Error and Correlation as fitness functions for Scaled Symbolic Regression in Grammatical Evolution

Aidan Muphy
Lero & Trinity College Dublin
Dublin, Ireland
murpha56@tcd.ie

Allan De Lima
Lero & University of Limerick
Limerick, Ireland
allan.delima@ul.ie

Douglas Mota Dias
Lero & University of Limerick
Limerick, Ireland
douglas.motadias@ul.ie

Conor Ryan
Lero & University of Limerick
Limerick, Ireland
conor.ryan@ul.ie

ABSTRACT

Linear scaling has greatly improved the performance of genetic programming when performing symbolic regression. Linear scaling transforms the output of an expression to reduce its error. Mean squared error and correlation have been used with scaling, often interchangeably and with assumed equivalence. We examine if this equivalence is justified by investigating the differences between an error-based metric and a correlation-based metric on 11 well-known symbolic regression benchmarks. We investigate the effect a change of fitness function has on performance, individuals size and diversity. Error-based scaling and Correlation were seen to attain equivalent performance and found solutions with very similar size and diversity on the majority of problem, but not all. In order to ascertain if the strengths of both approaches could be combined, we explored a double tournament selection strategy, where two tournaments are conducted sequentially to select individuals for recombination. Double tournament selection found smaller solutions and the best solution in five benchmarks, including finding the best solutions on both real-world dataset used in our experiments.

CCS CONCEPTS

• **Computing methodologies** → **Genetic programming**.

KEYWORDS

Grammatical Evolution, Symbolic Regression, Linear Scaling.

ACM Reference Format:
Aidan Muphy, Allan De Lima, Douglas Mota Dias, and Conor Ryan. 2023. Error and Correlation as fitness functions for Scaled Symbolic Regression in Grammatical Evolution. In *Genetic and Evolutionary Computation Conference Companion (GECCO '23 Companion), July 15–19, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583133.3590709

GECCO '23 Companion, July 15–19, 2023, Lisbon, Portugal

ACM ISBN 979-8-4007-0120-7/23/07.
https://doi.org/10.1145/3583133.3590709

1 INTRODUCTION

Symbolic regression (SR) aims to find a mathematical expression that best fits a given set of data points. Genetic programming (GP) has been used for SR since its invention, with several examples given in Koza's first book [5]. Many of the current state of the art SR methods are still GP based [6].

Assessing the strength of a SR method or comparing different approaches can be done using an error based metric, such as mean squared error (MSE) [11], or correlation based metric, such as the coefficient of determination [4]. Within linear scaling, it has long been assumed by many researchers that these are equivalent and can be used interchangeably to guide the search during evolution. Mathematically they are quite similar to each other, but not the same. However, it has not been established if they are identical as fitness functions for GP based SR.

Recently, scaled correlation has been shown to be a particularly effective method to guide the search when training data points are sparse, or noise has been added to the target [3]. Correlation has also been used extensively in GP for feature selection, feature construction, and to identify useful or important subtrees of solutions [1, 10].

Both LS and correlation try to accomplish the same thing, that is, permit the search to look for the best fit *shape* of the solution and then use a simple fitting procedure to match the potential solution to the target solution. This greatly simplifies the search process, as it does not need to find constants, and instead focuses on evolving the correct structure.

A major difference between both approaches is in their fitness landscape. Error will often have a single global maximum (error = 0). Correlation allows for an infinite number of optimal (r^2 = 0) and near-optimal solutions to be found. However, solutions with perfect correlation may have vastly different errors when scaled. The quality of a solution can only fully be assessed after it has undergone scaling and not during the search if using simple correlation as the fitness function, as we do in our experiments.

Another difference may manifest in solution size. As generations increase and solution quality increases solutions tend to bloat. For this reason, we have chosen grammatical evolution (GE) as our evolutionary technique. GE has been shown to be an efficient way to impart domain knowledge into into solutions through its grammar and producing interpretable solutions [8, 9].

