Symbolic Regression vs. the World

Prof. Fabrício Olivetti de França

Federal University of ABC
Center for Mathematics, Computation and Cognition (CMCC)
Heuristics, Analysis and Learning Laboratory (HAL)

01 October 2022

Topics

1. Regression Analysis
2. Symbolic Regression
3. Current State of SR
4. What else?
5. Scientific Discoveries
6. Conclusion

Regression Analysis

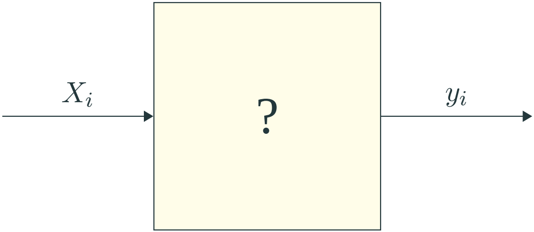
Regression analysis models the **association** between an outcome of interest with a set of predictors. This is useful for:

- **Prediction:** what is the expected price of my house?
- **Associations:** how much variation in price will I observe if richer people moves into my neighborhood?
- **Causal relationship:** by how much will the price change if I split this room in two?

Transparent x Opaque models

Depending on what we want, we have **transparent** and **opaque** models:

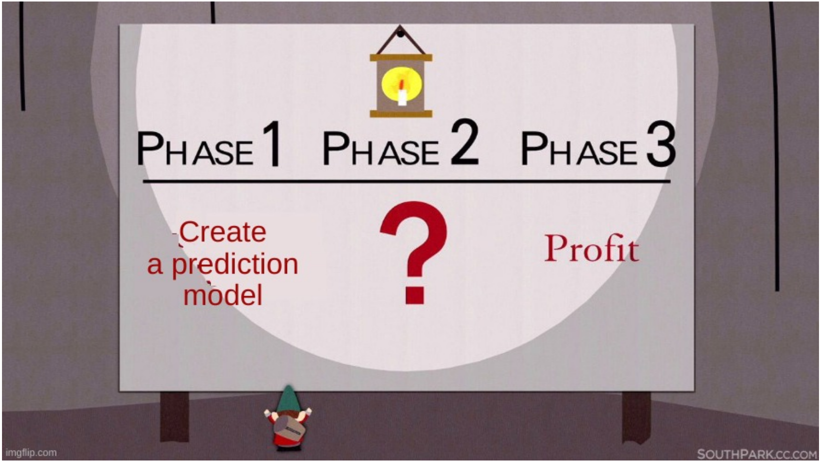
- It is possible to inspect the decision process and the behavior of **transparent** models
- In **opaque** models, this is obscured and external tools are needed to understand its behavior



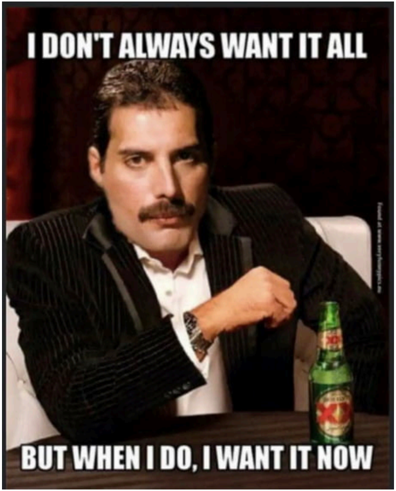
All about prediction

Opaque models (Deep Learning, SVM, Kernel Regression):

- Often associated with a higher predictive power (but not always true).
- If our only concern is prediction, they may be enough.



Transparent x Opaque models



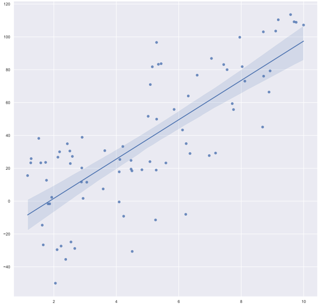
If the objective is to study associations, an opaque model may create a barrier to understand the strength of association of a predictor to the outcome.

Linear Regression

Linear Regression is usually the model of choice for this task:

$$\hat{f}(x, \beta) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$$

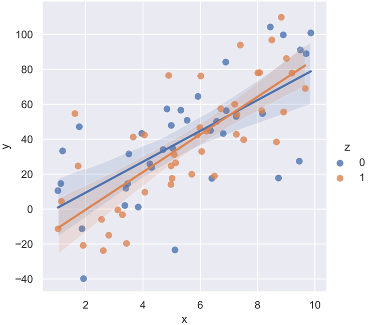
The strength of x_i is given by β_i.



