Northwestern University

Ab Initio Nonparametric Variable Selection for Scalable Symbolic Regression with Large p

Shengbin Ye [1,2] Meng Li [1]

[1]Department Statistics, Rice University

[2]Department of Statistics and Data Science, Northwestern University

RICE

Motivation

SR Algorithm		p	Has Irrelevant X
Transformer			
TPSR	(2023)	9	❌
RNN			
DySymNet	(2024)	9	❌
uDSR	(2022)	9	❌
DSR	(2021)	2	❌
Divide-and-conquer			
AIFeynman 2.0	(2020)	9	❌
Genetic Programming			
PySR	(2023)	6	❌
Operon	(2020)	5	❌
⋮		⋮	⋮

- Symbolic regression (SR) is **NP-hard**
- Most focus on **low-dimensional** problems (e.g., $p \leq 10$)
- Unrealistic settings: lack of irrelevant predictors
- No existing high-dimensional SR benchmark

PAN+SR: pre-screening framework for high-dimensional SR

Output y often depends on a **subset** $\mathcal{S}_0 \subseteq \{1, \ldots, p\}$ of p_0 relevant predictors:

$$y = f(\boldsymbol{X}) = f(\boldsymbol{X}_{\mathcal{S}_0}),$$

where $p_0 = |\mathcal{S}_0| \ll p$.

Building on the Parametrics Assisted by Nonparametrics (PAN) framework in Ye et al. (JASA, 2024), we propose the PAN+SR framework.

Main Idea

1. Nonparametric variable selection: $\boldsymbol{X} \mapsto \boldsymbol{X}_{\mathcal{S}}$
   - large $p \implies$ small p
   - SR search space ↓↓↓
2. Perform SR on low-dimensional dataset $(y, \boldsymbol{X}_{\mathcal{S}})$

PAN Criterion

Step 1 must select all $\mathcal{S}_0$: $\boxed{\mathcal{S} \supseteq \mathcal{S}_0}$

BART VIP Rank

We propose a novel Bayesian Additive Regression Tree (BART)-based variable selection method: BART VIP Rank

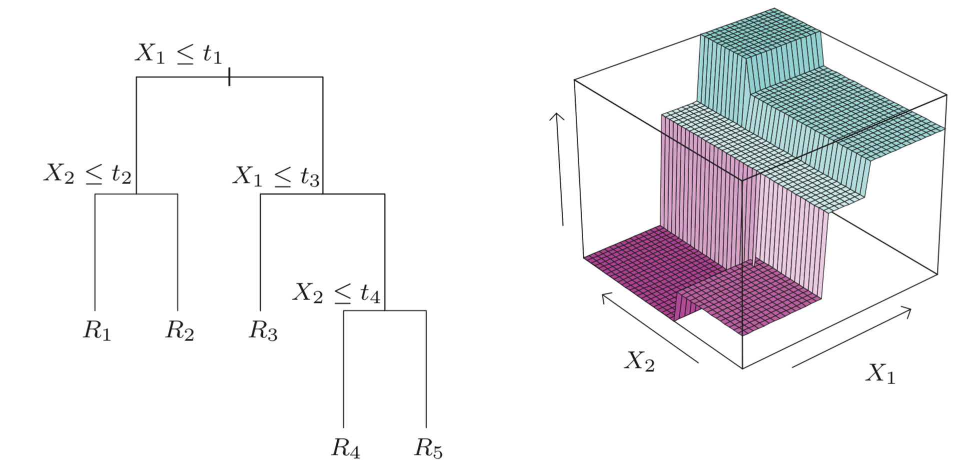

Figure 1. Visualization of BART.

Variable Inclusion Proportion (VIP)

A typical variable importance measure in BART is variable inclusion proportion (VIP):

$$q_j = \frac{1}{K}\sum_{k=1}^{K} \frac{c_{jk}}{c_{\cdot k}} \qquad \text{(avg prop of splits on } x_j\text{)}$$

- Arbitrary scale: how large is large?
- Tight range: $0 \leq q_j \leq 1$
  - Small perturbation in threshold $\implies$ different selections (sensitivity)

VIP Rank

Fit $L = 20$ independent BART models. Let $q_{j,\ell}$ = VIP of x_j in the ℓth fit, and let $R(q_{j,\ell})$ = ranking of $q_{j,\ell}$ within fit ℓ. Define the VIP Rank for x_j as the average ranking over L model fits:

$$\overline{R}_j = \frac{1}{L}\sum_{\ell=1}^{L} R(q_{j,\ell}).$$

Under mild assumptions,

$$\overline{R}_j = \begin{cases} (1+p_0)/2, & \text{if } x_j \text{ is relevant} \\ (p_0+1+p)/2, & \text{otherwise} \end{cases}$$

Say $p_0 = 4$ and $p = 204$. Then, $\overline{R}_j = 2.5$ if x_j is relevant vs. $\overline{R}_j = 104.5$ otherwise.