We investigate the performance and diversity of GE on several SR benchmarks using three different fitness functions: mean squared error, linear scaled mean squared error and correlation. With the results of one benchmark suggesting both squared error and correlation may not be exactly equivalent, we utilise novel a double tournament method of selection to try exploit the benefits of both. Two separate tournaments with different ranking criteria, scaled error and correlation, are investigated to ascertain if they improve the search.

2 BACKGROUND

2.1 Symbolic Regression Fitness Functions

The most common fitness function used in GE for SR experiments is MSE or Root MSE. MSE is a measure of how spread out the predicted data points are from the target points. Finding the error between the output of an expression and the target expression has been used by GP to guide the search since Koza [5]. A major paradigm shift occurred with the introduction by Keijzer of LS, where the output of an individual is recalibrated according to

$$Fitness_{\text{Scaled}}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + b\hat{y_i}))^2, \tag{1}$$

where y is the target function and $\hat{y}$ is the evolved expression with n data points. The coefficients a, the intercept, and b, the slope, are found by calculating a simple linear least-squares regression between the output of the training expression and the target output.

Scaling is fitness function agnostic. It is applied to the predicted outputs of the evolved expression, unaffected by what was used to guide the search for the expression. Kommenda et al. [4] and more recently Haut et al. [3] have highlighted the many benefits that a switch in fitness function from error based to correlation can have. An infinite number of optimal correlation solutions exist, in contrast to the singular optimum, zero error, found by error metrics.

The standard squared error for each individual is:

$$SE(y, \hat{y}) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2. \tag{2}$$

Correlation, more commonly described as r^2, is given by:

$$Correlation(y, \hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y_i})(\hat{y_i} - \bar{\hat{y}_i})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y_i})^2 \sum_{i=1}^{n}(\hat{y_i} - \bar{\hat{y}_i})^2}} \tag{3}$$

where y is the target function and $\hat{y}$ is the evolved expression and $\bar{y}$ is the mean with n data points.

It is clear from observing equations 2 and 3 that both expressions are related to each other,

$$Error \propto Correlation$$

. However, due to their difference in implementation described above, they may traverse the search space in very different methods, creating different types of individuals and may be more or less suited to different problem types. We investigate this potential difference.

2.2 Combining Fitness Functions

A novel double tournament as a means to combine both metrics [7] was used. We adapt this double tournament and create a novel approach which uses error and correlation, not size, as our tournament ranking metrics. This new approach will allow as much information about an individual to be considered during selection and may lead to the avoidance of premature convergence.

3 EXPERIMENTAL SETUP

The full experimental system used is shown in Table 1. The experiments were performed using GRAPE [2], an implementation of GE in Python. The target functions are sampled from well-known, and difficult-to-solve GP problems [12].

Table 1: List of the main parameters used to run GE

Parameter	Value	Parameter	Value
Generations	100	Population	500
Elitism	1%	Selection	Tournament (5)
Crossover	0.90	Mutation	0.05

3.1 Fitness functions

The first fitness function investigated is the standard mean squared error (MSE) for each individual, i.e.,

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2. \tag{4}$$

This metric is scaled in our second set of experiments to produce scaled MSE. We denote the results of these experiments *Err*.

Correlation uses the equation below to guide the search:

$$F(y, \hat{y}) = 1 - |Correlation(y, \hat{y})|. \tag{5}$$

After the search is finished, the best individual found from each run undergoes an adjustment, shown previously in Eq. 1. We denote these experiments as *Corr*.

3.2 Diversity Metrics

Structural diversity finds the number of unique expressions in the population this way:

$$Diversity_{\text{Structural}} = \frac{UniquePhenotypes}{TotalPhenotypes}. \tag{6}$$

That is to say, the number of unique phenotypes, not genotypes, in the population.

We report fitness diversity as

$$Diversity_{\text{Fitness}} = \frac{UniqueFitnessScores}{TotalFitnessScores}. \tag{7}$$

4 RESULTS

The full results of the experimentation can be seen in Table 2. *Mean Depth* refers to the mean depth of the best solutions found in 50 runs, *Mean* Div_S and *Mean* Div_F are the structural and fitness diversity, respectively. These diversity metrics of the population are recorded at generation 100, averaged across 50 runs. *Median Fit* and *Best* refer to the results of the test set.

