S.T. Yau High School Science Award

Research Report

The Team

Name of team member: Jialai She
School: Phillips Academy Andover
City, Country: Andover, United States of America

Name of team member:
School:
City, Country:

Name of team member:
School:
City, Country:

Name of supervising teacher: Dr. Gil Alterovitz
Job Title: Associate Professor
School/Institution: Harvard University
City, Country: Boston, United States of America

Title of Research Report

Beyond Additivity: Sparse Isotonic Shapley Regression toward Nonlinear Explainability

Date

08/19/2025

Beyond Additivity: Sparse Isotonic Shapley Regression toward Nonlinear Explainability

Jialai She

Abstract

Modern economic and financial analysis increasingly relies on complex models to capture nonlinear, high-dimensional, and sequential patterns in macroeconomic and market data. This complexity has outpaced existing explainability tools, creating a pressing need for robust, interpretable methods that can identify the true drivers of economic outcomes and inform policy or investment decisions. Shapley values, a widely used standard for feature attribution in explainable AI, face two key limitations in this context. First, the canonical Shapley framework assumes an additive worth function, yet real-world payoff structures—shaped by non-Gaussian distributions, heavy tails, feature dependence, or domain-specific loss scales—often violate this assumption, resulting in distorted attributions. Second, achieving sparsity by computing dense Shapley values and then applying ad hoc thresholding often leads to inconsistent results and is computationally demanding.

We propose Sparse Isotonic Shapley Regression (SISR), a unified framework for nonlinear, sparse model interpretation. SISR simultaneously learns a monotonic transformation to restore the additivity required by Shapley theory, without requiring a specified functional form, and enforces exact sparsity through L0 regularization. SISR's optimization leverages Pool-Adjacent-Violators for efficient isotonic regression and iterative normalized hard-thresholding for support selection, with theoretical guarantees for global convergence.

Empirical studies show that SISR accurately recovers both the true transformation and support, even in noisy settings. Notably, our results reveal for the first time that correlated or irrelevant features can induce substantial deviations from linearity in the payoff—a phenomenon common in economic applications. Across regression, classification, neural networks, and tree ensembles, SISR stabilizes attributions across payoff schemes and reliably filters out irrelevant features, while standard Shapley values often suffer severe rank and sign distortions. By jointly estimating nonlinear transformations and enforcing sparsity, SISR provides a theoretically grounded and practical approach to interpretable machine learning in economics.

Keywords: Shapley value, machine learning explainability, isotonic regression, sparsity pursuit

Acknowledgement

I wish to thank Dr. G. Alterovitz for his high-level advising on this project, which was conducted as part of the MIT PRIMES program, and his potdoc Dr. S. Pei for his encouraging feedback on an early draft. While I am grateful for their guidance, all ideas, topic selection, methodology, theoretical results, computation, implementation, analyses, and writing were conceived and executed solely by the author. I am also thankful to Dr. A. Owen, who first introduced me to some basic concepts of Shapley values early in my studies, which sparked my interest in this area.

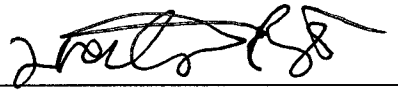
Commitments on Academic Honesty and Integrity

We hereby declare that we

1. are fully committed to the principle of honesty, integrity and fair play throughout the competition.
2. actually perform the research work ourselves and thus truly understand the content of the work.
3. observe the common standard of academic integrity adopted by most journals and degree theses.
4. have declared all the assistance and contribution we have received from any personnel, agency, institution, etc. for the research work.
5. undertake to avoid getting in touch with assessment panel members in a way that may lead to direct or indirect conflict of interest.
6. undertake to avoid any interaction with assessment panel members that would undermine the neutrality of the panel member and fairness of the assessment process.
7. observe the safety regulations of the laboratory(ies) where we conduct the experiment(s), if applicable.
8. observe all rules and regulations of the competition.
9. agree that the decision of YHSA is final in all matters related to the competition.

We understand and agree that failure to honour the above commitments may lead to disqualification from the competition and/or removal of reward, if applicable; that any unethical deeds, if found, will be disclosed to the school principal of team member(s) and relevant parties if deemed necessary; and that the decision of YHSA is final and no appeal will be accepted.

(Signatures of full team below)

X 

Name of team member:

X _____

Name of team member:

X _____

Name of team member:

X Gil Alterovitz Shaojun Pei

Name of supervising teacher:

Declaration of Academic Integrity

The participating team declares that the paper submitted is comprised of original research and results obtained under the guidance of the instructor. To the team's best knowledge, the paper does not contain research results, published or not, from a person who is not a team member, except for the content listed in the references and the acknowledgment. If there is any misinformation, we are willing to take all the related responsibilities.

Names of team members: Jialai She

Signatures of team members: 

Name of the instructor: Gil Alterovitz Shaojun Pei

Signature of the instructor: Gil Alterovitz Shaojun Pei

Date: 8/18/2025

Contents

1 Introduction and Motivation

Let $F = \{1, 2, \ldots, p\}$ denote the set of p features, and let $\nu : 2^F \to \mathbb{R}$ be the characteristic function or payoff function, with $\nu(A)$ representing the contribution or worth generated by a subset $A \subseteq F$ of features working together (often referred to as a *coalition* in game theory). A central question in economics and cooperative game theory is how to fairly allocate the value of a coalition to its individual members. The *Shapley value* (Shapley, 1953), a concept from Nobel laureate Lloyd Shapley, offers a theoretically grounded solution by assigning payoffs according to each member's average marginal contribution across all possible subsets.

In this paper, we denote the Shapley value for feature j by β_j for $1 \leq j \leq p$, quantifying the fair share or importance of feature j; for a subset $A \subseteq F$ we define β_A as the vector $[\beta_j]_{j \in A} \in \mathbb{R}^{|A|}$. For brevity, we also write ν_A as shorthand for $\nu(A)$ for any subset $A \subseteq F$. Shapley values establish a connection between the payoff function $\nu(A)$ and the underlying model parameters β_A. To make this dependence explicit, we introduce a function $V(\beta_1, \ldots, \beta_p; A)$, also denoted by $V_A(\{\beta_j\}_{j \in A})$, characterizing the deterministic, *noise-free* contribution associated with subset A:

$$\nu_A \sim V_A(\{\beta_j\}_{j \in A}), \tag{1}$$

where $\sim$ denotes approximate equality up to noise, a convention adopted throughout the paper.

In recent years, Shapley values have attracted substantial attention in machine learning, particularly in the field of *Explainable AI* (**XAI**) (Ancona et al., 2019). While assessing variable importance in simple regression is straightforward using traditional tools like T-tests and p-values, this task becomes a formidable challenge for the complex, "black-box" models now widely used to analyze sequential data in economics and finance. For sophisticated models—ranging from *tree-based ensembles* like random forests and boosted trees to *deep neural networks* like Long Short-Term Memory (LSTM) networks—standard inference methods are no longer applicable, making interpretation notoriously difficult.

Shapley values provide a model-agnostic framework for attributing predictive performance to individual features by casting feature contributions as a cooperative game. Specifically, for a prediction model $f(x)$, where $x \in \mathbb{R}^p$, researchers first design a payoff function ν_A over subsets $A \subseteq F$ to quantify the model's performance when using only the features in A. This reframes the explanation task as a "**credit allocation**" problem, enabling the use of Shapley values to quantify feature importance, and perhaps more importantly, to construct interpretable *surrogate models* based on restricted feature sets.

However, standard Shapley-based methods also face several limitations that restrict their practical utility in complex modeling scenarios.

