b(1)nk_bTln\left(\frac{V_2}{V_1}\right) + b(2)n^2k_bTln\left(\frac{V_2}{V_1}\right) + b(3)$
II.11.28	$\theta = 1 + \frac{n\alpha}{1-\frac{n\alpha}{3}}$	$\theta = b(1) + \frac{b(2)n\alpha}{1-\frac{n\alpha}{3}}$	$\theta = b(1) + \frac{b(2)n\alpha}{1-\frac{n\alpha}{3}} + \frac{b(3)n^2\alpha^2}{1-\frac{n\alpha}{3}}$
II.35.21	$M = n_p\mu_M\tanh\left(\frac{\mu_M B}{k_bT}\right)$	$M = b(1)n_p\mu_M\tanh\left(\frac{\mu_M B}{k_bT}\right) + b(2)$	$M = b(1)n_p\mu_M\tanh\left(\frac{\mu_M B}{k_bT} + b(2)\right) + b(3)$
III.8.54	$p_\gamma = sin\left(\frac{Et}{h}\right)^2$	$p_\gamma = b(1)sin\left(\frac{Et}{h} + b(2)\right)^2 + b(3)$	$p_\gamma = b(1)sin\left(\frac{Et}{h} + b(2)\right)^2 + b(3)sin\left(\frac{Et}{h}\right)^2 + b(5)$
III.14.14	$I = I_0\left(e^{\frac{qV_e}{k_bT}} - 1\right)$	$I = b(1)I_0\left(e^{\frac{qV_e}{k_bT}} - b(2)\right) + b(3)$	$I = b(1)I_0\left(e^{\frac{qV_e}{k_bT}} - b(2)\right) + b(3)I_0\left(e^{\frac{qV_e}{k_bT}}\right) + b(4)$

Table 8 The table reports the number of successful identifications out of 100 different noise realisations. Configuration parameters for the SR via GP algorithm

Population size	100
Number of generations	50
Max. time shift	1
k-step ahead prediction	10
Active functions	$*, +, -, \div, a^b; e^a$

Table 9 Models selected by SR via GP for x_{out}

	models
BIC	$x_{out}(t) = 0.9898 * x_{out}(t-1) - 0.032 * y_{out}(t-1) - 5,9e-4 \cdot \exp\left(y_{out}(t-1) \cdot x_{out}(t-1)\right)$ (46.1)
BIC_{Rob}	$x_{out}(t) = 1.007 * x_{out}(t-1) - 0.043 \cdot y_{out}(t-1) \cdot x_{out}(t-1)$(46.2)

Table 10 Models selected by SR via GP for y_{out}

	models
BIC	$y_{out}(t) = 0.9628 * y_{out}(t-1) - \exp\left(y_{out}(t-1)^2\right) * 5,9e-5-2.9e-4 \cdot \exp\left(y_{out}(t-1) \cdot x_{out}(t-1)\right) + 0.01739 \cdot y_{out}(t-1) \cdot x_{out}(t-1)$(49.1)
BIC_{Rob}	$y_{out}(t) = 0.9583 * y_{out}(t-1) + 0.02718 \cdot y_{out}(t-1) \cdot x_{out}(t-1)$(49.2)

$$\frac{dx_{out}}{dt} = 0.27 * y_{out}(t-1) - 1.64 \cdot y_{out}(t-1) \cdot x_{out}(t-1) \tag{49}$$

Equation (47) is exactly Eq. (42) that was used to generate the data. The model parameters can indeed be refined in a post-processing fitting to recover exactly the original equation.

The same has been done for the modelling of y_{out}.The results are reported in Table 10. In this case, Eq. (49.2) can be rewritten as:

$$\frac{y_{out}(t) - y_{out}(t-1)}{\Delta t} = \frac{0.0417}{\Delta t} * y_{out}(t-1) + \frac{0.02718}{\Delta t}\, y_{out}(t-1) \cdot x_{out}(t-1) \tag{50}$$

Approximating the finite difference as the derivative and substituting $\Delta t = 0.0261$, Eq. (50) becomes:

$$\frac{dy_{out}}{dt} = 1.59 * y_{out}(t-1) + 1.04 \cdot y_{out}(t-1) \cdot x_{out}(t-1) \tag{51}$$

Which is again equal to Eq. (43) implemented to generate the data. Again, without the proposed improvement of the fitness function, GPSR practically never manages to converge on the right model.