Table 2: Experimental Results. Each setup was run 50 times. The smallest mean depth and the best individual found for each problem are highlighted in bold. Median fitness Results which are underlined indicate that no other method performed significantly better than it, according to our Wilcoxon tests.

Problem Name	Fitness Function	Mean Depth	Mean Div_S	Mean Div_F	Median Fitness	Best Fitness
Keijzer-5	MSE	9.24	0.52	0.46	0.01582	0.00233
	Err	10.82	0.59	0.47	$\underline{6.91 \times 10^{-5}}$	1.85×10^{-9}
	Corr	11.02	0.58	0.47	$\underline{0.00010}$	1.86×10^{-9}
	Err & Corr	**8.42**	0.10	0.09	0.00179	$\mathbf{1.78 \times 10^{-9}}$
	Corr & Err	9.18	0.10	0.09	0.00175	7.26×10^{-9}
Keijzer-13	MSE	**11.26**	0.30	0.28	11.421	2.149
	Err	12.78	0.18	0.15	$\underline{6.718}$	**1.338**
	Corr	13.72	0.18	0.14	$\underline{7.846}$	2.304
	Err & Corr	14.66	0.07	0.06	11.346	1.734
	Corr & Err	15.2	0.06	0.05	$\underline{8.071}$	1.773
Korns-5	MSE	8.5	0.66	0.64	0.01676	0.00069
	Err	7.17	0.37	0.32	$\underline{2.68 \times 10^{-31}}$	**0.0**
	Corr	**5.5**	0.07	0.05	$\underline{4.66 \times 10^{-30}}$	**0.0**
	Err & Corr	-	-	-	-	-
	Corr & Err	-	-	-	-	-
Korns-12	MSE	**7.9**	0.50	0.49	$\underline{1.1098}$	**1.0819**
	Err	15.94	0.72	0.69	1.1145	1.0852
	Corr	15.38	0.69	0.66	1.1135	1.0852
	Err & Corr	14.72	0.21	0.20	1.1113	1.0839
	Corr & Err	14.36	0.21	0.20	1.1119	1.0832
Nguyen-5	MSE	9.94	0.19	0.18	0.00091	1.75×10^{-5}
	Err	9.08	0.11	0.10	$\underline{4.61 \times 10^{-6}}$	**0**
	Corr	**8.44**	0.11	0.09	$\underline{3.55 \times 10^{-6}}$	**0**
	Err & Corr	-	-	-	-	-
	Corr & Err	-	-	-	-	-
Nguyen-7	MSE	**8.0**	0.12	0.11	0.00119	0.00034
	Err	9.52	0.09	0.08	$\underline{0.00011}$	6.28×10^{-6}
	Corr	9.66	0.09	0.07	$\underline{5.75 \times 10^{-5}}$	$\mathbf{4.23 \times 10^{-6}}$
	Err & Corr	8.28	0.03	0.03	0.00016	1.35×10^{-5}
	Corr & Err	8.8	0.03	0.03	0.00014	9.14×10^{-6}
Pagie-1	MSE	11.28	0.33	0.30	0.03493	0.02471
	Err	10.78	0.29	0.23	0.02542	0.01545
	Corr	11.16	0.29	0.24	$\underline{0.02514}$	0.01753
	Err & Corr	**7.9**	0.08	0.05	0.02695	0.02029
	Corr & Err	9.08	0.09	0.06	$\underline{0.02603}$	**0.01218**
Vladislavleva-1	MSE	13.2	0.42	0.42	0.02724	0.00518
	Err	10.96	0.20	0.18	$\underline{0.02153}$	0.00165
	Corr	12.3	0.19	0.17	$\underline{0.02529}$	0.00197
	Err & Corr	9.58	0.04	0.04	$\underline{0.02171}$	**0.00118**
	Corr & Err	**9.36**	0.04	0.04	$\underline{0.02220}$	0.00164
Vladislavleva-4	MSE	18.52	0.80	0.78	0.03119	0.01857
	Err	9.44	0.54	0.50	$\underline{0.01818}$	**0.00598**
	Corr	9.08	0.56	0.52	$\underline{0.02034}$	0.00724
	Err & Corr	7.9	0.07	0.06	0.02644	0.00821
	Corr & Err	**6.36**	0.04	0.04	0.02658	0.01409
Dow	MSE	10.72	0.24	0.24	0.01608	0.01118
	Err	10.16	0.16	0.15	0.01024	0.00759
	Corr	**8.94**	0.15	0.15	0.01024	0.00736
	Err & Corr	10.42	0.05	0.05	$\underline{0.00977}$	**0.00729**
	Corr & Err	9.44	0.05	0.05	0.01010	0.00747
Concrete	MSE	**10.34**	0.44	0.44	232.041	151.904
	Err	11.1	0.11	0.10	$\underline{177.363}$	145.996
	Corr	13.5	0.11	0.10	$\underline{180.612}$	151.150
	Err & Corr	12.2	0.05	0.05	$\underline{179.361}$	149.714
	Corr & Err	11.84	0.06	0.05	$\underline{181.983}$	**142.505**

