Beyond Additivity: Sparse Isotonic Shapley Regression toward Nonlinear Explainability

Jialai She

Phillips Academy

Acknowledgment: Thanks to Dr. Alterovitz, Dr. Xie, Dr. Pei, and the PRIMES program

Background and Motivation

- Shapley values: fair credit in cooperative games (Shapley 1953).
  - Feature set: $F = \{1, \ldots, p\}$; coalition payoff $\nu_A = \nu(A)\ \forall A \subseteq F$.
  - Shapley value β_j quantifies fair share/importance of feature j
  - V_A: link between observed payoffs and underlying feature contributions: $\nu_A = V_A(\{\beta_j\}_{j \in A}) + \text{noise}$
- Widely used for **feature attribution** in explainable AI (**XAI**)
  - Regression: p-values. Complex '**black-box**' *trees/neural nets*?
- Challenge a): **Additivity** of payoffs: $V_A(\{\beta_j\}_{j \in A}) = \sum_{j \in A} \beta_j$
  - Diverse payoff methods (R^2, TreeSHAP, Sobol indices, etc.) exist; Shapley is applied *blindly* without checking its **axioms**!
  - Fryer (21)'s "taxicab" payoff $V_A = \max_{j \in A} \beta_j$ (*winner-takes-all*)
- Challenge b): **Dense**, non-sparse attributions: $\beta \in \mathbb{R}^p$, p large
  - Most features negligible. Dense Shapley + thresholding = greedy!
- Need **unified** "nonlinear + sparse" attribution.

Recasting Shapley: A Statistical Framework

- A weighted least squares formulation recovers the Shapley values:

$$\min_{\beta,c} \sum_A w_{\text{SH}}(A)(\nu_A - \sum_{j\in A}\beta_j - c)^2 \text{ s.t. } c = \nu_\emptyset, c + \sum_{j=1}^{p}\beta_j = \nu_F,$$

  where $w_{\text{SH}}(A) = \frac{p-1}{\binom{p}{|A|}|A|(p-|A|)}$. See Lundberg & Lee (17).

- Under $w_{\text{SH}}(\emptyset) = +\infty, w_{\text{SH}}(F) = +\infty$, $\nu_A \leftarrow \nu_A - \nu_\emptyset$, we obtain

$$\nu_A \sim \mathcal{N}(\mu_A, \sigma_A^2), \mu_A = \sum_{j\in A}\beta_j^*, \sigma_A^2 \propto \binom{p}{|A|}|A|(p-|A|)\,(\propto \frac{1}{w_{\text{SH}}(A)})$$

- Many payoff functions violate the **Gaussian** assumption due to range constraints, skewness, heavy tails, & heterogeneity.
- We propose to consider $T(\nu_A) \sim \mathcal{N}(\sum_{j\in A} T(\beta_j^*), \sigma_A^2)$.

Univariate T-Mappings for **Nonlinear** Payoffs

- Our proposal restores additivity in a *transformed* domain using a univariate, invertible function $T(\cdot)$:

$$V_A(\{\beta_j\}_{j\in A}) = T^{-1}(\sum_{j\in A} T(\beta_j)).$$

- Simple T-mapping captures rich, **multivariate** payoff structures.
- Special case: $T(x) = |x|^d$, then $V_A = \|\beta_A\|_d$.
  - $d = 1, 2$: ℓ_1/ℓ_2-ball. $d \to \infty$: $V_A = \max_{j\in A} |\beta_j|$ (winner-takes-all)
- Exponential/log/odds transforms for other nonlinearities.
- Nicely, our method will learn $T(\cdot)$ in a purely *data-driven* manner, and **bypass** the need to specify its analytical form.

Sparse Isotonic Shapley Regression (SISR)

$$\min_{\beta, T(\cdot)} \sum_{A \subseteq 2^F} w_{\mathrm{SH}}(A)\{T(\nu_A) - \sum_{j \in A} T(\beta_j)\}^2$$

$$\text{s.t. } \|\beta\|_0 \leq s, T \in \mathcal{M}, \sum_{j=1}^{p} (T(\beta_j))^2 = 1$$

- **Monotonicity**: $\mathcal{M}$ constrains T to be increasing $\rightarrow$ preserve ranking of feature importance: $\beta_i \geq \beta_j \Rightarrow T(\beta_i) \geq T(\beta_j)$.
- **Normalization**: The scale constraint prevents degeneracy and anchors the model scale (& yields a closed form update for β!)
- **Sparsity**: We enforce explicit support control via ℓ_0, selecting the most relevant features during optimization, not after.
  - $\lambda \sum |T(\beta_j)|$: over-shrinks and requires a fine λ grid search. s: an *upper bound*, simple to specify & tune (**RIC**, Foster & George 94)

