How to Handle Error Bars in Symbolic Regression for Data Mining in Scientific Applications

A. Murari[1], E. Peluso[2(✉)], M. Gelfusa[2], M. Lungaroni[2], and P. Gaudio[2]

[1] Consorzio RFX-Associazione EURATOM-ENEA per la Fusione,
Corso Stati Uniti, 4, 35127 Padova, Italy
[2] Associazione EURATOM-ENEA - University of Rome “Tor Vergata”,
Via del Politecnico 1, 00133 Rome, Italy
emmanuele.peluso@uniroma2.it

Abstract. Symbolic regression via genetic programming has become a very useful tool for the exploration of large databases for scientific purposes. The technique allows testing hundreds of thousands of mathematical models to find the most adequate to describe the phenomenon under study, given the data available. In this paper, a major refinement is described, which allows handling the problem of the error bars. In particular, it is shown how the use of the geodesic distance on Gaussian manifolds as fitness function allows taking into account the uncertainties in the data, from the beginning of the data analysis process. To exemplify the importance of this development, the proposed methodological improvement has been applied to a set of synthetic data and the results have been compared with more traditional solutions.

Keywords: Genetic programming · Symbolic regression · Geodesic distance · Scaling laws

1 Introduction

One of the main objectives of data analysis in scientific applications consists of the task of extracting from the data mathematical expressions for the problem at hand, to be compared with theoretical models and computer simulations. In the last years, symbolic regression (SR) via genetic programming (GP) has proved to be a very useful approach to this task. The method allows exploring the available databases and identifying among the mathematical expressions produced, the Best Unconstrained Empirical Model Structure (BUEMS) to describe the phenomena of interest. The main advantage of the proposed approach consists of practically eliminating any assumption about the mathematical form of the models. The obtained equations are therefore data driven and not hypothesis driven. The basic elements of symbolic regression via genetic programming are covered in Section 2.

A. Gammerman et al. (Eds.): SLDS 2015, LNAI 9047, pp. 347–355, 2015.
DOI: 10.1007/978-3-319-17091-6_29

In the last years, SR via GP has been successfully applied in various fields and has obtained significant results even in the analysis of very complex systems such as high temperature plasmas for the study of thermonuclear fusion [1]. On the other hand, the methodology is still in evolution and would benefit from upgrades aimed at increasing both its versatility and the quality of its results.

In this paper, an approach to handle the error bars in the measurements is proposed, to increase the scientific relevance of the method. Indeed in many scientific studies, such as the extraction of scaling laws from large data sets, very often the measurements are taken as perfect or at least the consequence of their uncertainties is not properly evaluated. In this work, it is shown how the error bars of the measurements can be taken into account in a principled way from the beginning of the data analysis process using the concept of the Geodesic distance on Gaussian Manifolds (GD). The idea, behind the approach proposed, consists of considering the measurements not as points, but as Gaussian distributions. This is a valid assumption in many physics applications, because the measurements are typically affected by a wide range of noise sources, which from a statistical point of view can be considered random variables. Each measurement can therefore be modelled as a probability density function (pdf) of the Gaussian type, determined by its mean μ and its standard deviation σ. The distance between the measurements and the estimates of the various models can therefore be calculated using the GD using the Rao formula (see Section 4). This distance can be used as fitness function in the SR code to converge on the most suitable model taking into account the error bars of the measurements in a principled way, since the beginning of the data analysis process.

With regard to the structure of the paper, SR via GP is introduced in the next section. The mathematical formalism of the GD on Gaussian manifolds is presented in Section 3. The potential of the proposed technique to tackle the uncertainties in the data is illustrated with a series of numerical tests using synthetic data. Some examples of application are described in detail in Section 4. Summary and directions of future activities are the subject of the last section of the paper.

2 The Basic Version of Symbolic Regression via Genetic Programming

As mentioned in the previous section, this paper describes the refinement of advanced statistical techniques for the extraction of mathematical expressions from large databases to investigate the behaviour of scientific phenomena. The main advantage of the basic tools consists of practically eliminating any assumption about the form of the models. This section describes briefly the mathematical basis of the tools implemented to perform the analysis presented in the rest of the paper.