(i) Moving beyond additive frameworks: Given a prediction model f, various methods have been proposed to construct the payoff function ν_A for Shapley-value analysis. (a) A fundamental approach involves retraining the model on every subset of features A and defining ν_A based on the reduction in statistical accuracy (such as R^2 in regression) (Lipovetsky and Conklin, 2001; Covert et al., 2020). However, this may be computationally prohibitive for modern AI models due to its exponential cost. (b) To circumvent retraining, Lundberg and Lee (2017) approximates ν_A by masking absent features in a fixed neural network and imputing them with samples from a background (or baseline) distribution. (c) In tree ensembles, TreeSHAP efficiently marginalizes missing features by splitting the model's output across branches, weighted by training data proportions (Lundberg et al., 2018). (d) Many more payoff constructions exist—for example, sampling guided by conditional and marginal feature distributions (Covert et al., 2020) and derivative-based methods for scalable computation (Duan and Okten, 2025). Notably, each surrogate method approximates risk reduction under specific assumptions, such as feature independence, distributional priors, or model smoothness. Once ν_A is constructed, researchers often mechanically apply the Shapley formula to compute feature attributions.

However, the theoretical justification for Shapley values relies on several foundational "axioms''—efficiency, symmetry, linearity, and nullity (Shapley, 1953)—which are not easily testable and are *rarely* validated in practice. In particular, Shapley's framework implicitly assumes an **additive structure** (Lundberg and Lee, 2017):

$$\nu_A \sim \sum_{j\in A} \beta_j \quad \text{or} \quad V_A(\{\beta_j\}_{j\in A}) = \sum_{j\in A} \beta_j. \tag{2}$$

But the so-called additive feature attribution is not guaranteed to hold in real-world constructions of coalition values. For example, we can reformulate the abstract Shapley axioms and principles into a multivariate Gaussian assumption (cf. Section 2), but many of the constructions mentioned previously are prone to violating this assumption due to non-Gaussian characteristics such as bounded ranges, heavy tails, and skewness. In particular, Fryer et al. (2021) recently proposed a realistic "taxicab" payoff defined by a *winner-takes-all* dynamic that is in stark contrast to (2):

$$V_A(\{\beta_j\}_{j\in A}) = \max_{j\in A} \beta_j, \ \forall A \subseteq 2^F. \tag{3}$$

Such nonlinear relations are prevalent in applications but fundamentally violate the

additive model underpinning standard Shapley value estimation.

(ii) Embedding sparsity into value attribution: In many real-world applications with a large number of features, a substantial proportion contribute only negligibly—or are effectively irrelevant—to the overall outcome, making them unnecessary to explain in practice (Strumbelj and Kononenko, 2014; Covert et al., 2020). Exploiting the structural parsimony can enhance both statistical accuracy and interpretability of Shapley values. In implementation, leveraging sparsity helps to reduce iteration complexity, thanks to a substantially smaller effective model size, along with mitigating communication costs and storage requirements in high-dimensional settings.

However, most existing approaches adopt a **greedy** strategy by first computing dense Shapley values based on the *full* model, followed by post hoc ranking or thresholding to achieve sparsity (Cohen et al., 2007; Jothi et al., 2021; Fryer et al., 2021; Au et al., 2022). For large p, such multi-step procedures are not only inefficient but may also fail to provide faithful explanations or meaningful selection (see, e.g., Covert et al. (2020); Slack et al. (2020); Ma and Tourani (2020)). To the best of our knowledge, *no* widely adopted framework integrates sparsity as an intrinsic property into Shapley-value estimation, let alone in the context of an unknown transformation. These challenges indicate the need for developing unified approaches that *simultaneously* enforce sparsity and ensure coherent Shapley-based attributions.

This paper aims to develop a novel nonlinear explanation framework for applying the Shapley mechanism in a way that simultaneously aligns individual feature contributions with appropriately transformed worths across all subsets, and promotes sparsity by eliminating irrelevant features to enhance both computational efficiency and statistical accuracy. The contributions of our work are as follows:

- Our research is the first to demonstrate that common factors such as the presence of irrelevant features and inter-feature dependencies can induce a payoff transformation that deviates substantially from linearity, even when using standard payoff constructions (e.g., R^2-based worths). This finding underscores the need for nonlinear explainability frameworks.

- We propose Sparse Isotonic Shapley Regression (**SISR**), the first framework to *jointly* address nonlinearity and sparsity in Shapley attributions. By learning a monotonic transformation and enforcing an ℓ_0 constraint simultaneously, our integrated approach overcomes the limitations of ad-hoc methods.

- SISR learns the transformation of payoffs without requiring a predefined analytical form. This is achieved through efficiently leveraging the Pool-Adjacent-Violators algorithm, allowing the model to adapt to diverse real-world payoff structures.
- The optimization algorithm developed for SISR features simple, closed-form updates and is accompanied by global convergence guarantees. The incorporation of sparsity improves computational efficiency
- Through extensive experiments across various datasets and payoff schemes, we show that SISR significantly stabilizes feature attributions and correctly identifies relevant features, mitigating the severe rank and sign distortions often observed with standard Shapley value applications.

The rest of the paper is organized as follows. Section 2 proposes a novel Sparse Isotonic Shapley Regression model to address challenges related to domain adaptation and high dimensionality. In Section 3, an optimization-based algorithm is developed to address the functional challenge and the nonsmooth sparsification, with established theoretical guarantees. Section 4 provides valuable data-driven insights drawn from experiments in various scenarios. We conclude in Section 5.

In accordance with submission guidelines, the author has also submitted a separate work titled, "*Structured Modeling of Cancer Pharmacogenomic Outcomes under Latent Confounding*," to the biology category.

2 Proposed Method

The Shapley axioms and principles have been interpreted by economists in various ways (Algaba et al., 2019). Here, we recast the Shapley framework as a statistical assumption on the data-generating process. To begin, let us revisit a motivating *weighted least squares* formulation of Shapley value estimation as derived in Lundberg and Lee (2017), echoing earlier developments in econometrics (Charnes et al., 1988):

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p, c\in\mathbb{R}} \sum_{A\subseteq 2^F,\, A\neq\emptyset,\, A\neq F} w_{\mathrm{SH}}(A) \left(\nu_A - \sum_{j\in A}\beta_j - c\right)^2 \\ \text{subject to} \quad c = \nu_\emptyset, \quad c + \sum_{j=1}^{p}\beta_j = \nu_F, \end{aligned} \tag{4}$$

where the Shapley weights are given by

$$w_{\text{SH}}(A) = \frac{p-1}{\binom{p}{|A|}|A|(p-|A|)}. \tag{5}$$

Perhaps surprisingly, it can be shown that the optimal solution $\hat{\beta}$ to (4) recovers the exact Shapley values (Lundberg and Lee, 2017), which are traditionally derived based on the concept of *marginal contributions* across all possible feature coalitions.

If we define

$$w_{\text{SH}}(\emptyset) = +\infty, \quad w_{\text{SH}}(F) = +\infty \tag{6}$$

as an extension of (5), then (4) can be written as

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

where A can take any subset of the power set 2^F. Note that when $A = \emptyset$, $\sum_{j \in A} \beta_j = 0$ by convention and $\hat{c} = \nu_\emptyset$. It is thus convenient to define the *baseline-adjusted* coalition values:

$$\nu_A^c = \nu_A - \nu_\emptyset, \quad \forall A \subseteq 2^F, \tag{7}$$

which saves one parameter in the subsequent optimization:

$$\min_{\beta} \sum_{A \subseteq 2^F} w_{\text{SH}}(A) \left(\nu_A^c - \sum_{j \in A} \beta_j \right)^2.$$

For notational simplicity, we will write 'ν_A^c' as just 'ν_A', assuming that all ν values have been properly shifted according to (7) unless otherwise specified.