Err denotes scaled MSE, Corr where correlation was used to guide the search. Err & Corr denotes the first tournament conducted in the double tournament used scaled MSE fitness to rank the individuals, while Corr & Err denotes the opposite, that is, the correlation score was used to rank the individuals in the first tournament.

The best results for *Mean Depth* and *Best* are highlighted in bold, for readability. Wilcox signed rank statistical tests were performed on each pair of fitness functions, with a p-value of 0.05 chosen to decide significance. All median fitness methods underlined signify that the method was not significantly outperformed by any other method.

Unsurprisingly, standard, unscaled MSE was the worst performing fitness function. It was outperformed by every other method on all benchmarks except one, Korns-12, where it found the best single individual and significantly outperformed every other approach. Scaled MSE (Err) and correlation were seen to perform very similarly and did not statistically outperform each other on 10 of the 11 benchmarks considered. Both were able to evolve perfect solutions for the Korns-5 and Nguyen-5 problems. Scaled MSE found better individuals than correlation on the Keijzer-5, Keijzer-13, Pagie-1, Concrete and both Vladislavleva benchmarks while correlation found the better on the Nguyen-7 and Dow problems.

The inconsistent and poor performance of both correlation and scaled MSE on the Pagie-1 and Korns-12 problems motivated the use of a double tournament method. The introduction of the double tournament showed promising results. The scaling-first double tournament method found the best solution in the Keijzer-5, Vladislavleva-1 and Dow problems. The correlation-first double tournament found the best solution in the Pagie-1 and Concrete problems.

MSE was the best at maintaining population diversity in every problem except two. The best solutions it found were smallest on average on the Keijzer-13, Korns-12 and Nguyen-7 problems. *Mean Depth*, *Mean* Div_S and *Mean* Div_F were very close for scaled MSE and correlation on every problem. Both were seen to have between 45%-70% as diverse populations compared to MSE. Correlation-first double tournament found the smallest individual in two benchmarks, as did the scaling first double tournament.

The population diversity was seen to be drastically smaller when using a double tournament. While a reduction was to be expected as it is a more greedy approach, particularly as the tournament size remained the same at 5, such a drastic drop was unexpected.

5 CONCLUSION

We investigated the common belief that a scaled correlation-based fitness function is equivalent to a scaled error-based fitness function on several symbolic regression problems using grammatical evolution. The performance of a correlation fitness function combined with scaling was seen to be very close to that of traditional error-based linear scaling and produce solutions of similar complexity. However, correlation based was seen to outperform linear scaling on the Pagie-1 problem, despite linear scaling finding a better individual. Surprisingly, no difference in size or population diversity was seen between the methods.