Algorithm

1. Fit $L = 20$ independent BART models on $(y, \boldsymbol{X})$
2. Calculate BART VIP Rank $\overline{\boldsymbol{R}} = (\overline{R}_1, \ldots, \overline{R}_p) \in \mathbb{R}^p$
3. Apply Agglomerative Hierarchical Clustering on $\overline{\boldsymbol{R}}$
4. Cut dendrogram to form 2 clusters: $\mathcal{C}_{\text{low}}$ and $\mathcal{C}_{\text{high}}$
5. Select x_j if $\overline{R}_j \in \mathcal{C}_{\text{low}}$

High-Dimensional SR Benchmark

We design a high-dimensional SR benchmark using 22 real-world datasets from PMLB and 100 synthetic datasets based on *Feynman Lectures on Physics*.

	SRBench	✨ New Benchmark
Irrelevant predictors $\mathcal{S}_1$	❌	✅
No. of predictors p	2 to 9	102 to 459
No. of relevant predictors p_0	2 to 9	2 to 9
Sample size n	100,000	500, 1000, 1500, 2000
Signal-to-noise ratio (SNR)	10 to noiseless	0.5 to noiseless
No. of trials per setting	10	10

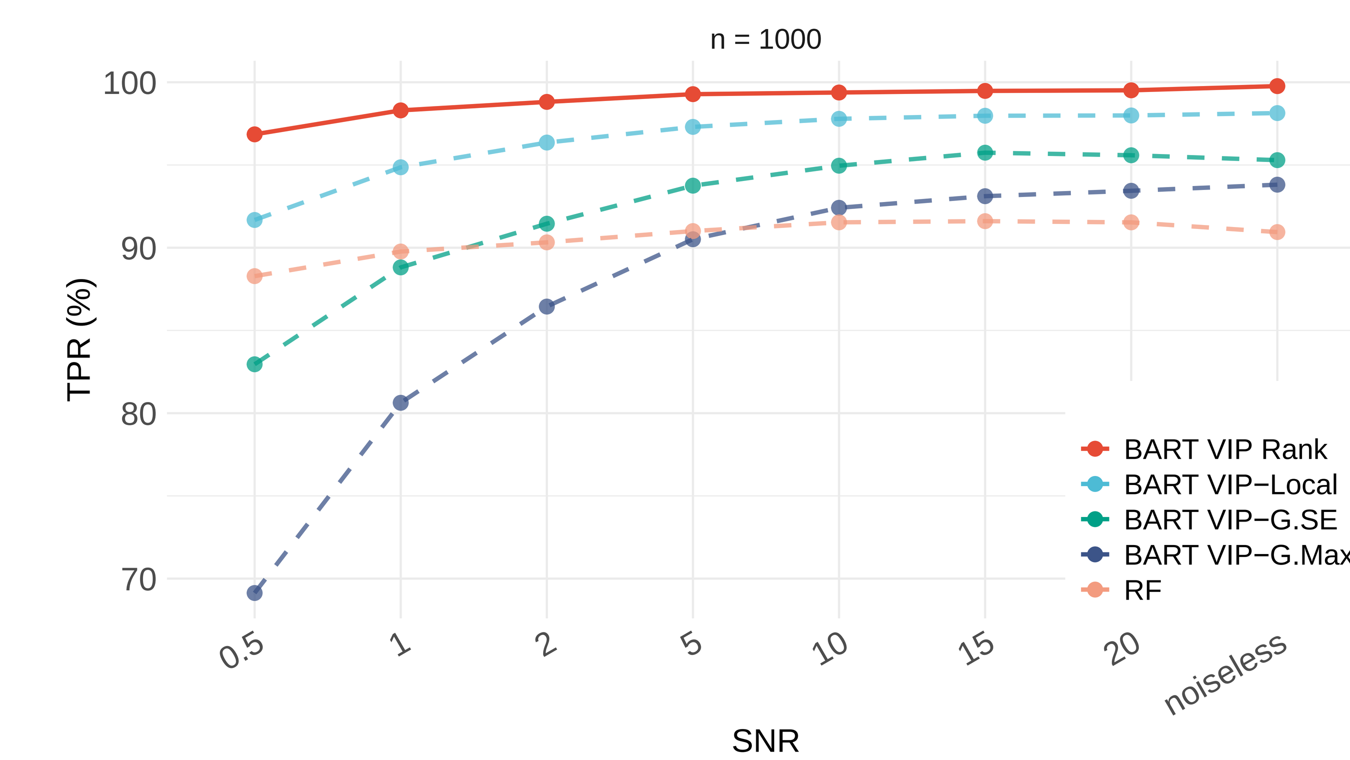

Figure 2. True positive rates on high-dimensional Feynman datasets.