This model can be extended to:

- Include interactions.
- Nonlinear transformations.
- Different distributions.
- etc.

Manual task requiring exploratory analysis, still limited to a generalized linear form.



Nonlinear associations

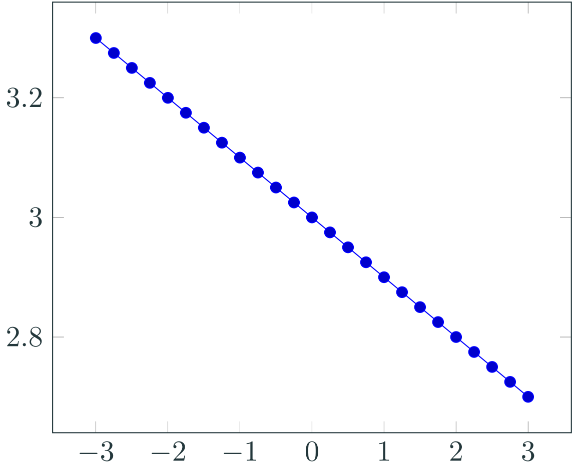
When we augment the model with non-linearity, the association analysis gets more complicated as the strength is not constant:

$$y = 3x_1 - 0.1x_1x_2$$

What is the strenght of association of x_1?

Nonlinear associations

It is $3 - 0.1x_2$, we can actually plot that one!

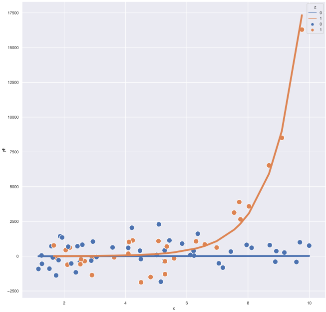


Symbolic Regression

Symbolic Regression

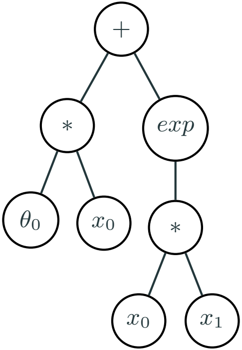
Symbolic Regression searches for a function form together with the numerical coefficients that best fits the outcome.

$$f(x, \theta) = \theta_0 x_0 + e^{x_0 x_1}$$



- Genetic Programming is the most common algorithm to search for the expression
- Represents the solution as an expression tree.

$$f(x, \theta) = \theta_0 x_0 + e^{x_0 x_1}$$



A very simple search meta-heuristic:

```
gp gens nPop = do
  p <- initialPopulation nPop
  until (convergence p) repeat do
      parents   <- select p
      children  <- recombine parents
      children' <- perturb children
      p         <- reproduce p children'
```

Two NP-Hard problems[1]:

- Search for the correct function form $f(x, \theta)$.
- Find the optimal coefficients θ^*.

[1] Virgolin, Marco, and Solon P. Pissis. "Symbolic Regression is NP-hard." arXiv preprint arXiv:2207.01018 (2022).

If we fail into one of them we may discard promising solutions.

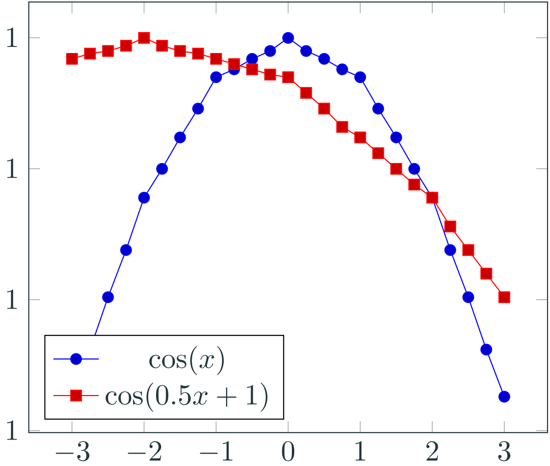


Figura 1: The function $\cos(\theta_1 x + \theta_2)$ may behave differently depending on the choice of θ.

Pros:

- It can find the generating function of the studied phenomena.
- Automatically search for interactions, non-linearity and feature selection.