8 Discussion and conclusions

The identification of the most adequate equations, to model a phenomenon given the available data, is a fundamental task in both science and engineering. The holy grail of model selection is the definition of a universal indicator that would always produce the correct hierarchy of the candidate models. In the effort to identify such a general indicator, powerful and conceptually sound families of criteria have been devised, in the framework of information theory and Bayesian statistics. The AIC is the representative of the first class and BIC of the second. These criteria, and all the numerous variants derived from them, are all informed by the conceptual objective of identifying a general solution valid for all applications.

In practice, the most widely used versions of the model selection criteria AIC and BIC assume that the data are affected by Gaussian, zero sum additive noise. This is a consequence of the fact that, in most practical applications in science and engineering, it is often very difficult, if not impossible, to compute the likelihood of the data given the model. In this situation, the AIC and BIC have been reformulated as reported in Eqs. (4) and (5). These versions of the criteria can fail badly, when the data does not verify the underlying assumptions and mainly for three orders of reasons. First, because the statistical information about the residuals, limited to their MSE and variance, can become insufficient to discriminate properly between the candidate models. The second main weakness of the popular versions of the AIC and BIC, resides in their rudimental quantification of complexity, which can result in major blunders. The last aspect, rendering the traditional versions of AIC, BIC and their families sometimes inadequate, is the implicit assumption that all the parts of the data pdfs are equally relevant and of equal quality. All these limitations are particularly dangerous and penalising when investigating real life complex systems. The practitioners would therefore certainly benefit from more flexible but at the same time not arbitrary versions of MSC.

Adopting an objective utility-based approach to model selection, various improvements of the AIC and BIC families of criteria have been devised, which alleviate all the previously mentioned limitations of the traditional forms. The proposed modifications constitute an array of tools, giving the user ample choice of the most adequate indicators to implement, depending on the application, the goals of the analysis and the prior information available. In most situations, the new versions of the criteria provide better discriminatory capability. The weighting of the entries and the robust statistics indicators proposed in Sect. 5 are preliminary measures that have always a positive effect on the final selection. Exploiting all the statistical information contained in the residuals and not only the MSE, when it is not possible to calculate the likelihood of the models, is also very beneficial. A better quantification of the model complexity can also be very useful in many contexts. Moreover, the requirements of the developed improvements, in terms of both data quality and computational resources, are not significantly more demanding than those of the original versions of the criteria. Of course, in full harmony with the conceptual framework proposed, none of the proposed advances can claim absolute validity or be applied blindly. Their deployment requires serious considerations and conscious decisions by the analysts about their potential validity. On the other hand, the context, in which each one is more appropriate, can be determined quite clearly in an objective way. Moreover, the appropriate improvements can be very valuable particularly in difficult situations, including complex systems, high uncertainties and feedback loops. It is also worth pointing out that, even if in the paper the discussion

has been particularised for regression, the same improvements can be also very useful for classification and density estimation.

To avoid difficulties in the application of the proposed enhanced criteria, no recourse has been made to the prior probability of the models and their Bayes factors. Assigning a probability to a model is an issue, which has been the subject of much discussion for many years (Key, Pericchi, and Smith, 1999). Difficulties such as the Jefferys-Lindley paradox and the use of improper priors are not solved to the present day and noninformative priors present their own issues (Spanos 2013). The fact that the developed indicators do not require the definition of a prior probability of the model is therefore to be considered a positive. On the other hand, coherently with the aim of providing a series of objective utility-based tools, if solid prior probabilities can be assigned to the candidate models, it would be important to allow the user to take advantage of that information. To this end, one can resort to a solution analogous to the odds of the Bayes factors. Indicating with π_k the prior of the k model, the inverse odd ratios (INVODDR) for the various candidates can be calculated as:

$$\mathrm{INVODDR_k} = \pi_k \mathrm{CRIT_k} / \left(\sum \pi_i \mathrm{CRIT_i} \right) \tag{52}$$

Where CRIT indicates any of the proposed criteria and the sum is over the candidate models. Given the fact that the proposed indicators are cost functions to be minimised, the model with the lowest odd ratio is the one to be selected. Of course, great care must be taken, because a strong subjective element can be introduced by the selection of the priors π_i.