It is helpful to reinterpret the weighted least squares formulation of Shapley values as a probabilistic model:

$$\begin{aligned} \nu_A &\sim \mathcal{N}(\mu_A, \sigma_A^2) \\ \mu_A &= \sum_{j \in A} \beta_j^*, \\ \sigma_A^2 &\propto \binom{p}{|A|} |A|(p-|A|) \, (\propto \frac{1}{w_{\text{SH}}(A)}), \end{aligned} \tag{8}$$

and all ν_A's are independent. Here, β_j^* denotes the true Shapley value for the jth feature.

By reformulating the Shapley axioms and principles into assumption (8), we gain insight into why numerous payoff functions may not meet the model criteria. Indeed, due to issues such as range constraints, skewness, heavy tails, and heterogeneity, it is natural to question the appropriateness of the multivariate Gaussianity across different definitions of coalition values.

In our view, one viable solution is to apply a transformation that promote Gaussianity. Let's consider an alternative Shapley value model in a *transformed domain*:

$$T(\nu_A) \sim \mathcal{N}(\sum_{j\in A} T(\beta_j^*),\, \sigma_A^2), \tag{9}$$

where $T(\cdot)$ is an unknown transformation. Under this model,

$$\mathbb{E}\left[T(\nu_A)\right] = \sum_{j\in A} T(\beta_j^*), \tag{10}$$

which defines a "T-additive" framework for nonlinear settings. To model this structure, we propose a new Shapley framework termed Functional Shapley Regression, which jointly estimates β and $T(\cdot)$ by solving

$$\min_{\beta, T(\cdot)} \sum_{A\subseteq 2^F} w_{\text{SH}}(A)\Big\{T(\nu_A) - \sum_{j\in A} T(\beta_j)\Big\}^2 \text{ subject to } \beta \in \mathcal{C},\, T(\cdot) \in \mathcal{T}, \tag{11}$$

where the objective minimizes the Shapley-weighted sum of squared differences between the transformed coalition values $T(\nu_A)$ and the transformed linear sum $\sum_{j\in A} T(\beta_j)$ over all subsets $A \subseteq 2^F$. Here ,we use $\mathcal{C} \subseteq \mathbb{R}^p$ to denote the constraint set for β, and $\mathcal{T}$ to denote the class of admissible transformation functions. By the notational convention for $A = \emptyset$, (11) automatically enforces

$$T(0) = 0,$$

corresponding to $T(\nu_\emptyset) = 0$ (recall all ν_A have been centered).

The remark below illustrates that our framework, in contrast to the common additive model (see, e.g., Lundberg and Lee (2017)), accommodates a broader range of multivariate structures in the payoff function, enabling nonlinear explainability.

Remark 1 (**Univariate T-Mappings** for **Multivariate Structure**). *Introducing a* univariate *transformation $T(\cdot)$ enables a remarkably rich class of models capable*

of capturing complex multivariate *relationships between* ν_A *and* $\{\beta_j : j \in A\}$*, well beyond the standard additive form.*

Specifically, under the T*-transformed model* (9)*, assuming the existence of the inverse transformation* T^{-1} *and using the notation* V_A *(cf.* (1)*) we have*

$$V_A(\{\beta_j\}_{j\in A}) = T^{-1}\Big(\sum_{j\in A} T(\beta_j)\Big) \quad \text{or} \quad \nu_A \sim T^{-1}\Big(\sum_{j\in A} T(\beta_j)\Big). \tag{12}$$

If T *is a nondegenerate linear map like the identity map, the model reduces to the conventional additive Shapley game, where the coalition value* ν_A *in* (12) *is essentially a simple sum of individual contributions. However, a general transformation allows* (12) *to adapt to a broad range of application domains.*

For instance, consider a monomial transformation $T(x) = |x|^d$ *for some* $d > 0$*, which induces a multivariate "d-norm" relationship:*

$$V_A(\{\beta_j\}_{j\in A}) = \big(\sum_{j\in A} |\beta_j|^d\big)^{1/d} = \|\beta_A\|_d.$$

Varying the degree d *recovers a spectrum of geometric structures, e.g.,*

(i) $d = 1$*: the* ℓ_1*-norm polytope,* $V_A(\{\beta_j\}_{j\in A}) = \sum_{j\in A} |\beta_j|$*;*

(ii) $d = 2$*: the* ℓ_2*-norm ball,* $V_A(\{\beta_j\}_{j\in A}) = \big(\sum_{j\in A} \beta_j^2\big)^{1/2}$*;*

(iii) $d \to \infty$*: the* ℓ_∞*-norm cube,* $V_A(\{\beta_j\}_{j\in A}) = \max_{j\in A} |\beta_j|$*.*

In particular, under nonnegativity constraints ($\nu_A \geq 0$*,* $\beta_j \geq 0$*), the* ℓ_∞ *case corresponds to the* winner-takes-all *mechanism as first noted in Fryer et al. (2021), where the coalition value is dominated by the largest individual contribution (practically, monomial transformations with large degrees* d *can closely approximate such behavior). This is motivating, as the examples here are highly nonlinear and incompatible with a linear Shapley game. Yet, with a univariate transformation, they can be incorporated into the* T*-Shapley framework. Additional examples are the exponential form* $T(x) = \exp(x) - 1$ *and the odds form* $T(x) = \Phi(x)/(1 - \Phi(x))$ *with* Φ *a distribution function of a continuous random variable.*

In sum, an appropriately chosen $T(\cdot)$ *establishes a versatile* nonlinear *modeling mechanism that enhances the additive expressiveness of Shapley values for XAI. Another advantage of the proposed approach is that it* bypasses *the need for a predefined analytic transformation, instead learning it directly from the data (cf. Section 3).*

In this paper, we focus a specific instance of (11), referred to as the "**S**parse **I**sotonic **S**hapley **R**egression" (**SISR**):

$$\begin{aligned} &\min_{\beta, T(\cdot)} \sum_{A \subseteq 2^F} w_{\text{SH}}(A) \Big\{ T(\nu_A) - \sum_{j \in A} T(\beta_j) \Big\}^2 \\ &\text{subject to } \|\beta\|_0 \leq s, T \in \mathcal{M}, \sum_{j=1}^{p} (T(\beta_j))^2 = 1, \end{aligned} \tag{13}$$

where $\mathcal{M}$ denotes the class of strictly increasing functions and $1 \leq s \leq p$ specifies the user-defined upper bound on the true model sparsity. (13) incorporates three critical modeling considerations.

Monotonicity. We impose a monotonicity constraint on $T(\cdot)$ to preserve the relative ordering of feature importance values:

$$\beta_i \geq \beta_j \quad \Rightarrow \quad T(\beta_i) \geq T(\beta_j).$$

This ensures that the learned transformation respects the relative contribution levels of individual features. The structure closely resembles **isotonic regression** (Robertson et al., 1988), which seeks a weighted least squares fit under monotonicity constraints and has widespread applications in psychometrics, epidemiology, yield curve estimation, risk modeling and credit scoring studies. Compared with enforcing smoothness in T, our monotonicity approach avoids any need for a basis expansion or other parametric representation. Pursuing T reinterprets the data in a transformed domain where feature contributions recover an additive Shapley structure.