Cons:

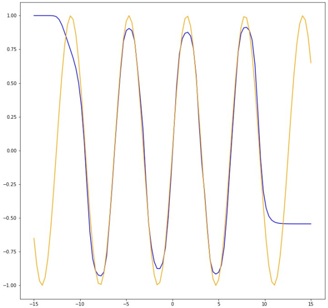
- It can find an obscure function that also fits the studied phenomena.
- The search space can be difficult to navigate.
- Not gradient-based search, it can be slower than opaque models.

Is it worth it?

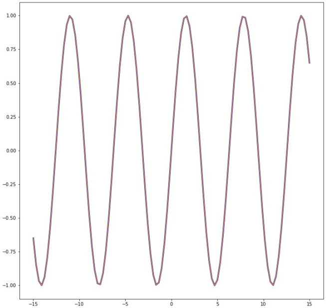
As we can define the primitives, we can choose how *expressive* the model will be.
Consider the $\sin(x)$ function. GP can find the correct model if it contains this function in its primitives.

Is it worth it?

3 layers neural network with sigmoid activation trained on the interval $x \in [-10, 10]$, took 300 seconds and returned this model:



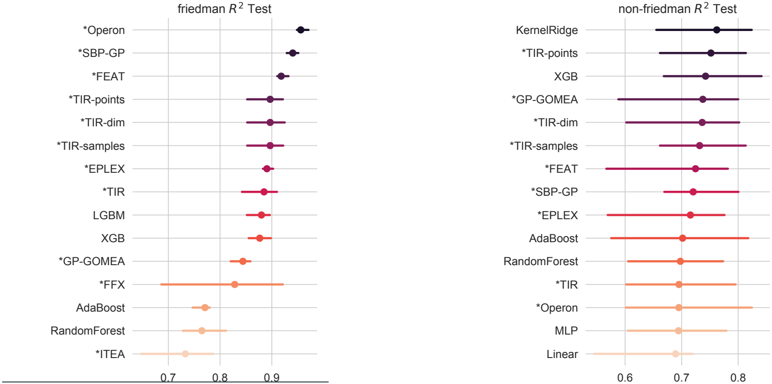
TIR Symbolic Regression model, took 10 seconds and returned this model:



Current State of SR

Is SR competitive?

Benchmark[2] of 22 regression algorithms using 122 benchmark problems, 15 of them are SR algorithms.

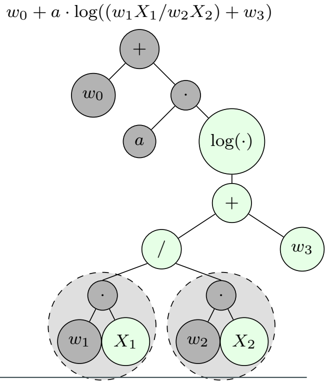


[2] La Cava, William, et al. “Contemporary Symbolic Regression Methods and their Relative Performance.” Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). 2021.

Different approaches

- Many different ideas to improve current results.
- Using nonlinear least squares or ordinary least squares to find θ.
- Constraining the representation.
- Using information theory to improve recombination and perturbation.
- Incorporating multi-objective, diversity control, etc.

Operon C++[3] is a C++ implementation of standard GP and GP with nonlinear least squares for coefficient optimization.



- Competitive runtime, good accuracy
- Supports multi-objective optimization, many hyper-parameters to adjust to your liking
- May overparameterize the model

[3] Burlacu, Bogdan, Gabriel Kronberger, and Michael Kommenda. "Operon C++ an efficient genetic programming framework for symbolic regression." Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion. 2020.

UFABC

Constraint the generated expressions to the form[4]:

invertible function

$$f_{TIR}(\mathbf{x}, \mathbf{w_p}, \mathbf{w_q}) = g\left(\frac{p(\mathbf{x}, \mathbf{w_p})}{1 + q(\mathbf{x}, \mathbf{w_q})}\right)$$

IT expressions

linear coefficient

$$f_{IT}(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{m} w_j \cdot (f_j \circ r_j)(\mathbf{x})$$

transformation function **interaction function**

$$r_j(\mathbf{x}) = \prod_{i=1}^{d} x_i^{k_{ij}}$$

strength of interaction

[4] Fabrício Olivetti de França. 2022. Transformation-interaction-rational representation for symbolic regression. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '22). Association for Computing Machinery, New York, NY, USA, 920–928. https://doi.org/10.1145/3512290.3528695