Equivalent Reformulations

- **Challenges**: (i) functional T, (ii) double nonconvex constraints
- Let $\gamma_j = T(\beta_j)$. As $T(0) = 0$, the optimization problem becomes

$$\min_{\gamma, T} \sum_{A \subseteq 2^F} w_{\text{SH}}(A)\Big(T(\nu_A) - \sum_{j \in A} \gamma_j\Big)^2 \text{ s.t. } \|\gamma\|_0 \le s, \|\gamma\|_2 = 1, T \in \mathcal{M}$$

- As $T(\cdot)$ is only evaluated at observed $\nu_A (A \subseteq 2^F)$, we **discretize** the problem by introducing $t = \big[T(\nu_A)\big] \in \mathbb{R}^{2^p}$, $\nu = [\nu_A]$, $\delta = [\sum_{j \in A} \gamma_j] = Z\gamma \in \mathbb{R}^{2^p}$ with $Z \in \mathbb{R}^{2^p \times p}$ the 'incidence matrix'
- Introducing $E(\nu) = \{(i,j) : \nu_i \le \nu_j\}$ to encode the pairwise ordering and $W = \text{diag}\{w_{\text{SH}}(A)\}_{A \subseteq 2^F}$, it suffices to solve

$$\begin{aligned} \min_{\gamma, t} \frac{1}{2}(t - \delta)^\top W (t - \delta) \text{ s.t. } & \delta = Z\gamma, \\ & \|\gamma\|_0 \le s, \ \|\gamma\|_2 = 1, \ t_i \le t_j \ \forall (i,j) \in E(\nu) \end{aligned}$$

Two-Block Alternating Optimization

- For $\vec{t}$, the problem reduces to a *weighted* isotonic regression (de Leeuw et al 09). We solved it by an efficient *stack*-based weighted variant of Pool-Adjacent-Violators Algorithm (**PAVA**).
- For γ, we use a "surrogate function" trick to deal with δ, leading to an *iterative* **normalized** hard-thresholding algorithm:

$$\gamma^{\text{new}} = \frac{\mathcal{H}(y; s)}{\|\mathcal{H}(y; s)\|_2}, \quad \text{where } y = \gamma^{\text{old}} - \frac{1}{\rho} Z^\top W (Z\gamma^{\text{old}} - t).$$

where $\mathcal{H}(\cdot; s)$ keeps the s largest entries of y, sets others to zero.

A Global Convergence Theorem

- We rigorously **proved** the algorithm has global convergence
- Let $\rho \geq \|Z^\top W Z\|_2$, with $\|\cdot\|_2$ the matrix **spectral norm**.
- For any initial point $\gamma^{(0)}$ satisfying $\|\gamma^{(0)}\|_0 \leq s$ and $\|\gamma^{(0)}\|_2 = 1$, the sequence $\{\gamma^{(k)}\}$ generated by the SISR algorithm produces non-increasing (and thus convergent) function values:

$$l(\gamma^{(k+1)}) \leq l(\gamma^{(k)}) \quad \text{for all } k \geq 0.$$

- Furthermore, if $\rho > \|Z^\top W Z\|_2$, $\|\gamma^{(k+1)}) - \gamma^{(k)}\|_2 \to 0$ as $k \to \infty$.

SISR Algorithm Outline

Input: $\nu = [\nu_A]_{A \subseteq 2^F} \in \mathbb{R}^{2^p}$ (baseline-adjusted, such that $\nu_\emptyset = 0$), sparsity level s, initial vector $t^{(0)} \in \mathbb{R}^{2^p}$, and the design matrix $Z \in \mathbb{R}^{2^p \times p}$ and diagonal weight matrix W as constructed before.

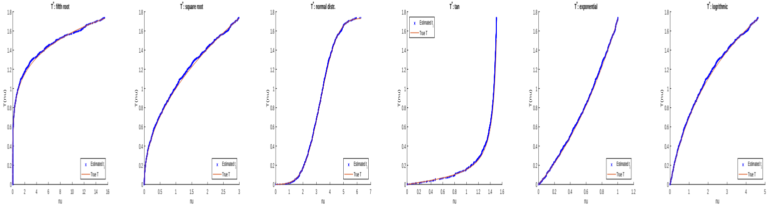
Initialize $t \leftarrow t^{(0)}$, $\gamma \leftarrow 0$
$\rho \leftarrow \|Z^\top W Z\|_2$
repeat
  while not converged **do**
    $\xi \leftarrow \mathcal{H}(\gamma - \frac{1}{\rho} Z^\top W (Z\gamma - t); s)$
    $\gamma \leftarrow \frac{\xi}{\|\xi\|_2}$
  end while
  $\delta \leftarrow Z\gamma$
  Update t by fitting a weighted isotonic regression (W, Z) to δ
until convergence
return t, γ