The objective of the method consists of testing various mathematical expressions to model a certain phenomenon, given a database. The main stages to perform such a task are reported in Figure 1. First of all, the various candidate formulas are expressed as trees, composed of functions and terminal nodes. The function nodes can be standard arithmetic operations and/or any mathematical functions, squashing terms as well as user-defined operators [2,3]. The terminal nodes can be independent variables or constants (integer or real). This representation of the formulas allows an easy implementation of symbolic regression with Genetic Programming. Genetic Programs (GPs) are computational methods able to solve complex and non-linear optimization problems [2,3]. They have been inspired by the genetic processes of living organisms. In nature, individuals of a population compete for basic resources. Those individuals that achieve better surviving capabilities have higher probabilities to generate descendants. As consequence, better adapted individuals' genes have a higher probability to be passed on to the next generations.

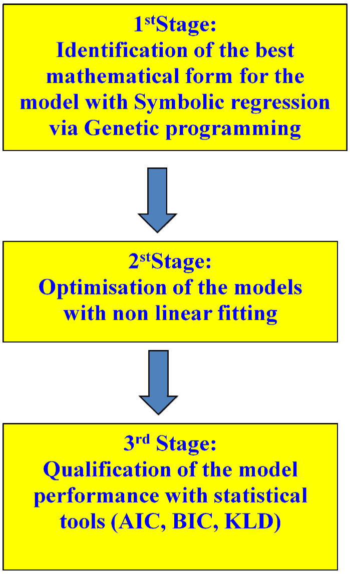
1^{st}Stage:
Identification of the best mathematical form for the model with Symbolic regression via Genetic programming
2^{st}Stage:
Optimisation of the models with non linear fitting
3^{rd} Stage:
Qualification of the model performance with statistical tools (AIC, BIC, KLD)

Fig. 1. The main steps of the proposed methodology to identify the best models without assumption on their mathematical form

GPs emulate this behaviour. They work with a population of individuals, e.g mathematical expressions in our case. Each individual represents a possible solution to a problem. A fitness function (FF) is used to measure how good an individual is with respect to the environment. A higher probability to have descendants is assigned to those individuals with better FF. Therefore, the better the adequacy (the value of the FF) of an individual to a problem, the higher is the probability that its genes are propagated.

In our application, the role of the genes is played by the basis functions used to build the trees representing the various formulas. The list of basis functions used to obtain the results described in the rest of the paper is reported in Table I. Evolution is achieved by using operators that perform genetic like operations such as reproduction, crossover and mutation. Reproduction involves selecting an individual from the current population and allowing it to survive by copying it into the new population. The crossover operation involves choosing nodes in two parent trees and swapping the respective branches thus creating two new offsprings. Mutations are random modifications of parts of the trees.

Table 1. Types of function nodes included in the symbolic regression used to derive the results presented in this paper, x_i and x_j are the generic independent variables

Function class	**List**
Arithmetic	c (real and integer constants),+,-,*,/
Exponential	exp(x_i),log(x_i),power(x_i, x_i), power(x_i,c)
Squashing	logistic(x_i),step(x_i),sign(x_i),gauss(x_i),tanh(x_i), erf(x_i),erfc(x_i)

To derive the results presented in this paper, the Akaike Information Criterion (AIC) has been adopted [4] for the FF. The AIC, is a well-known model selection criterion that is widely used in statistics, to identify the best models and avoid overfitting. The AIC form used in the present work is:

$$AIC = 2k + n \cdot \ln(RMSE/n) \tag{1}$$

where RMSE is the Root Mean Square Error, k is the number of nodes used for the model and n the number of entries in the database (DB). The AIC allows rewarding the goodness of the models, by minimizing their residuals, and at the same time penalising the model complexity by the dependence on the number of nodes. The better a model, the smaller its AIC.

Having optimised the models with non-linear fitting, what remains is the qualification of the statistical quality of the obtained scaling laws. To this end, a series of statistical indicators have been implemented. They range from model selection criteria, such as the Bayesian Information Criterion (BIC), to statistical indicators, such as the Kullback-Leibler divergence (KLD) [5,6].