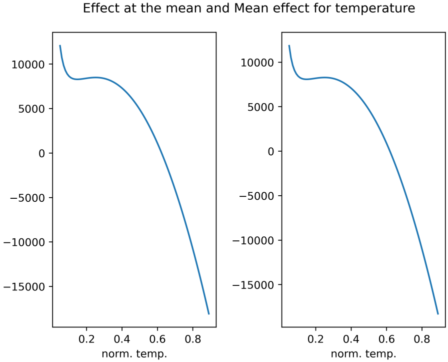

VS. THE WORLD
An epic of epic epicness.
WWW.SCOTTPILGRIMTHEMOVIE.COM

What else?

More benefits to SR

- As a middle ground between opaque and clear model, it can be interpreted
- We can make sure it conforms to our prior-knowledge
- Standard statistical tools can also be applied

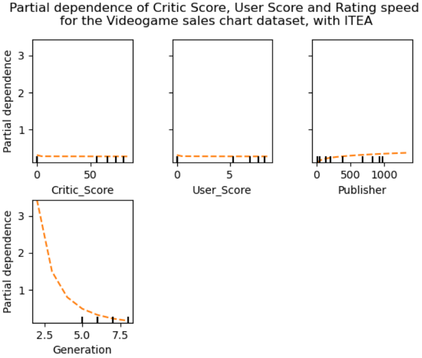
Partial effect at the mean[5] or the mean of the partial effects.



[5] Aldeia, Guilherme Seidyo Imai, and Fabrício Olivetti de França. "Interpretability in symbolic regression: a benchmark of explanatory methods using the Feynman data set." Genetic Programming and Evolvable Machines (2022): 1-41.

Dataset: Video Game Sales

Or a PDP plot[6] if you want to.

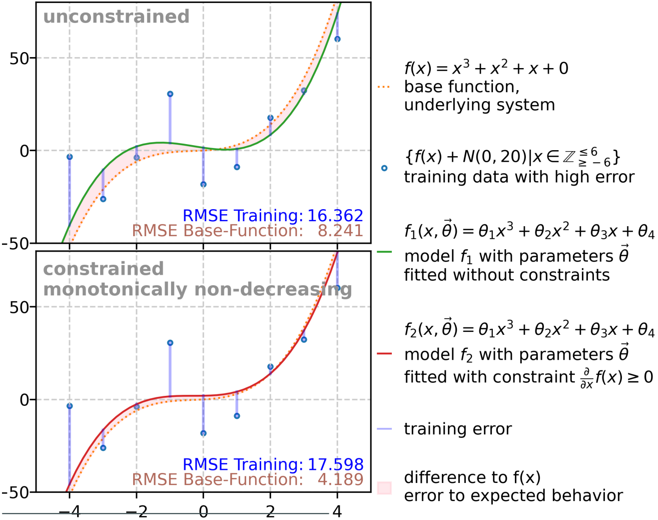


[6]Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

Dataset: Video Game Sales - Debug it!

$$
\begin{aligned}
&0.4593106521142636\\
&+0.08\log(1+\text{publisher}^3\text{gen}^{-3})\\
&-0.09\log(1+\text{critic\_score}^2\text{user\_score}^3\text{gen}^3\text{PC}^3)\\
&+0.21\log(1+\text{user\_count}^3\text{gen}^{-3})\\
&-1.77\log(1+\text{gen})
\end{aligned}
$$

Shape-constraint[7]

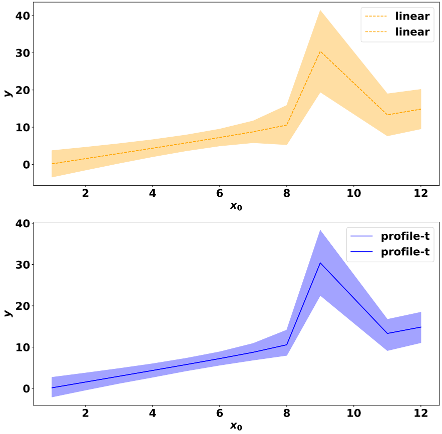


[7] Kronberger, Gabriel, et al. "Shape-Constrained Symbolic Regression—Improving Extrapolation with Prior Knowledge." Evolutionary Computation 30.1 (2022): 75-98.