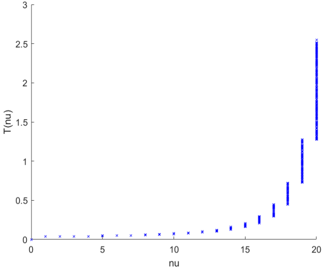
Data Insights

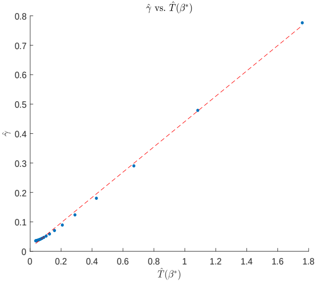
- **Domain adaptation:** Strong evidence that SISR accurately recovers the underlying transformation $\hat{T}$.
- **Sparsity recovery:** Robust SISR support recovery even in challenging settings; lower sparsity yields faster computation.
- **Feature dependence and irrelevance:** Correlation drives curvature; sparsity yields piecewise behavior with distinct-slope segments. (Even without correlation, irrelevant features can distort raw worths, undermining additivity.)
- **Prostate data:** Plain Shapley can assign spurious importance with correlated or irrelevant predictors; SISR's importance aligns with corroborating diagnostics.
- **Boston housing:** Standard Shapley ranks/signs shift with payoff scaling; SISR remains stable, preserving interpretation.

Domain Adaptation

- SISR accurately recovers the true T^* for various **nonlinear** forms (top: 5th root, sqrt, normal, tan, exp, log; bottom: max)





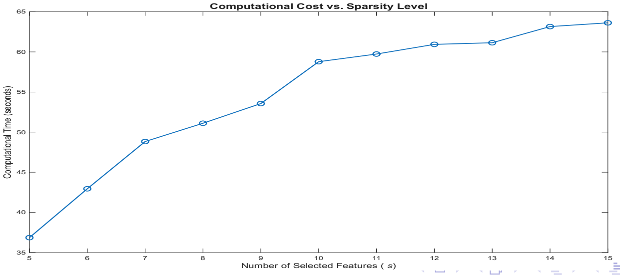


Sparsity Recovery

- **Affn**: $\langle \hat{\gamma}, \gamma^* \rangle \times 100$, affinity to measure estimation accuracy
- **Supp**: $|\text{supp}(\hat{\gamma}) \cap \text{supp}(\gamma^*)|/s^* \times 100\%$, support recovery rate for correct feature identification. (Tuning: RIC)
- The support recovery rate remains surprisingly strong even under challenging cases—SISR consistently identifies the correct features

noise scale	5e-3		1e-2		5e-2		1e-1		2e-1	
	Affn	Supp	Affn	Supp	Affn	Supp	Affn	Supp	Affn	Supp
$p = 10$	99.6	100%	99.5	100%	97.9	100%	88.7	98.7%	66.2	80.7%
$p = 15$	99.9	100%	97.8	100%	79.9	100%	70.9	98.0%	57.6	73.3%
$p = 20$	87.9	100%	80.3	100%	68.9	100%	63.2	96.0%	54.3	65.3%
$p = 25$	74.0	100%	70.5	100%	65.5	100%	60.6	90.7%	52.1	62.0%

- Lower sparsity levels generally lead to faster computation



R^2-Payoffs

- Standard R^2-constructed payoffs fail to satisfy Shapley!
  - Both correlated and irrelevant features induce strong nonlinearity
  - **Correlation** drives curvature; **sparsity** introduces breaks.
- SISR adaptively learns the transformation needed to calibrate payoffs and restore additivity. $(\theta : x_i \sim N(0, \Sigma), \Sigma_{ij} = \theta^{|i-j|})$

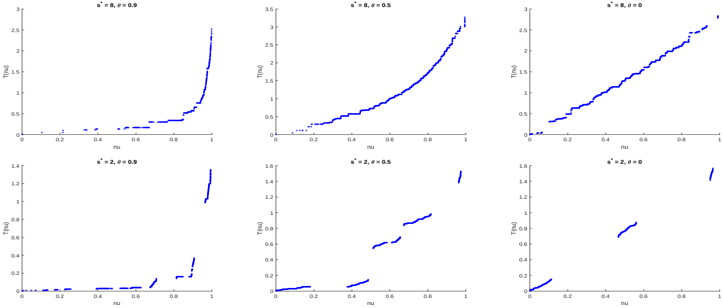


Figure: Estimated $\hat{T}(\nu)$ across varying **sparsity** levels ($s = 8, 2$, top to down) and feature **correlation** strengths ($\theta = 0.9, 0.5, 0$, left to right).