In terms of future developments, as a consequence of their performance and their simple implementation, the proposed new versions of the selection criteria are expected to be deployed quite systematically in various fields of complex science, ranging from high temperature plasmas (Ongena 2004; Murari 2015; Puiatti 2002; Saarelma 2018) (Martini 2007) to atmospheric physics (Gaudio 2013). Other interesting applications could be found in the regularization of recent tomographic inversion methods (Craciunescu 2009; Odstrčil 2012) and machine learning. From a methodological perspective, the proposed criteria could be further improved by implementing more advanced metrics, such as the geometric distance (Amari and Nagaoka 2000; Craciunescu 2016; Murari 2013) (Craciunescu 2018) and the Venn definition of probability (Dormido-Canto 2013). It should also be mentioned that the proposed tools and techniques can be applied equally well to the outputs of large-scale simulations, which nowadays can produce enormous amounts of data very difficult to interpret (Dubois 2018). The solutions proposed are also expected to contribute to other delicate learning tasks, such as transfer learning, which have become quite important in modern societies (Robert and Adrian 1995).

Author contributions A.M conceptualization and writing; M.G and R.R numerical tests; M.L, L.S. and P.G. numerical tests; M.G coordination; M.L and R.R prepared figures; A.M. theoratical developments; All authors reviewed the manuscript

Funding Open access funding provided by Università degli Studi di Roma Tor Vergata within the CRUI-CARE Agreement. No funding was received to assist with the preparation of this manuscript.

Data availability The datasets generated during and/or analysed during the current study are either publicly available online (https://osf.io/drwcq/) or available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control. https://doi.org/10.1109/TAC.1974.1100705

Amari S, Nagaoka H (2000) Methods of Information Geometry. Oxford University Press, Oxford

Ando T (2010) Bayesian model selection and statistical modeling. CRC Press, Boca Raton

Arndt, C. (2004). *Information Measures, Information and its Description in Science and Engineering.* Springer Series: Signals and Communication Technology. doi:978–3–540–40855–0

Åström KJ, Murray RM (2008) What is feedback?", Feedback Systems: An Introduction for Scientists and Engineers. Princeton University Press, Princeton

Bailly F, Longo G (2011) Mathematics and the Natural Sciences. Imperial College Press, London

Bartlett P, Mendelson S (2002) Rademacher and gaussian complexities: risk bounds and structural results. J Mach Learn Res 3:463–482

Baudot P, Tapia M, Bennequin D, Goaillard JM (2019) Topological Information Data Analysis. Entropy 21(9):869

Bomze IM (1983) Lotka-Volterra equation and replicator dynamics: A two-dimensional classification. Biol Cybern 48(3):201–211. https://doi.org/10.1007/bf00318088

Bousquet, O. (2004). Introduction to statistical learning theory. Biol Cybern 3176(1):169–207. https://doi.org/10.1007/978-3-540-28650-9_8

Breiman L (2001) Statistical modeling: the two cultures. Stat Sci 16:199–231. https://doi.org/10.1214/ss/1009213726

Cavanaugh JE, Neath AA (2019) The Akaike information criterion. Wires Comput Stat 11(3):e1460

Chen Q, Xue B, Zhang M (2020) Rademacher complexity for enhancing the Generalisation of Genetic Programming for symbolic regression. IEEE Transactions on Cybernetics. https://doi.org/10.1109/TCYB.2020.3004361

Claeskens G (2016) Statistical model choice. Annu Rev Stat Appl 3(1):233–256

Corder GW, Foreman DI (2014) Nonparametric Statistics: A Step-by-Step Approach. Wiley, New York

Craciunescu T (2009) A comparison of four reconstruction methods for JET neutron and gamma tomography. Nucl Instrum Methods Phys Res 605(3):374–383. https://doi.org/10.1016/j.nima.2009.03.224