Normalization. A normalization is imposed on the transformed feature contributions, $\sum_{j=1}^{p} (T(\beta_j))^2 = 1$. This prevents degeneracy (e.g., trivial solutions such as $T \equiv 0$) and anchors the scale of the model. An appealing feature of (13) is its invariance to the overall scaling of $\{\nu_A\}$, and the normalization constant is fixed at 1 without loss of generality. Moreover, Section 3 will show that imposing such a spherical constraint yields computational benefits, enabling a closed-form solution for the attribution-update and improving implementability.

Sparsity. (13) directly incorporates **sparsity** into the Shapley estimation process. Rather than relying on multi-step methods that first estimate a dense Shapley vector and then rank features (Slack et al., 2020), the formulation constrains the support of $\hat{\beta}$ while pursuing the transformation during the iterative optimization process (cf.

Algorithm 1). This *unified* treatment ensures that sparsity, domain adaptation, and Shapley coherence are achieved simultaneously, avoiding inconsistencies by post hoc selection. Unlike the popular ℓ_1-penalty $\lambda \sum |\beta_j|$, which requires cumbersome λ-tuning and induces unwanted shrinkage, our ℓ_0 regularization provides direct control over the model's sparsity level. This is advantageous in fields like bioinformatics, where practitioners often require a pre-specified number of features. For cases where s must be tuned, we find the RIC criterion (Foster and George, 1994) is an effective selection method within our Shapley framework.

Before concluding this section, we introduce a *reparameterization* trick that proves beneficial for both modeling and computation. Define

$$\gamma_j = T(\beta_j). \tag{14}$$

Since T is strictly increasing and $T(0) = 0$ (the loss would become infinite if $T(0) \neq 0$), (13) can be rewritten as

$$\min_{\gamma, T(\cdot)} \quad \sum_{A \subseteq 2^F} w_{\text{SH}}(A)(T(\nu_A) - \sum_{j \in A} \gamma_j)^2 \text{ s.t. } \|\gamma\|_0 \leq s, \|\gamma\|_2 = 1, T \in \mathcal{M}. \tag{15}$$

The corresponding model assumption is thus $T^*(\nu_A) \sim \mathcal{N}(\sum_{j \in A} \gamma_j^*,\ \sigma_A^2)$ for all $A \subseteq 2^F$, where the genuine transformation function T^* is monotonic with $T^*(0) = 0$, and ν_A are assumed to be independent across different subsets A. The *starred* quantities represent the underlying statistical truth of interest to estimate. Assume $\gamma^* \in \mathbb{R}^p$ satisfies $\|\gamma^*\|_0 \leq s^*$ and $\|\gamma^*\|_2 = 1$ with $1 \leq s^* \leq p$ and s is specified as an upper bound on s^*. After estimating $\hat{\gamma}$ and $\hat{T}$ from (15), one can recover the β-scores by applying the inverse transformation $\hat{\beta}_j = \hat{T}^{-1}(\hat{\gamma}_j)$. This reconstructs the multivariate relationship between ν_A and the set of feature contributions in the original scale, yielding $\nu_A \approx \hat{T}^{-1}(\sum_{j \in A} \hat{T}(\hat{\beta}_j))$, to offer interpretable Shapley-based attributions.

3 Optimization Algorithm

The optimization of SISR involves two main challenges: (i) a functional estimation component, and (ii) a combinatorial sparsity constraint coupled with a nonconvex normalization constraint. We show that the functional challenge can be addressed by a discretization technique, which, rather than introducing an approximation, preserves full equivalence. To handle the two constraints on γ, we develop a surrogate function framework. These efforts lead to an iterative procedure that combines the

pool-adjacent-violators with a normalized hard thresholding. Each step has implementation ease and the sparse structure ensures that the overall algorithm remains efficient in high-dimensional settings.

First, since $T(\cdot)$ is only evaluated at the observed values ν_A in the objective function, we "discretize" (15) by introducing the vector

$$t = \big[T(\nu_A)\big]_{A\subseteq 2^F} \in \mathbb{R}^{2^p}.$$

In defining this vector, one should fix a specific order over subsets $A \subseteq F$; we follow the conventional lexicographic binary ordering to arrange the entries of t. Correspondingly, we define

$$\nu = \big[\nu_A\big]_{A\subseteq 2^F} \in \mathbb{R}^{2^p}, \quad \delta = \Big[\sum_{j\in A}\gamma_j\Big]_{A\subseteq 2^F} = Z\gamma \in \mathbb{R}^{2^p},$$

where $Z \in \mathbb{R}^{2^p\times p}$ is the "incidence matrix" indicating which features are active in each subset A, aligned with the same ordering used for t. Henceforth, we also write ν_i (and likewise δ_i) to denote the entry corresponding to the ith subset. Additionally, introduce the diagonal weight matrix

$$W = \mathrm{diag}\{w_{\mathrm{SH}}(A)\}_{A\subseteq 2^F}. \tag{16}$$

With this notation in place, we study the following optimization problem:

$$\begin{aligned} \min_{\gamma\in\mathbb{R}^p,\, t\in\mathbb{R}^{2^p}} \quad & \frac{1}{2}(t-\delta)^\top W(t-\delta) \\ \text{subject to} \quad & \delta = Z\gamma, \quad \|\gamma\|_0 \le s, \quad \|\gamma\|_2 = 1, \\ & t_i \le t_j \quad \text{for all } (i,j)\in E(\nu) = \{(i,j) : \nu_i \le \nu_j\}, \end{aligned} \tag{17}$$

where E encodes the pairwise ordering constraints induced by ν, due to the monotonicity of the transformation T. This formulation replaces strict monotonicity with a non-decreasing constraint, a mild adjustment that facilitates numerical implementation. In the following of the section, we design a two-block alternating optimization algorithm.

First, with δ fixed, the optimization over t corresponds to the (weighted) *isotonic regression*

$$\min_{t\in\mathbb{R}^{2^p}} \quad \frac{1}{2}(t-\delta)^\top W(t-\delta) \text{ subject to } \ t_i \le t_j \text{ for all } (i,j)\in E, \tag{18}$$

where the goal is to obtain a monotonic fit to δ under a weighted squared-error loss defined by W. The problem can be solved using any standard Quadratic Programming (QP) solver, but it is more efficiently handled by the Pool-Adjacent-Violators Algorithm (**PAVA**)(de Leeuw et al., 2009), which leverages the structure of the monotonicity constraints for improved computational performance.

Next, we focus on the γ-optimization.

Theorem 1. *Let $\mathcal{H}(\cdot; s)$ denote the* hard-thresholding *operator associated with cardinality s, defined as follows: for a vector $y \in \mathbb{R}^p$, $\mathcal{H}(y; s) = z$ where $z_i = y_i$ if $|y_i|$ is among the s largest entries of $|y_1|, \ldots, |y_p|$, and $z_i = 0$ otherwise, and the* normalized hard-thresholding *operator $\mathcal{H}^\circ(y; s) = \mathcal{H}(y; s)/\|\mathcal{H}(y; s)\|_2$ if $\|\mathcal{H}(y; s)\|_2 \neq 0$. Then, for the optimization problem with $y \neq \vec{0}$, $1 \leq s \leq p$,*

$$\min_{\beta} \frac{1}{2}\|y - \beta\|_2^2 \quad \textit{subject to} \quad \|\beta\|_0 \leq s,\ \|\beta\|_2 = 1,$$

the vector obtained by normalized hard-thresholding,

$$\hat{\beta} = \mathcal{H}^\circ(y; s) = \frac{\mathcal{H}(y; s)}{\|\mathcal{H}(y; s)\|_2}$$

is a global optimizer.