Combining fitness and correlation metrics, through the use of a novel double tournament, was explored. Both orders of tournaments, fitness first and correlation first, were conducted and led to the identification of the best individuals on the Keijzer-5, Pagie-1, Vladislavleva-1, Dow and Concrete benchmarks. The double tournaments also consistently found these improved solutions with smaller individuals. Double tournament selection was also seen to result in very poor diversity, both structural and fitness based.

There are many avenues for future research. There is some evidence that the performance of scaled correlation fitness functions and scaled error based fitness functions are not interchangeable and justifies further research on which problems are suited for which, particularly on high dimensional real-world benchmark problems. These problems may benefit more from the combination of correlation and error, as was observed in our experiments. Combining both using a double tournament yielded promising results but may need refining.

ACKNOWLEDGMENTS

The authors are supported by Science Foundation Ireland grants 20/FFP-P/8818 and 16/IA/4605.

REFERENCES

[1] Muhammad Sarmad Ali, Meghana Kshirsagar, Enrique Naredo, and Conor Ryan. 2022. Automated grammar-based feature selection in symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 902–910.

[2] Allan de Lima, Samuel Carvalho, Douglas Mota Dias, Enrique Naredo, Joseph P Sullivan, and Conor Ryan. 2022. GRAPE: Grammatical Algorithms in Python for Evolution. *Signals* 3, 3 (2022), 642–663.

[3] Nathan Haut, Wolfgang Banzhaf, and Bill Punch. 2023. Correlation Versus RMSE Loss Functions in Symbolic Regression Tasks. In *Genetic Programming Theory and Practice XIX*. Springer, 31–55.

[4] Michael Kommenda, Gabriel Kronberger, Christoph Feilmayr, Leonhard Schickmair, Michael Affenzeller, Stephan M Winkler, and Stefan Wagner. 2012. Application of symbolic regression on blast furnace and temper mill datasets. In *Computer Aided Systems Theory–EUROCAST 2011: 13th International Conference, Las Palmas de Gran Canaria, Spain, February 6-11, 2011, Revised Selected Papers, Part I 13*. Springer, 400–407.

[5] John R. Koza. 1992. *Genetic Programming - On the Programming of Computers by Means of Natural Selection.* MIT Press. I–XVIII, 1–419 pages.

[6] William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H Moore. 2021. Contemporary symbolic regression methods and their relative performance. *arXiv preprint arXiv:2107.14351* (2021).

[7] Sean Luke and Liviu Panait. 2002. Fighting bloat with nonparametric parsimony pressure. In *International conference on parallel problem solving from nature*. Springer, 411–421.

[8] Aidan Murphy, Gráinne Murphy, Jorge Amaral, Douglas MotaDias, Enrique Naredo, and Conor Ryan. 2021. Towards incorporating human knowledge in fuzzy pattern tree evolution. In *Genetic Programming: 24th European Conference, EuroGP 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings*. Springer, 66–81.

[9] Aidan Murphy, Gráinne Murphy, Douglas Mota Dias, Jorge Amaral, Enrique Naredo, and Conor Ryan. 2022. Human in the Loop Fuzzy Pattern Tree Evolution. *SN Computer Science* 3, 2 (2022), 1–14.

[10] Aidan Murphy and Conor Ryan. 2020. Improving Module Identification and Use in Grammatical Evolution. In *2020 IEEE Congress on Evolutionary Computation, CEC 2020*, Yaochu Jin (Ed.). IEEE Computational Intelligence Society, IEEE Press.

[11] Aidan Murphy, Ayman Youssef, Krishn Kumar Gupt, Muhammad Adil Raja, and Conor Ryan. 2021. Time is on the side of grammatical evolution. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–7.

[12] David R White, James McDermott, Mauro Castelli, Luca Manzoni, Brian W Goldman, Gabriel Kronberger, Wojciech Jaśkowski, Una-May O'Reilly, and Sean Luke. 2013. Better GP benchmarks: community survey results and proposals. *Genetic Programming and Evolvable Machines* 14, 1 (2013), 3–29.