Explaining potential epistasis in genomic data using symbolic representations of complex black box models

AAKARSH ANAND[1], PRATEEK ANAND[1], Boyang Fu[2], Sriram Sankararaman[2,3,4,5]

[1]BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA [2]Department of Computer Science, UCLA
[3]Department of Human Genetics, David Geffen School of Medicine, UCLA [4]Department of Computational Medicine, David Geffen School of Medicine, UCLA
[5]Bioinformatics Interdepartmental Graduate Program, UCLA

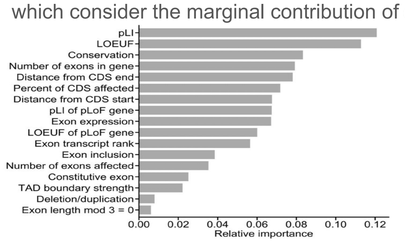
Abstract

Epistasis, known as the interaction among genetic variants, has long been hypothesized to play a major role in explaining missing heritability. Though recent studies have found many candidate variants demonstrating epistasis signals in the UKBiobank, it remains a controversial question how to interpret the findings. Nonlinear models have shown potential in capturing these signals, but we require additional explanation methods to understand the projected relationships. Here, we utilize symbolic pursuit, a form of symbolic regression that provides a closed-form, interpretable model which generalizes first order explanations. Furthermore, we extend this study by applying Taylor expansions to the model, balancing interpretability with performance while improving its generalizability. We found the method performed reliably and was consistent with other methods across a variety of simulated data. This work contains strong implications for its use on large genomic datasets and its ability to capture nonlinear interactions without prior knowledge of the genetic architecture.

Explanation Methods

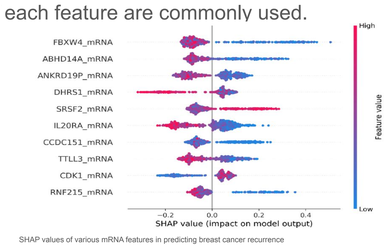
Explanation methods provide the user with importance values that convey how much each feature contributes to the output.

In critical domains such as medicine and defense, explanation methods are commonly used as a means of understanding the decisions made by black box models to gain trust in their underlying algorithms. For their ease of implementation, 1st order explanations which consider the marginal contribution of each feature are commonly used.



Importances of features used to predict pathogenicity of structural variants

StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants, 2020



SHAP values of various mRNA features in predicting breast cancer recurrence

Direct comparison shows that mRNA-based diagnostics incorporate information which cannot be learned directly from genomic mutations, 2020

By only considering marginal contributions, 1st order explanations are insufficient to explain nonlinear models. Some issues:

1. Unable to account for nonlinearity in data
2. Inaccurate, non-generalizable local fidelity
3. Misleading feature importances

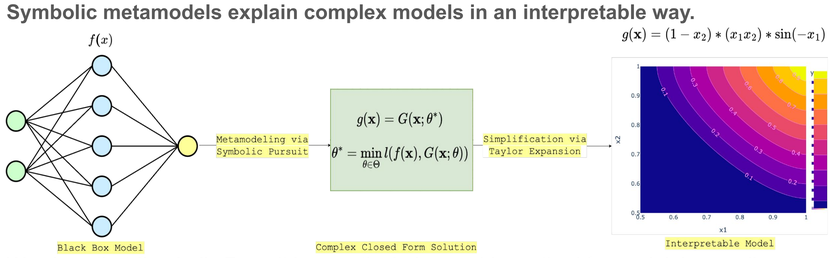
Symbolic Pursuit (SP)

Why can't we modify existing methods to explain nonlinear models?

For example, using the gold standard (SHAP), can we manually construct "interaction" features and obtain their importances? There are two problems with this approach:

1. Which features are interacting with each other, and to what extent?
   - In the real world, we have no prior knowledge of the data, so this is impossible.
2. How will this affect runtime?
   - Adding custom features will cause scalability issues, as SHAP values require iterating through the power set of size 2^d feature coalitions to obtain fair rankings.