Proof. Let $A \subseteq \{1, \ldots p\}$ and assume $\beta_{A^c} = 0$ and $\|\beta\|_2 = 1$. Because

$$\begin{aligned}
\|y - \beta\|_2^2 =& \|y\|_2^2 + \|\beta\|_2^2 - 2\langle y, \beta\rangle \\
=& \|y\|_2^2 + 1 - 2\langle y, \beta\rangle \\
=& \|y\|_2^2 + 1 - 2\langle y_A, \beta_A\rangle \\
\geq& \|y\|_2^2 + 1 - 2\|y_A\|_2\|\beta_A\|_2 = \|y\|_2^2 + 1 - 2\|y_A\|_2,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality and the equality is achieved at $\beta_A = y_A/\|y_A\|_2$.

Therefore, $\min_{\beta:\beta_{A^c}=0,\|\beta\|_2=1} \|y - \beta\|_2^2 = \|y\|_2^2 + 1 - 2\|y_A\|_2$ for any $A : |A| = s$. Minimizing over A gives an index set corresponding to the s largest entries of $|y_1|, \ldots, |y_p|$, thereby the normalized hard thresholding operator $\mathcal{H}^\circ(y; s)$. □

We are now ready to develop an iterative algorithm for updating γ with t held fixed. Define the below objective function

$$l(\gamma) = \frac{1}{2}(Z\gamma - t)^\top W(Z\gamma - t).$$

A straightforward calculation yields the gradient:

$$\nabla l(\gamma) = Z^\top W(Z\gamma - t).$$

To facilitate optimization, we construct a new "surrogate function":

$$g(\gamma, \gamma^-) = l(\gamma^-) + \langle \nabla l(\gamma^-), \gamma - \gamma^- \rangle + \frac{\rho}{2}\|\gamma - \gamma^-\|_2^2.$$

where $\rho > 0$ should be properly large (cf. Theorem 2). Define an iterative scheme:

$$\gamma^{(k+1)} = \arg\min_{\gamma} \; g(\gamma, \gamma^{(k)}) \quad \text{subject to } \|\gamma\|_0 \le s, \; \|\gamma\|_2 = 1.$$

Using Theorem 1, the update step admits a closed-form expression:

$$\begin{aligned} \gamma^{(k+1)} &= \mathcal{H}^\circ(y; s) = \frac{\mathcal{H}(y; s)}{\|\mathcal{H}(y; s)\|_2}, \text{ with} \\ y &= \gamma^{(k)} - \frac{1}{\rho}\nabla l(\gamma^{(k)}) = \gamma^{(k)} - \frac{1}{\rho} Z^\top W(Z\gamma^{(k)} - t). \end{aligned} \tag{19}$$

Theorem 2. *Let $\rho \ge \|Z^\top W Z\|_2$, where $\|\cdot\|_2$ denotes the martrix spectral norm. For any initial point $\gamma^{(0)}$ satisfying $\|\gamma^{(0)}\|_0 \le s$ and $\|\gamma^{(0)}\|_2 = 1$, the sequence $\{\gamma^{(k)}\}$ generated by (19) produces non-increasing (and thus convergent) function values:*

$$l(\gamma^{(k+1)}) \le l(\gamma^{(k)}) \quad \textit{for all } k \ge 0.$$

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

Proof. First, simple algebra shows

$$\begin{aligned} g(\gamma, \gamma^-) - l(\gamma) &= \frac{\rho}{2}\|\gamma - \gamma^-\|_2^2 - (l(\gamma) - l(\gamma^-) - \langle \nabla l(\gamma^-), \gamma - \gamma^- \rangle) \\ &= \frac{\rho}{2}\|\gamma - \gamma^-\|_2^2 - \frac{1}{2}(\gamma - \gamma^-)^\top H(\xi)(\gamma - \gamma^-) \\ &= \frac{1}{2}(\gamma - \gamma^-)^\top (\rho I - H(\xi))(\gamma - \gamma^-), \end{aligned}$$

where we applied the mean-value theorem, $H(\xi)$ denotes the Hessian matrix of l at ξ which is between γ and γ^-. Thus under the choice of ρ, $l(\gamma^{(k+1)}) \le g(\gamma^{(k+1)}, \gamma^{(k)})$ for any $k \ge 0$.

By the optimality of $\gamma^{(k+1)}$, $g(\gamma^{(k+1)}, \gamma^{(k)}) \le g(\gamma^{(k)}, \gamma^{(k)}) = l(\gamma^{(k)})$ and the first conclusion follows. Moreover, from the inequality: $l(\gamma^{(k)}) - l(\gamma^{(k+1)}) \ge \frac{1}{2}(\gamma^{(k+1)} - \gamma^{(k)})^\top(\rho I - H(\xi))(\gamma^{(k+1)} - \gamma^{(k)}) \ge \frac{\rho - \|Z^\top W Z\|_2}{2}\|\gamma^{(k+1)} - \gamma^{(k)}\|_2^2$, we obtain the second result. □

Algorithm 1 Sparse Isotonic Shapley Regression (**SISR**) Algorithm

Input: $\nu = [\nu_A]_{A \subseteq 2^F} \in \mathbb{R}^{2^p}$ (baseline-adjusted, such that $\nu_\emptyset = 0$), sparsity level s, design matrix $Z \in \mathbb{R}^{2^p \times p}$, diagonal weight matrix W (cf. (16)), and an initial vector $t^{(0)} \in \mathbb{R}^{2^p}$ (e.g., $C\nu$ with a large C if $\|\nu\|_\infty$ is small, to improve precision, and $C = 1$ otherwise).

1: Initialize $t \leftarrow t^{(0)}$, $\gamma \leftarrow 0$
2: $\rho \leftarrow \|Z^\top W Z\|_2$
3: **repeat**
4: **while** not converged **do**
5: $\xi \leftarrow \mathcal{H}(\gamma - \frac{1}{\rho} Z^\top W (Z\gamma - t); s)$
6: $\gamma \leftarrow \frac{\xi}{\|\xi\|_2}$
7: **end while**
8: $\delta \leftarrow Z\gamma$
9: Fit isotonic regression (18) with δ, W, Z to update t
10: **until** convergence
11: **return** t, γ

A summary of our algorithmic procedure is outlined in Algorithm 1. Some practical implementation notes: (i) The provided values ν_A have been baseline adjusted as described in (7) (i.e., a preprocessing $\nu_A \leftarrow \nu_A - \nu_\emptyset$ for all $A \subseteq 2^F$ is assumed). We take C as 1e+4 if $\|\nu\|_\infty \leq 10$. (ii) For $A = \emptyset$ or $A = F$, although $w_{\text{SH}}(A)$ take infinite weights in theory, practically one can assign a weight equal to a large multiplier (e.g., 10) times the largest non-infinite weight (cf. (5)), which is often numerically sufficient to enforce $\hat{T}(\nu_F) \doteq \sum_{j=1}^p \hat{T}(\hat{\beta}_j)$. (iii) It is unnecessary to explicitly form the diagonal matrix W; only the diagonal weights are required. Likewise, the sparsity of matrix Z can be utilized. Additionally, key quantities such as $Z^\top W$ and $Z^\top W t$ can be precomputed prior to the iterative updates to improve computational efficiency. (iv) In Step 9), we employ a self-implemented, *stack*-based weighted PAVA for improved efficiency. (v) The paired data $(\nu_i, \hat{t}_i)$ approximate T and form the basis for visualizing $\hat{T}(\cdot)$ and its inverse $\hat{T}^{-1}(\cdot)$ via a smooth monotonic interpolation of unique, sorted pairs. By averaging any duplicate coordinates before interpolation, the method guarantees a well-defined increasing mapping suitable for a variety of applications. Overall, Algorithm 1 is straightforward to implement and scales very well in practice.