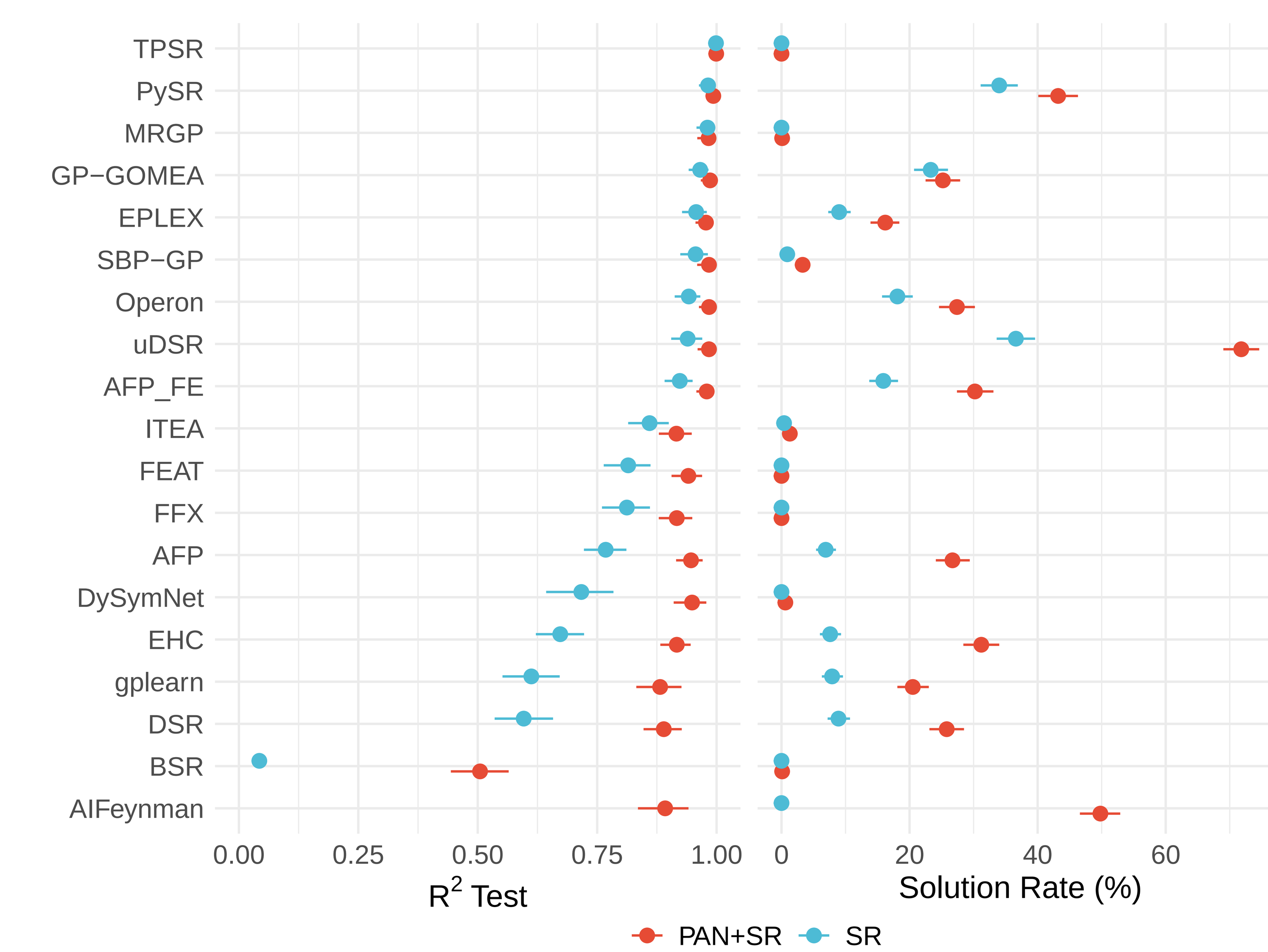

Figure 3. Performance of PAN+SR vs standalone SR on high-dimensional Feynman datasets.