3 Geodesic Distance to Include the Effects of the Error Bars and Application to Scaling Laws

As seen in the previous sections, the goal of SR via GP is to extract the most appropriate formulas to describe the available data. To achieve this, typically the RMSE of the distances between the data and the model predictions is used in the FF. In this way, SR is implicitly adopting the Euclidean distance. The (dis)similarity between data points and predictions is measured with the Euclidean distance, which has a precise geometrical meaning and a very long historical pedigree. However, it implicitly requires considering all data as single infinitely precise values. This assumption can be appropriate in other applications but it is obviously not the case in science, since all the measurements typically present error bars. The idea is to develop a new distance between data, which would take into account the measurement uncertainties. The additional information provided by this distance should hopefully render the final results more robust. In particular, using the GD the final equations are more general and less vulnerable to the detrimental effects of outliers (see Section 4).

The idea, behind the approach proposed in this paper, consists of considering the measurements not as points, but as Gaussian distributions. This is a valid assumption in many scientific applications, because the measurements are affected by a wide range of noise sources, which from a statistical point of view can be considered random variables. Since the various noises are also typically additive, they can be expected to lead to measurements with a global Gaussian distribution around the most probable value, the actual value of the measured quantity. Each measurement can therefore be modelled as a probability density function (pdf) of the Gaussian type, determined by its mean μ and its standard deviation σ:

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

In this work, having in mind applications to experimental data, it is assumed that the measured values are the most likely ones, within the probability distribution representing the measurement errors. In the case of Gaussian distributions, the most likely value is the mean. Therefore, the individual measurements are assumed to be the means of the Gaussian distributions representing their uncertainties. This is the typical hypothesis adopted in the application of the GD.

Modelling measurements not as punctual values, but as Gaussian distributions, requires defining a distance between Gaussians. The most appropriate definition of distance between Gaussian distributions is the geodesic distance (GD), on the probabilistic manifold containing the data, which can be calculated using the Fischer-Rao metric [7]. For two univariate Gaussian distributions $p_1(x|\mu_1,\sigma_1)$ and $\mathrm{p}_2(\mathrm{x}|\mu_2,\sigma_2)$, parametrized by their mean μ_i and standard deviations $\sigma_i(i = 1{,}2)$, the geodesic distance GD is given by:

$$\mathrm{GD}(\mathrm{p}_1||\mathrm{p}_2) = \sqrt{2}\ln\frac{1+\delta}{1-\delta} = 2\sqrt{2}\tanh^{-1}\delta,\ \text{where } \delta = \left[\frac{(\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2}\right]^{\frac{1}{2}} \tag{3}$$

In the case of multiple independent Gaussian variables, it is easy to prove that the square GD between products of distributions is given by the sum of the squared GDs between corresponding individual distributions.

The last thing required consists of inserting the GD into the SR. To this end, a good solution has proved to be to insert the GD in the FF according to the following formula:

$$AIC = 2k + \textstyle\sum_i GD_i \tag{4}$$

Where the symbols have the same meaning as in formula (1) and the index i runs over the entries of the database.

4 Example of Application: Scaling Laws

To exemplify the flexibility and power of the techniques described in the previous section, they are applied to the problem of deriving scaling laws from the data. Section 4.1 contains an introduction to the problem of data driven scaling laws. In Section 4.2 the capability of the upgraded methodology to identify also scaling laws is proven using synthetic data.

4.1 Scaling Laws

In various fields of science, such as Magnetic Confinement Nuclear Fusion, many systems to be studied are too complex to allow a treatment on the basis of first principle physical models. Therefore, to overcome the gap between theoretical understanding and data, in order to profit from the deluge of experimental measurements of the last decades, data driven models have become quite popular. They are particularly useful to assessing how system properties scale with dimensions or other parameters. One major hypothesis is implicitly accepted in the vast majority of the most credited scaling laws reported in the literature [9,10]: the fact that the scaling laws are in form of power law monomials. Indeed the available databases have typically been studied using traditional log regression. This implies that the final scaling laws are assumed "a priori" to be power law monomials of the form: $B^{\beta}C^{\chi} \dots D^{\delta}G^{-\gamma}H^{-\lambda} \dots K^{-\mu}$; where the capital letters represent physical quantities [11]. This assumption on the form of these scaling equations is often not justified neither on theoretical nor on experimental grounds. Moreover, the use of log regression to analyse the data typically does not take into account the uncertainties in the measurements properly. The two developments described in the previous sections overcome these limitations. SR via GP allows the freedom to identify scaling laws of very different mathematical forms and the GD can take properly into account the error bars in the measurements. The profit of combining these two approaches (SR via GP and a Gaussian manifold based metric) in the proposed tool are shown in the next Section using synthetic data.