Craciunescu T (2016) Geodesic distance on Gaussian manifolds for the robust identification of chaotic systems. Nonlinear Dyn 86(1):677–693. https://doi.org/10.1007/s11071-016-2915-x

Craciunescu T, Peluso E, Murari A, Gelfusa M (2018) Maximum likelihood bolometric tomography for the determination of the uncertainties in the radiation emission on JET TOKAMAK. Rev Scientific Instruments 89(5):053504. https://doi.org/10.1063/1.502788

D'Espargnat B (2002) On Physics and Philosophy. Princeton University Press, Ocford

Ding J (2018) Model selection techniques – an overview. IEEE Signal Process Mag 35(6):16–34. https://doi.org/10.1109/MSP.2018.2867638

Dormido-Canto S (2013) Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER. Nucl Fusion 53(11):113001

Dubois G (2018) Modeling and Simulation. CRC Press, Boca Raton

Farcomeni A, Greco L (2013) Robust methods for data reduction. Chapman and Hall/CRC Press, Boca Raton

Gaudio, P., and et al. (2013). *Design and development of a compact Lidar/Dial system for aerial surveillance of urban areas.* Proceedings of SPIE - The International Society for Optical Engineering.
Guiasu S (1986) Grouping data by using the weighted entropy. J Stat Plan Inference 15:63–69
Hettmansperger TP, McKean JW (1998) Robust nonparametric statistical methods. John Wiley, New York
Hingston P., L. Barone, and Z. Michalewicz (Editors), Design by Evolution, Natural Computing Series, 200 8, Springer, ISBN 3540741097Huber, P. J. (1981). *Robust statistics.* New York: John Wiley and Sons, Inc.
Huber PJ (1981) Robust statistics. John Wiley & Sons, Inc, New York
Karpinski M, Macintyre A (1997) Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. J Comput Syst Sci 54(1):169–176. https://doi.org/10.1006/jcss.1997.1477
Kelbert, M., Stuhl, I., and Suhov, Y. (2017). Weighted Entropy and its Use in Computer Science and Beyond. Analytical and Computational Methods in Probability Theory - 1st International Conference, ACMPT 2017, Proceedings.
Kenneth PB, Anderson DR (2002) Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer, Berlin
Key, J. T., Pericchi, L. R., and Smith, A. F. (1999). Bayesian model choice: what and why. *Bayesian statistics*.
Koza JR (1992) Genetic Programming: on the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge
Lofti S, Izmailov P, Benton G, Goldblum M, Wilson AG (2022) Proceedings of the 39th International Conference on Machine Learning. PMLR 162:14223–14247
Mark C, Metzner C, Lautscham L et al (2018) Bayesian model selection for complex dynamic systems. Nat Commun 9:1803. https://doi.org/10.1038/s41467-018-04241-5
Martini S et al (2007) Active MHD control at high currents in RFX-mod Nucl. Fusion 47:783
McDonald, J. D., Shalizi, C. R., and Schervish, M. (2011). Estimated VC dimension for risk bounds. *Neural Computation*.
Meyers RA (2009) Encyclopedia of Complexity and Systems Science. Springer, New York
Miller AJ (2002) Subset selection in regression. CRC Press, Boca Raton
Mitchell M (2009) Complexity: A Guided Tour. Oxford University Press, Oxford
Murari A (2012) A statistical methodology to derive the scaling law for the H-mode power threshold using a large multi-machine database. Nucl Fus 52(6):063016
Murari A (2013) Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions. Nucl Fus. https://doi.org/10.1088/0029-5515/53/3/033006
Murari A (2019) On the use of entropy to improve model selection criteria. Entropy 21(4):394. https://doi.org/10.3390/e21040394
Murari A (2021) Alternative definitions of complexity for practical applications of model selection criteria. Complexity. https://doi.org/10.1155/2021/8887171
Murari A, Peluso E, Gelfusa M, Lupelli I, Gaudio P (2015) A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks. Nucl Fus 55(7):073009
Murari A, Lungaroni M, Peluso E et al (2019) A model falsification approach to learning in non-stationary environments for experimental design. Sci Rep 9:17880
Murari A, Peluso E, Lungaroni M (2020) Data driven theory for knowledge discovery in the exact sciences with applications to thermonuclear fusion. Sci Rep. https://doi.org/10.1038/s41598-020-76826-4
Odstrčil M (2012) Modern numerical methods for plasma tomography optimisation. Nucl Inst Methods Phys Res Sect A-Accel Spectrom Detect Assoc Equip 686:156–161
Ongena J (2004) Towards the realization on JET of an integrated H-mode scenario for ITER. Nucl Fus 44(1):124–133. https://doi.org/10.1088/0029-5515/44/1/015
Piqueira, J.R.C.. (2018). Dynamic Complexity Measures: Definition and Calculation. https://doi.org/10.20944/preprints201801.0099.v1
Puiatti M (2002) Radiation pattern and impurity transport in argon seeded ELMy H-mode discharges in JET. Plasma Phys and Control Fusion. https://doi.org/10.1088/0741-3335/44/9/305
Ricardo L-R, Mancini H, Calbet X (1995) A statistical measure of complexity. Phys Lett A 209(5–6):321–326
Robert KE, Adrian E (1995) Bayes factors. Raftery J Am Stat Assoc 90(430):773–795
Rossi R (2020) Upgrading model selection criteria with goodness of fit tests for practical applications. Entropy 22(4):447
Rousseeuw PJ, Hubert M (2011) Robust statistics for outlier detection. Wiley Interdiscip Rev: Data Mining and Knowl Discov 1(1):73–79. https://doi.org/10.1002/widm.2
Saarelma S (2018) Integrated modelling of H-mode pedestal and confinement in JET-ILW. Plasma Phys Control Fusion. https://doi.org/10.1088/1361-6587/aa8d45