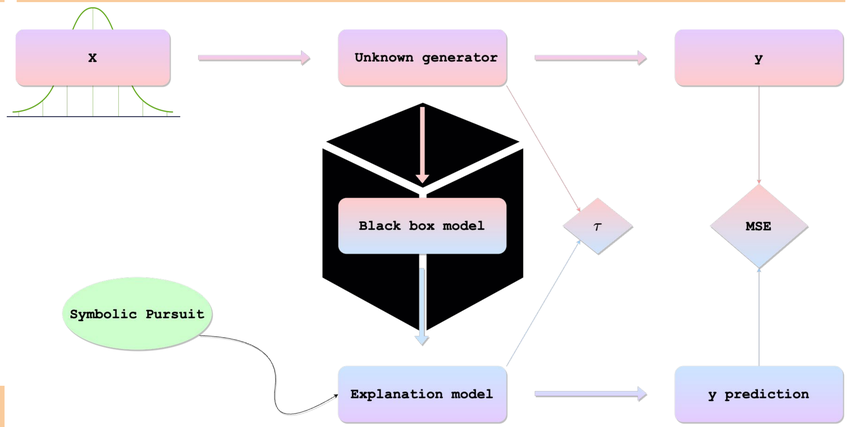
Symbolic metamodels explain complex models in an interpretable way.



We show that Symbolic Pursuit is a particularly capable method for explaining nonlinear models. This algorithm:

1. Does not require prior knowledge of the data or its relationship to the output
2. Provides a closed-form equation that explains interactions unlike 1st order explanations
3. Utilizes a broader hypothesis class

Methodology



Simulated Data (used to train 200 models for each order across which all results are averaged)

X = (100 instances, 3 features) drawn from random distribution 0 - 1
β drawn from random distribution 0 - 1

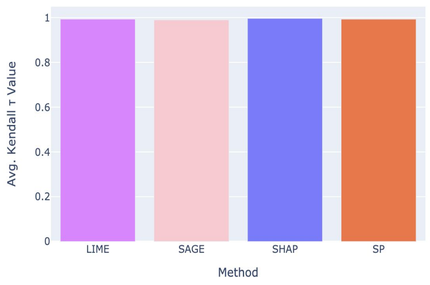
1. 1st order: $y = X\beta$
   - β: length 3
2. 2nd order: $y = X_{trans}\beta$
   - X_{trans}: degree 2 polynomial transform of X
   - β: length 10

Metrics of Interest

Model performance measured using:

1. Mean Squared Error (MSE): between SP predictions and targets
2. Kendall Tau Correlation (τ): correlation between feature importances and known weights (β)

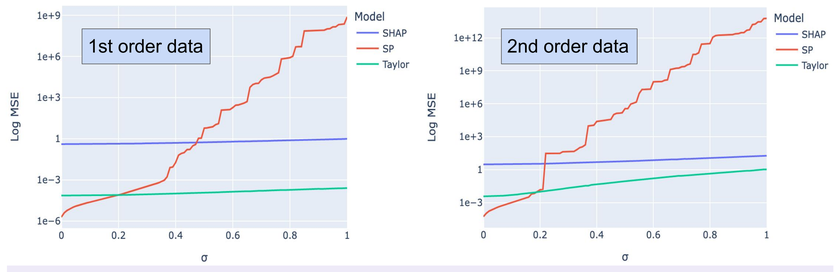
Linear Comparison



SP provides consistent, accurate explanations of linear data.

To verify the potential of SP, we first compare its performance to other notable linear explainers on a 1st order dataset. We note a high average correlation between original and predicted importances for all explainers across 200 trials.

SP Taylor Expansions Improve Generalizability



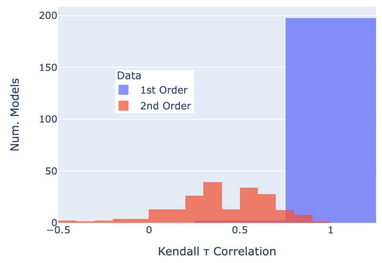
Taylor expansions control the method's overfitting and improve generalizability.