4 Data-Driven Insights

4.1 Domain Adaptation

To propose a convenient noisy data generation scheme, let's revisit the statistical model defined in Section 2, where the expectation $\mathbb{E}[T^*(\nu)] = Z\gamma^*$, with $T(\cdot)$ applied componentwise. Assume without loss of generality that Z is structured using a bit generation process, with each row corresponding to the binary representation of $i-1$ (e.g., the second row is $[1, 0, \ldots, 0]$). For $\gamma^* = c_0[2^0, 2^1, \ldots, 2^{p-2}, 2^{p-1}]^\top$, this yields

$$\mathbb{E}[T^*(\nu)] = c_0[0, 1, \ldots, 2^p - 2, 2^p - 1]^\top$$

where $c_0 = \sqrt{\frac{3}{4^p-1}}$ ensures that γ^* is normalized. To simulate this in experiments, we generate *noisy* versions ν_A for all subsets A using

$$\nu = Q(c_1 \cdot \sigma(U)) \in \mathbb{R}^{2^p}$$

where $U \in \mathbb{R}^{2^p}$ contains entries uniformly distributed between 0 and $c_0(2^p - 1)$, approximately $\sqrt{3}$ when p is sufficiently large. Here, σ the permutation that arranges the elements of U in ascending order. An accurate estimator, $\hat{T}$ or $\hat{t}$, should then closely approximate the inverse transformation

$$T^* = Q^{-1}/c_1.$$

The inclusion of c_1 is to ensure flexibility.

Figure 1 presents the results under 6 different functional forms for the true transformation T^*: *square root* ($T^* = (\cdot)^{1/2}$), *fifth root* ($T^* = (\cdot)^{1/5}$), *exponential* ($T^* = \exp(\cdot) - 1$), *logarithmic* ($T^* = \log(\cdot + 1)$), *tangent* ($T^* = \tan(\cdot)/c_1, c_1 = 10$), and *normal distribution* ($T^* = \Phi(\cdot + c_2)/c_1, Q(\cdot) = \Phi^{-1}(c_1\cdot) - c_2, c_1 = 1/\sqrt{3}, c_2 = Q(c_1\sigma_{\min})$). Encouragingly, across all cases, the estimated transformation $\hat{T}(\nu)$ closely aligns the ground-truth $T^*(\nu)$, providing strong empirical evidence for the effectiveness of SISR in accurately recovering the underlying transformation structure.

Finally, an additional experiment was conducted using data generated according to $\nu_A = \max_{j \in A} \beta_j$, where $\beta_j = j$. The resulting estimated transformation is displayed in Figure 2. The recovered transformation exhibits a pronounced increasing trend and nonlinearity. The estimates are nearly perfectly correlated with the transformed ground truths, and the best-fit line passes through the origin up to a scaling factor.

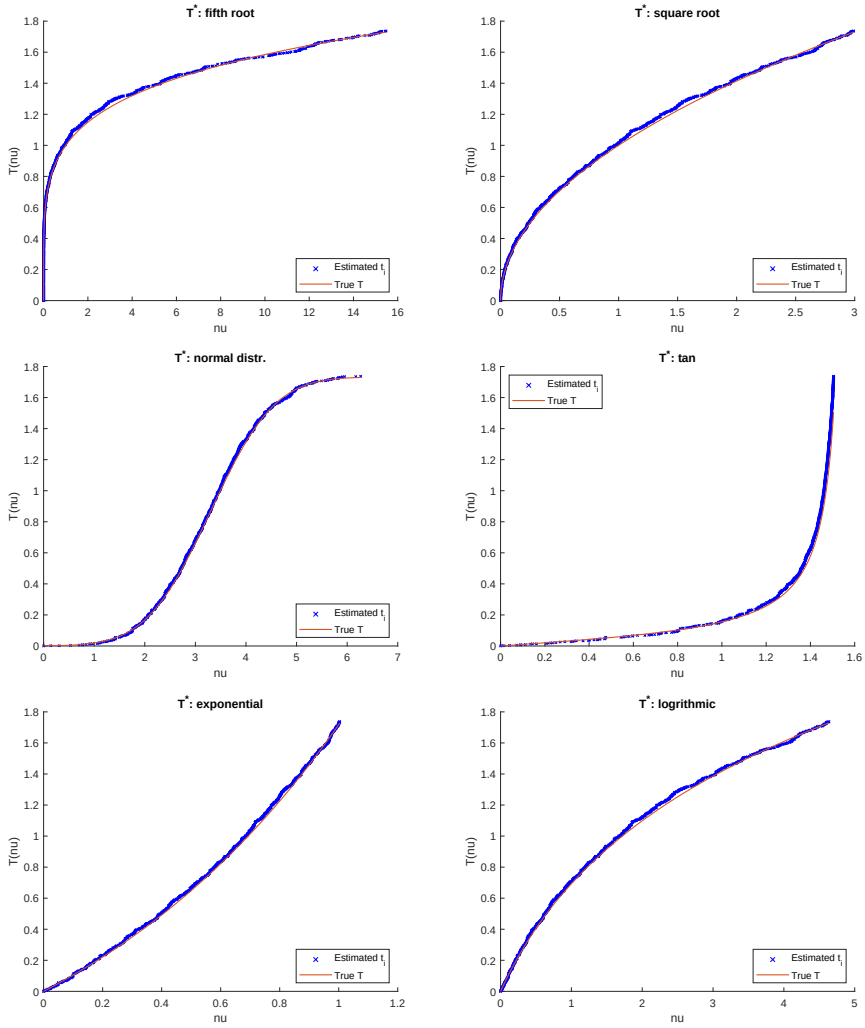


Figure 1: Estimated monotonic transformation $\hat{T}(\nu)$ (in *blue*) versus the true transformation T^* (in *red*), for $p = 10$ under 6 different functional forms for T^*: the fifth root, square root, normal distribution, tangent, exponential, and logarithmic transformations.

4.2 Sparsity Recovery

We generate data according to the sparse-γ model in Section 2, with the true transformation set as the cubic root, $T^*(\cdot) = \sqrt[3]{\cdot}$. The ground-truth coefficients are given by $\gamma^* = [1/\sqrt{3},\ 1/\sqrt{3},\ 1/\sqrt{3},\ 0, \ldots, 0]^\top$, a relatively weak signal with sparsity level $s^* = 3$. The noise variance is defined as $\sigma_A^2 = \sigma_0^2 / w_{\mathrm{SH}}(A)$, with varying values