4.2 Numerical Results

Synthetic databases have been generated to test the Euclidean metric and Geodesic metric as fitness function for SR via GP. Four formulas have been investigated , each one composed of two terms with three independent variables and one dependent variable, as showing the equations (5), (6), (7), (8). The range of variation of the three independent variables is also reported $x_1 \in [0.015, 3.9]$, $x_2 \in [0.044, 1.97]$, $x_3 \in [0.268, 2.178]$. Values are randomly generated, uniformly distributed within the variation range of the independent variables, in relation to the size of the database. The study is performed by choosing the number of entries in the DB, then checking if the genetic algorithm, with the different metrics, gives in output the formulas used to generate the data.

$$f_1 = \cos(x_1 \cdot x_2) + \sin x_1^{0.5} \tag{5}$$

$$f_2 = \cos\left(\frac{x_1}{x_2}\right) + 2x_3 \cdot \{1/[1 + exp(-0.8 \cdot x_2)]\} \tag{6}$$

$$f_3 = x_2^{1.5} \cdot x_3 - 0.5 \cdot x_1^{0.5} \cdot \cos x_3 \tag{7}$$

$$f_4 = x_1 \cdot x_2 \cdot exp(-x_3) + 2x_3 \tag{8}$$

After the choice of the formula and the size of the inputs, it is possible to choose the type of noise to be applied to the analytical formulation: Uniform, Gaussian and Asymmetric. In the case of uniform noise, it is possible to set the percentage of noise compared to the average value of the chosen analytical formulation. Hence, a random uniform vector is generated, including values between +/- the percentage of the mean value of the used formula. This noise vector is then applied to the dependent variable to generate the noise-affected database. In the case of Gaussian noise, the standard deviation of the distribution of the noise can be selected. This way, a random noise of Gaussian distribution, which has zero mean and standard deviation equal to a percentage of the average value of the chosen analytical formulation, is generated. After that, this noise vector is applied to the dependent variable to generate the noise-affected database. The last type of noise is asymmetric. It consist of two random Gaussian distributions, in which the user can choose the standard deviations of the two distributions, always equal to a percentage of the average value of the analytical formulation. And more, the weight of the first distribution can be chosen, while the second is complementary to the first. The first Gaussian distribution always has zero mean, while the second has mean equal to two times the sum of the two standard deviations. Also, in this case the noise vector is then applied to the dependent variable to generate the noise-affected database.

Once the noise-affected databases are generated, as described before, Symbolic Regressions adopting the Euclidean metric and the Geodesic metric (introduced in section3) are used to study the behaviour of the algorithm. In all the cases in which the noise distributions are Gaussian noise or Uniform noise, the two metrics have provide exactly the same results. This is true for independently from the size of the databases. On the other hand, when the database contains outliers there is a significant improvement in the results of the Geodesic metric compared to the Euclidean metric. Table II shows the results of the SR conducted to database of 50 inputs (very few). This is challenging scenario to check which metric gives the best results in relation the presence of outliers (asymmetric database). In the columns the table shows the characteristic parameters of the asymmetric noise (described previously), the metric used and the standard deviations (for GD metric).The accuracy of algorithm in finding the right formula is shown in the last column. The results that can be found are classified as follows: Global, the algorithm finds perfectly the formulation expected; (1/2) the algorithm finds, at least, the first term of the formulation; (2/2) the algorithm finds, at least, the second term of the formulation; Negative, when the algorithm does not find any terms of the formulation.