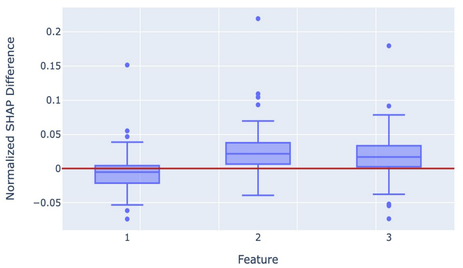
While SP importance attributions effectively model data similar to its training environment, the complex nature of Meijer-G functions often causes the final expression to be overfitted, unintuitive, and non-generalizable. This does not help with the real world case where we wish to explain instances similar to those already learned:

$$X_{new} = X_{learned} + \varepsilon$$

We perturb our input data as $X_{noise} = X + N(0, \sigma)$ and observe the average performance of taylor expansions vs. the original SP model. We note a significant decrease in MSE of Taylor Expansions for increasing noise across 1st and 2nd order data, indicating high generalizability.

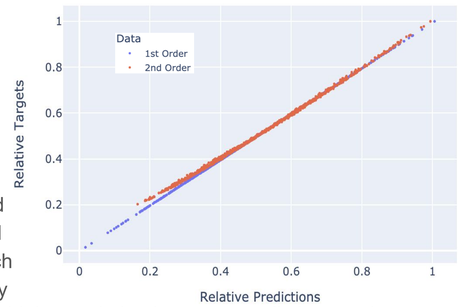
SP Taylor Expansion Performance on Higher Order Data



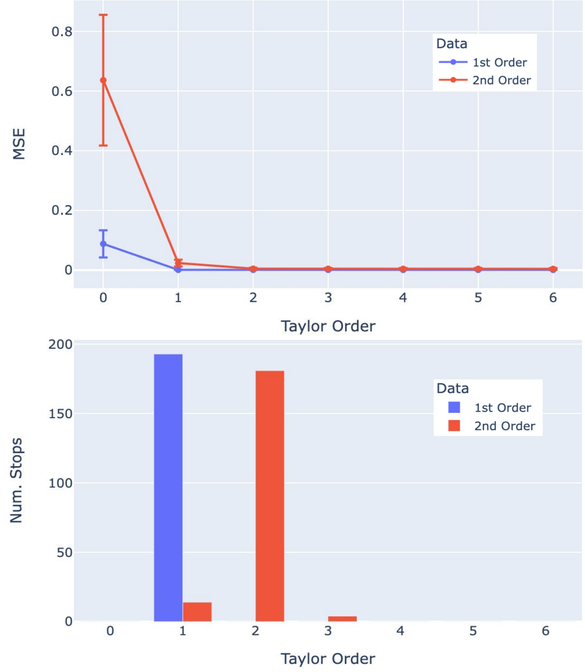


SP Taylor Expansions may reveal alternate, viable solutions.

Correlation experiments yielded less favorable τ values for 2nd order expansions. However, further comparisons of the difference between SHAP attributions for the original 2nd order data and its SP expansion yielded results close to 0, indicating that SP still captured the overall importances of each feature in the data, but in a different way from how it was originally modeled. The high correlation between predictions and targets further corroborates this finding. In real world settings, true coefficients are not known, and there are multiple solutions to model complex data. In this regard, SP's low τ values are of little consequence.



SP Taylor expansions match the characteristics of the data.



Higher Taylor orders result in decreasing MSE. However, the difference is negligible past the original order of the data. We also observe the optimal stopping point for each type of data: the Taylor expansion that has the lowest MSE where the subsequent order causes < 50% improvement. All stopping points were order 1 for 1st order data, and most were order 2 for 2nd order data. These results provide hopeful implications for SP's performance on data with unknown distributions, as iteratively taking Taylor expansions will allow one to find the simplest expression that accurately models the data as well.

Conclusions/Future Directions

The above findings show theoretical promise in the underlying strategies of SP. However, the impact on performance when analyzing extremely large datasets remains to be seen. As the high dimensionality of genetic data could cause problems for SP, one idea we hope to explore is using an ensemble of neural networks to summarize groups of features before generating explanations. Regardless, the current performance of SP shows strong potential for effective complex model interpretations which have never been seen before.

References

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
Covert, Ian, Scott M. Lundberg, and Su-In Lee. "Understanding global feature contributions with additive importance measures." *Advances in Neural Information Processing Systems* 33 (2020): 17212-17223.
Alaa, Ahmed M., and Mihaela van der Schaar. "Demystifying black-box models with symbolic metamodels." *Advances in Neural Information Processing Systems* 32 (2019).
Crabbe, Jonathan, et al. "Learning outside the black-box: The pursuit of interpretable models." *Advances in Neural Information Processing Systems* 33 (2020): 17838-17849.