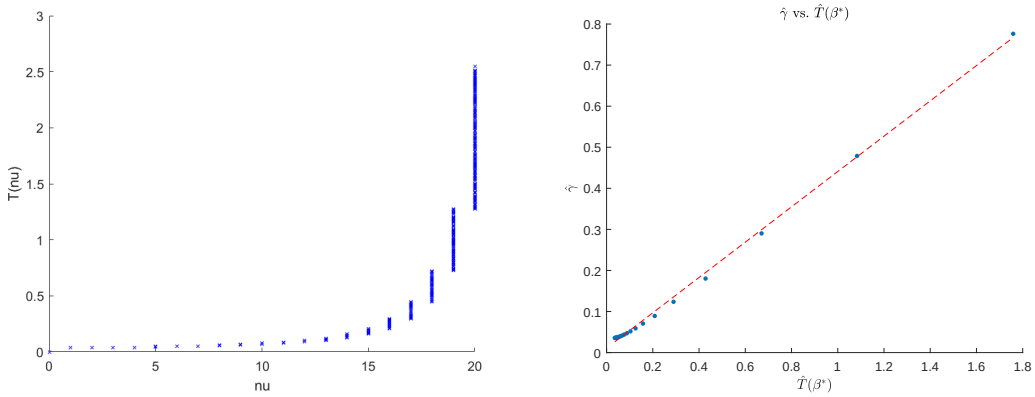


Figure 2: Estimated monotonic transformation $\hat{T}(\nu)$ (left) and comparison between $\hat{\gamma}$ vs $\hat{T}(\beta^*)$ (right, showing an almost perfect correlation of 1.00) for $p = 20$ under a winner-takes-all setting.

of σ_0. The sparsity level upper bound s in running SISR is set to $1.5\,s^*$. Performance is evaluated using two metrics. The first measures the alignment or affinity between the two unit-norm vectors: $\langle \hat{\gamma}, \gamma^* \rangle \times 100$ (denoted by **Affn**), serving as an index of estimation accuracy. The second metric is the support recovery rate: $|\text{supp}(\hat{\gamma}) \cap \text{supp}(\gamma^*)|/s^* \times 100\%$ (denoted by **Supp**), reflecting the proportion of correctly identified nonzero components in the true support. Table 1 reports results for varying values of p and σ_0. All results are averaged over 100 simulation runs.

As shown in the table, both performance metrics decline with increasing feature dimension p and noise level σ_0, as expected due to greater model complexity and reduced signal-to-noise ratio (SNR). Although not reported, running the algorithm without sparsity enforcement (i.e., $s = p$) yields noticeably worse affinity scores in high SNR settings (e.g., the first setting row of Table 1). Compared to the affinity scores, the support recovery rate remains surprisingly strong even under challenging conditions, indicating that SISR consistently identifies the correct features.

We further investigated the impact of the sparsity level s on computational time. In this experiment, we varied s from 5 to 15, while fixing $s^* = 3$, $p = 15$, $\sigma_0 = 5e\text{-}3$. As illustrated in the figure, lower sparsity levels generally lead to faster computation, highlighting the efficiency gains achievable when enforcing proper sparsity in the model.

Table 1: Affinity score (**Affn**) and support recovery rate (**Supp**) across different values of p and noise level σ_0. Larger values reflect better performance.

	$\sigma_0 = $ 1e-3		$\sigma_0 = $ 5e-3		$\sigma_0 = $ 1e-2	
	Affn	Supp	Affn	Supp	Affn	Supp
$p = 10$	99.6	100%	99.6	100%	99.5	100%
$p = 15$	99.8	100%	99.9	100%	97.8	100%
$p = 20$	99.9	100%	87.9	100%	80.3	100%
$p = 25$	87.9	100%	74.0	100%	70.5	100%
	$\sigma_0 = $ 5e-2		$\sigma_0 = $ 1e-1		$\sigma_0 = $ 2e-1	
	Affn	Supp	Affn	Supp	Affn	Supp
$p = 10$	97.9	100%	88.7	98.7%	66.2	80.7%
$p = 15$	79.9	100%	70.9	98.0%	57.6	73.3%
$p = 20$	68.9	100%	63.2	96.0%	54.3	65.3%
$p = 25$	65.5	100%	60.6	90.7%	52.1	62.0%

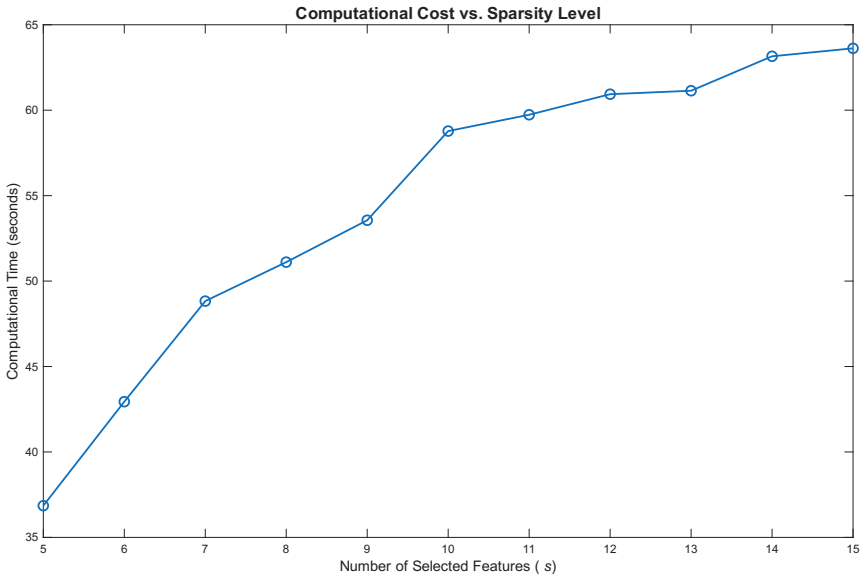


Figure 3: Computational time versus sparsity level.

4.3 R^2-Payoffs in Regression/Logistic Regression

In regression settings, coalition values for feature subsets are commonly defined using the *coefficient of determination* (R^2) obtained from retraining the model on each subset, reflecting the scaled improvement in model fit (Lipovetsky and Conklin, 2001; Covert et al., 2020). Contrary to conventional expectations, our results unveil a novel insight: such a standard construction can fail to yield an inherently additive Shapley framework, especially when features are dependent or include irrelevant ones, which are almost certain to occur in practice.

To illustrate this phenomenon, let's consider the following simulation setup: $y = X\alpha^* + \epsilon$, where $\epsilon_i \sim \mathcal{N}(0, 1)$ and each row of $X \in \mathbb{R}^{n\times p}$ is drawn from a multivariate normal distribution with mean zero and Toeplitz covariance $\Sigma_{ij} = \theta^{|i-j|}$ with $\theta = 0.5$. The true coefficient vector is set as $\alpha^* = [3,\ 3,\ \ldots,\ 3]^\top$, and the sample size is $n = 5p$. For each subset A, we fit a regression model using the predictors indexed by A and define ν_A as the resulting R^2 value. Logistic regression is also considered in the classification setting, $y_i \sim \text{Bernoulli}(\pi_i)$ and $\text{logit}(\pi) = X\alpha^*$, where we generate y according to a Bernoulli model and define ν_A using the *deviance*-based pseudo-R^2.

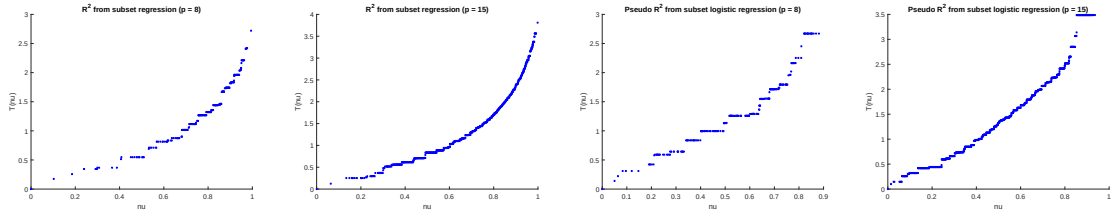


Figure 4: Estimated monotonic transformation $\hat{T}(\nu)$ using regression-based R^2 and logistic regression-based pseudo-R^2 as the coalition worth function, for $p = 8, 15$.