Table 2. Summary table of tests carried out on the database of 50 inputs (where there are five points outliers) to the equations from (5) to (8), containing details of all the noise information and the metric listed in this section

Type Noise	Eq.	σ I Gauss [%mean value]	σ II Gauss [%mean value]	Weight I Gauss ∈[0,1]	Metric	GD σ data [%]	GD σ model [%]	Goals
Asymmetric	(5)	15	30	0.9	EUC	---	---	(1/2)
Asymmetric	(5)	15	30	0.9	GD	20	0.1	Global
Asymmetric	(6)	10	50	0.9	EUC	---	---	Negative
Asymmetric	(6)	10	50	0.9	GD	20	0.1	(1/2)
Asymmetric	(7)	15	30	0.9	EUC	---	---	(1/2)
Asymmetric	(7)	15	30	0.9	GD	10	0.1	(1/2)
Asymmetric	(8)	10	50	0.9	EUC	---	---	Negative
Asymmetric	(8)	10	50	0.9	GD	10	0.1	Global

5 Discussion and Conclusions

In this paper, symbolic regression via genetic programming has been tested on different custom noise-affected databases to check the most challenging scenarios in relation to the fitness function (FF) of the algorithm. In the study, many variations of the parameters of the noise have been carried out on all four formulas presented. The obtained results demonstrate that the two metric, Euclidean and Gaussian, return the same performance, for all levels of Gaussian noise and dimensions of the database investigated. On the other hand, when the Geodesic distance is used, the method is significantly more robust to the presence of outliers because the GD tends to discriminate the data too far from the main trend.

So, to summarise, in the case of databases affected by Gaussian noise, SR via GP performs equally well irrespective of the fact the Euclidean or the Geodesic distance is used. In these cases, the using the GD simply provides an alternative way to double-check the quality of the results as it can be seen in Table III. In the case of small databases, affected by a significant percentage of outliers, the use of the GD improves the resilience and the reliability of the results compared to the RMSE. In these situations, as in the case of classification as reported in [12], the use of the GD provides a competitive advantage compared to the traditional RMSE based on the Euclidean distance.

Table 3. Summary table of tests carried out on the database of 50 or 500 inputs with Gaussian noise, using the formulations from (5) to (8), containing details of all the noise information and the metric listed in this section. Results show how both metric can be used as a double check of the quality of the results.

Inputs DB	Eq.	σ I Gauss [%mean value]	Metric	GD σ data [%]	GD σ model [%]	Goals
50	(5)	30	EUC	---	--	(1/2)
50	(5)	30	GD	30	0.1	(1/2)
50	(6)	30	EUC	---	---	(1/2)
50	(6)	30	GD	30	0.1	(1/2)
50	(7)	30	EUC	---	---	(1/2)
50	(7)	30	GD	30	0.1	(1/2)
50	(8)	30	EUC	---	---	(2/2)
50	(8)	30	GD	30	0.1	(2/2)
500	(5)	20	EUC	---	---	Global
500	(5)	20	GD	20	0.1	Global
500	(6)	20	EUC	---	---	(1/2)
500	(6)	20	GD	20	0.1	(1/2)
500	(7)	20	EUC	---	---	(1/2)
500	(7)	20	GD	20	0.1	(1/2)
500	(8)	20	EUC	---	---	Global
500	(8)	20	GD	20	0.1	Global

References

1. Wesson, J.: Tokamaks, 3rd edn. Clarendon Press Oxford, Oxford (2004)
2. Schmid, M., Lipson, H.: Science, vol. 324, April 2009
3. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
4. Hirotugu, A.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974)
5. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapmans & Hall (1986)
6. Burnham, K.P., Anderson, D.R.: Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer (2002)
7. Amari, S., Nagaoka, H.: Methods of information geometry. Translations of mathematical monographs, vol. 191. American Mathematical Society (2000)
8. Connor, J.W., Taylor, J.-B.: Nuclear Fusion 17, 5 (1977)
9. Murari, A., et al.: Nucl. Fusion **53,** 043001 (2013), doi:10.1088/0029-5515/53/4/043001
10. Murari, A., et al.: Nucl. Fusion **52,** 063016 (2012), doi:10.1088/0029-5515/52/6/063016
11. Barenblatt, G.I.: Scaling. Cambridge University Press (2003)
12. Murari, A., et al.: Nucl. Fusion **53,** 033006, (9 p.) (2013)