Schmid M, Lipson H (2009) Distilling free-form natural laws from experimental data. Science 324(5923):81–85. https://doi.org/10.1126/science.1165893
Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464. https://doi.org/10.1214/aos/1176344136
Spanos A (2013) Who should be afraid of the Jeffreys-Lindley paradox? Philos Sci 80(1):73–93. https://doi.org/10.1086/668875
Stoica P, Selen Y (2004) Model-order selection: a review of information criterion rules. IEEE Signal Process Mag 21(4):36–47. https://doi.org/10.1109/MSP.2004.1311138
Sumathi S, Hamsapriya T, Surekha P (2008) Evolutiory intelligence. Springer Verlag, Berlin
Turchin P (2003) Complex Population Dynamics: a Theoretical/Empirical Synthesis. Princeton University Press, Princeton
Udrescu, S., and Tegmark, M. (2020). AI Feynman: a Physics-Inspired Method for Symbolic Regression. *Science Advances*.
Vapnik V (2000) The nature of statistical learning theory. Springer, Berlin
Wang Z, Bovik AC (2009) Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. IEEE Signal Process Mag 26(1):98–117. https://doi.org/10.1109/MSP.2008.930649
Wesson J (2004) Tokamaks. Oxford Clarendon Press, Oxford
Wilcox R (2012) Introduction to robust estimation and hypothesis testing, Statistical Modeling and Decision Science. Elsevier/Academic Press, Amsterdam
Zhou Y, Herath HM (2016) Evaluation of alternative conceptual models for groundwater modelling. Geosci Front. https://doi.org/10.1016/j.gsf.2016.02.002

Authors and Affiliations

Andrea Murari[1,2] · Riccardo Rossi[3] · Luca Spolladore[3] · Michele Lungaroni[3] · Pasquale Gaudio[3] · Michela Gelfusa[3]

✉ Michela Gelfusa
gelfusa@ing.uniroma2.it

Andrea Murari
andrea.murari@istp.cnr.it

Riccardo Rossi
r.rossi@ing.uniroma2.it

Luca Spolladore
luca.spolladore@uniroma2.it

Michele Lungaroni
michele.lungaroni@uniroma2.it

Pasquale Gaudio
gaudio@ing.uniroma2.it

[1] Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padua, Italy

[2] Istituto per la Scienza e la Tecnologia Dei Plasmi, CNR, Padua, Italy

[3] Department of Industrial Engineering, University of Rome "Tor Vergata", Via del Politecnico 1, Rome, Italy