Case Study: Prostate Cancer

- Response: log(cancer volume) (`lcavol`). Predictors: age (`age`), seminal vesicle invasion (`svi`, binary), log(capsular penetration) (`lcp`), log(prostate specific antigen) (`lpsa`), and so on.

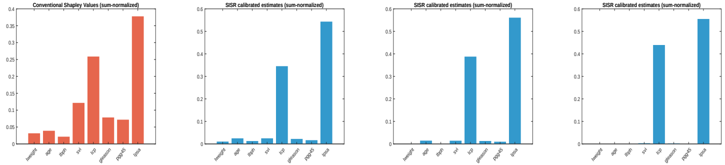


Figure: Raw Shapley (left) and SISR-calibrated values ($s = 8, 6, 4$); RIC identifies $s = 6$ as optimal.

- Both methods agree that `lcp` and `lpsa` are dominant
- `svi`: Naive Shapley ranks it 3rd ($> 10\%$), while SISR-**calibrated** $\hat{\gamma}$ gives it nearly **zero**.
  - Independent checks: Stepwise AIC/BIC both exclude `svi`; very last variable selected by LASSO; p-value $= 0.6$ in the full model.

Case Study: Boston Housing

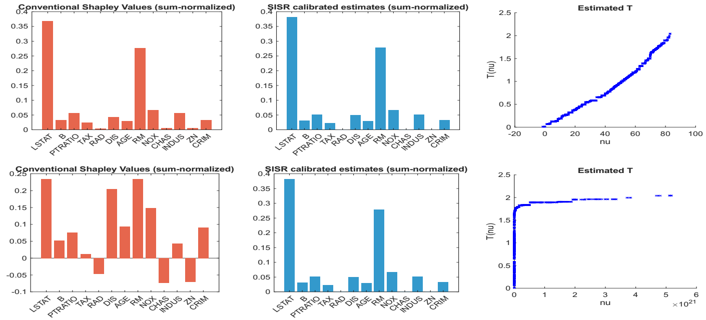


Figure: Top: MSE payoff, bottom: Exponential payoff

- Trained XGBoost to predict median home value
- Standard Shapley values are highly **sensitive** to the payoff scale!
  - The importance of `DIS` increases from minor to leading
  - `CHAS` and other variables even receive **negative** attributions
- In contrast to sign and rank changes, SISR remains **robust**, learning a nonlinear transformation that yields stable attributions

Limitations

- **Scalability**: Coalition-level objects live in dimension 2^p ($t \in \mathbb{R}^{2^p}$ and $Z \in \mathbb{R}^{2^p \times p}$), and the isotonic step runs over all subsets.
  - In practice, handling $p > 25$ is difficult without randomized/approximate strategies.
- **Outlier sensitivity:** Extreme coalition payoffs (ν_A) can contaminate the estimation of T
  - They can bend $\hat{T}$ toward extremes and biasing attributions
- **No uncertainty quantification:** We only provide point estimates for T and γ with no standard errors or CIs.

Future Directions

- **Randomized PAV:** Develop an approximate, streaming/distributed isotonic solver for $T(\cdot)$.
- **Robust SISR**: Replace the quadratic loss with *robust* alternatives (e.g., L1, Huber loss) to learn a transformation that is more resilient to data imperfections.
- **Payoff Pooling**: *Integrate* diverse payoff specifications to enhance stability of feature rankings, and construct an aggregated surrogate model

Conclusions and Contributions

- We introduced **SISR**, the first unified framework to address nonlinearity and sparsity in Shapley-based explanations.
- First to show that irrelevant features and inter-feature dependencies can induce nonlinearity in payoff structures.
- It learns a data-driven payoff transformation nonparametrically (via PAVA) *without* a predefined basis expansion or other parametric representation, and enforces sparsity inherently.
- Simple and efficient closed-form updates with **theoretical** convergence guarantees.
- Extensive experiments show SISR stabilizes attributions and correctly identifies relevant features, avoiding **rank/sign** distortions common in standard Shapley values, offering an interpretable framework for complex XAI attribution tasks.

Acknowledgements

I would like to express my deepest gratitude to my mentor, **Dr. Gil Alterovitz**, for his unwavering support and encouragement.

I am also sincerely grateful to **Dr. Ning Xie** and **Pei Shaojun** for their insightful suggestions and coordination throughout this project.

Finally, my thanks to the **PRIMES program** for providing this wonderful research opportunity.