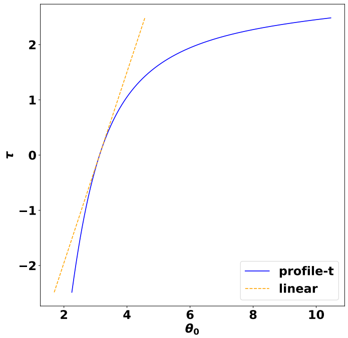
Unlike opaque models, we can calculate the confidence interval of our parameters and predictions using standard statistical tools:

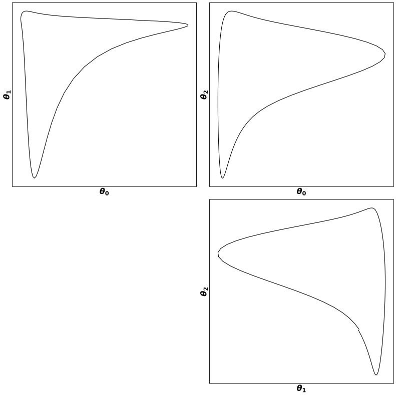
```
SSR 752.76 s^2 28.95
theta    Estimate      Std. Error.    Lower        Upper
0       -1.43e+01      4.29e+00     -2.31e+01    -5.44e+00
1        1.28e+01      2.49e+00      7.67e+00     1.79e+01

Corr. Matrix
[ 1.   -0.97]
[-0.97  1.  ]
```

Prediction intervals:



And many more!





Scientific Discoveries

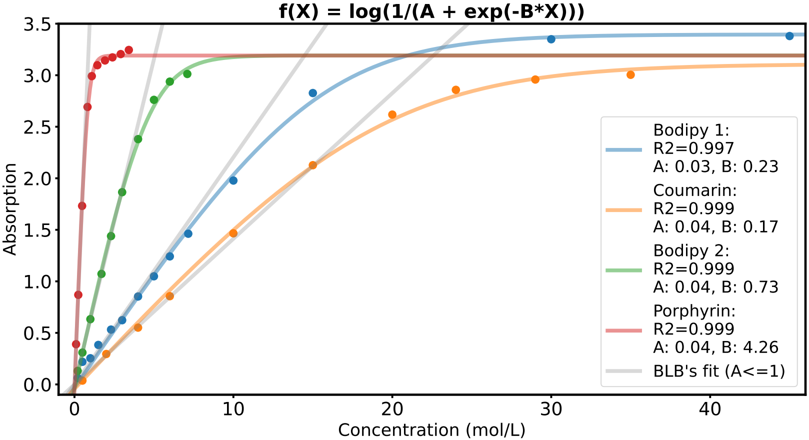
Symbolic Regression is becoming popular to find general laws for different sciences.

Chiefly if you can aggregate the knowledge from multiple views of your data and minimize the number of parameters:

- Neural Networks want overparameterized models to increase its flexibility to fit **any** data
- SR wants models with the right amount of parameters that fits **only** the data of a particular phenomena

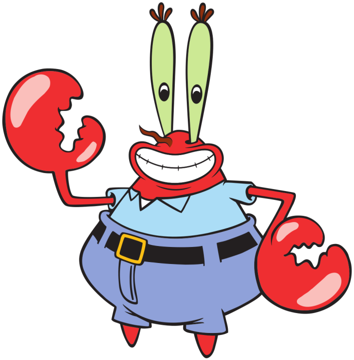
Beer's law relates the attenuation of light to the material through which the light is travelling.
It is presumed to be only due to absorption of a solution, as they do not scatter light of wave- lengths frequently used in analytical spectroscopy.

Using Multi-view Symbolic Regression we found many different alternative laws:

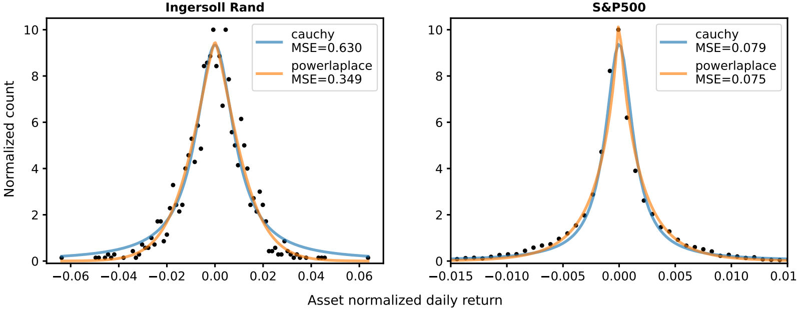


Economy: Distribution of Returns

The distribution of returns in stock prices $(p_{t+1} - p_t)$ is an important information for econometrics to determine the price of an option. A famous model is the Black-Scholes-Merton model, but this neglects the influence of rare and extreme events.



Economy: Distribution of Returns

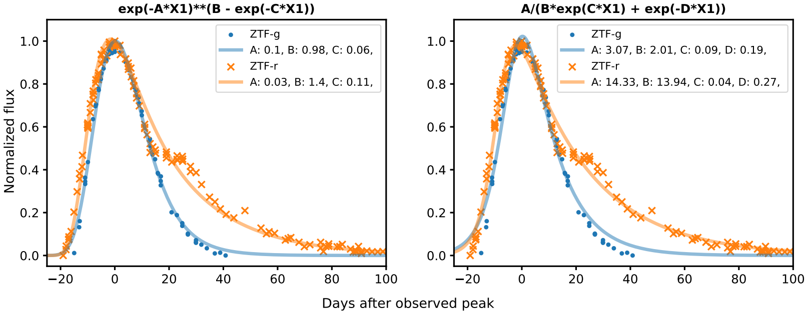


Models	Equation f(x)	$med(MSE)$	$MSE_{S\&P}$
Gaussian [?, ?]	$A \cdot \mathrm{e}^{-\frac{x^2}{B}}$	0.363	0.260
Laplace [?]	$A \cdot \mathrm{e}^{-B\|x\|}$	0.342	0.084
Cauchy [?]	$A \cdot B^2/(x^2 + B^2)$	0.305	0.079
Linear-Laplace	$(A - Bx) \cdot \mathrm{e}^{-C\|x\|}$	0.327	0.065
Exp-Laplace	$A \cdot \mathrm{e}^{Bx - C\|x\|}$	0.328	***0.063***
Power-Laplace	$A \cdot \mathrm{e}^{B\|x\|^C}$	***0.246***	0.075

Tabela 1: Best functions generated by MvSR. The last two columns respectively show the median MSE score of the functions fitted on individual normalized assets and the score when fitted on normalized the $S\&P500$ dataset. Bold numbers correspond to the best score of the column.

Tracking the change in radiation flux over time is an important tool to understand the physics of stellar objects.

One of the most studied types of extreme variability is supernovae (SNe), a catastrophic cosmic event caused by a star's explosion. The physical processes in a supernova are highly complex, and we may not yet be able to directly model individual supernova light curves.


exp(-A*X1)**(B - exp(-C*X1))
A/(B*exp(C*X1) + exp(-D*X1))
ZTF-g
A: 0.1, B: 0.98, C: 0.06,
ZTF-r
A: 0.03, B: 1.4, C: 0.11,
ZTF-g
A: 3.07, B: 2.01, C: 0.09, D: 0.19,
ZTF-r
A: 14.33, B: 13.94, C: 0.04, D: 0.27,
Normalized flux
Days after observed peak

UFABC

Equation f(t)	$< R^2 >$	No. parameters
$e^{-At\cdot(B-e^{-Ct})}$	0.990	3
$\frac{A}{(B\cdot e^{Ct}+e^{-Dt})}$	0.987	4
$\frac{A^{Bt}}{Ct+(-Dt+e^{Et})^2}$	0.992	5

Tabela 2: Summary of the best parametric functions generated using MvSR on SNIa lightcurves. The second column corresponds to the mean R^2 score over the 6 examples provided.


Regression
VS. THE WORLD
An epic of epic epicness.
WWW.SCOTTPILGRIMTHEMOVIE.COM

Conclusion

Conclusion

- Symbolic Regression can be a good alternative to clear and opaque models
- The current state-of-the-art is competitive with opaque models in accuracy
- With the analytical solutions, we can apply almost every statistical tool we do in linear regression models
- Can be adapted to your needs

- Preach about the good things of SR to scientific communities
- Create an easy-to-use tool similar to what we have in R and Python
- Make it easier to select between simpler or more accurate model from a set of choices

Questions?