As shown in Figure 4, the estimated transformation $\hat{T}(\nu)$ deviates significantly from linearity. For instance, in the regression case, even a simple logarithmic transformation fails to linearize the relationship, whereas a log-log transformation produces a nearly linear pattern—suggesting that the underlying transformation is super-exponential.

To examine whether model sparsity and feature correlation play a role in shaping the transformation T, we conducted a factorial simulation for linear regression with $p = 15$ and $s = p$. We fixed the coefficient vector to $\alpha^* = [3, 0, 3, 0, \ldots, 0]^\mathsf{T}$ with the sparsity level $s^* \in \{2, 8, 15\}$, and varied the correlation parameter $\theta \in \{0,\ 0.5,\ 0.9\}$. Hence the design ranges from independent ($\theta = 0$) to collinear ($\theta = 0.9$) predictors, and from very sparse to fully dense signals. Figure 5 displays the empirical

transformation $\hat{T}$ recovered in each setting.

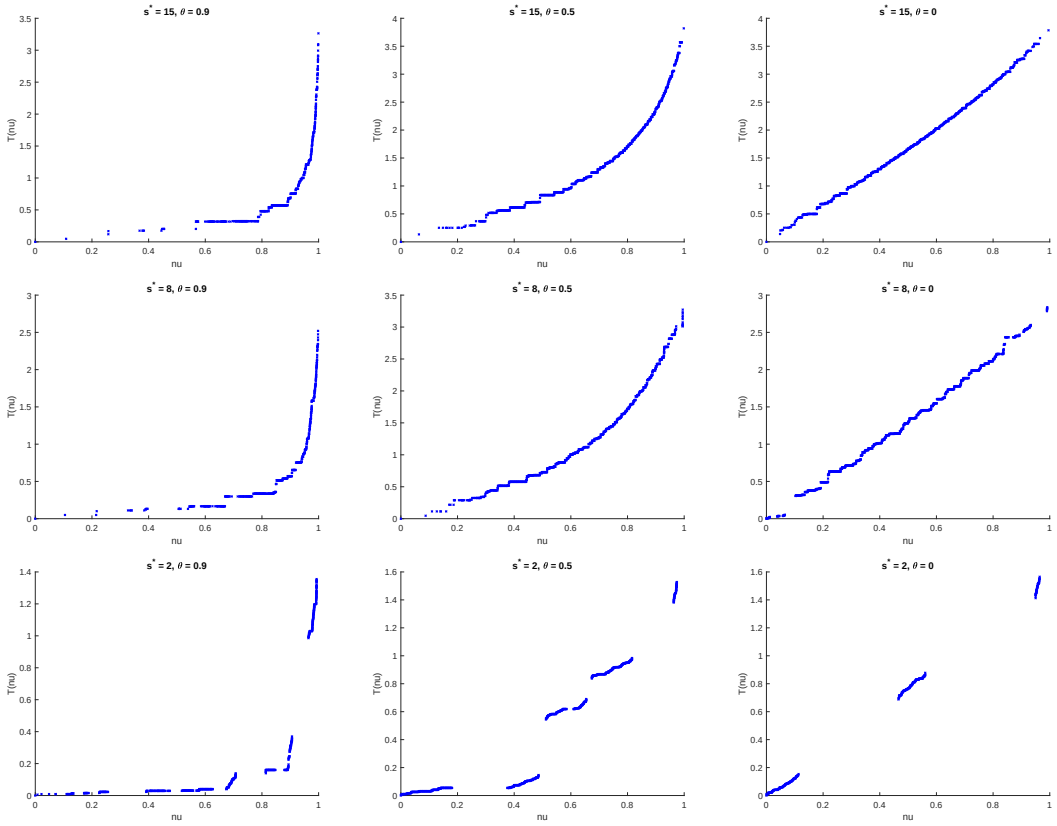


Figure 5: Estimated monotonic transformation $\hat{T}(\nu)$ across varying sparsity levels ($s = 15, 8, 2$, top to down) and feature correlation strengths ($\theta = 0.9, 0.5, 0$, left to right). The RIC criterion identifies $s = 6$ as optimal.

Two main patterns emerge. (i) **Correlation drives curvature:** as θ increases, $\hat{T}$ deviates sharply from linearity, even in dense models. (ii) **Sparsity introduces breaks:** at low s^* the curve becomes piecewise, with segment slopes that differ substantially—another marker of non-additivity. Notably, pronounced nonlinearity appears even in the independent, ultra-sparse case ($\theta = 0$, $s^* = 2$), showing that irrelevant features alone can distort raw worths. Hence a *monotone nonlinear* transformation is indispensable when translating R^2-based worth measures into the additive Shapley framework; without it, either strong correlation or feature irrelevance breaks the required additivity. To our knowledge, these results provide the first evidence that correlated features and the presence of irrelevant features can

substantially undermine the additivity assumption in Shapley frameworks.

4.4 Boston Housing

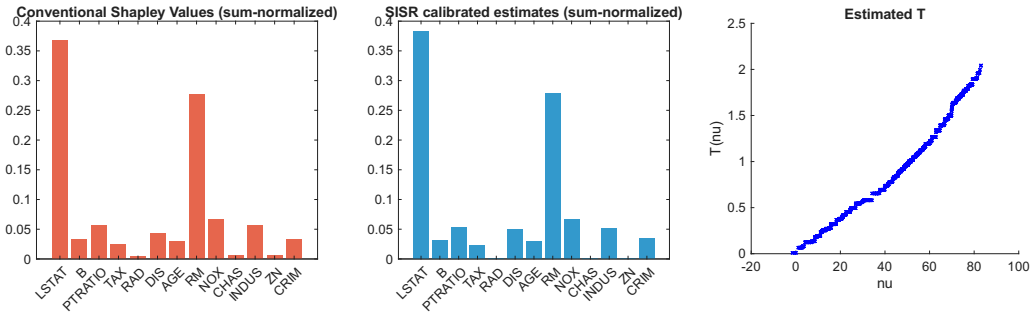
The dataset (Harrison and Rubinfeld, 1978) captures socioeconomic and environmental factors for a suburban area of Boston, including percentage of lower status of the population (`LSTAT`), weighted distances to five Boston employment centers (`DIS`), average number of rooms per dwelling (`RM`), bordering Charles River (`CHAS`), and others, accompanied by the response, the median property value. We trained an XGBoost regression model (Chen and Guestrin, 2016) using the hyper-parameter configuration given by Maniar (2023). After fitting the boosted tree ensemble, we computed feature attributions using the path-traversal algorithm in TreeSHAP (Lundberg et al., 2018). For each feature subset A we consider two payoff functions: the negative mean squared error $\nu_A^{\mathrm{MSE}} = -\,\mathrm{MSE}(A)$ and the negative exponential loss $\nu_A^{\exp} = -\exp(c\,\mathrm{MSE}(A))$ with $c = 50/\max_A \mathrm{MSE}(A)$. Figure 6 displays the resulting Shapley attributions.

The figure contrasts how the two attribution schemes respond to different payoff scales. Under the negative-MSE payoff, SISR has little to adjust—the scale is already compatible with linear additivity. In contrast, for the exponential payoff, SISR produces a highly nonlinear transformation, which compensates for the distortions and preserves essentially the same attribution pattern observed under the MSE scale. The conventional Shapley values shift noticeably: the importance of `DIS` increases from minor to leading, and `CHAS` and several other variables even receive negative attributions. These sign and rank changes substantially alter the qualitative interpretation of the game and reveal the standard procedure's sensitivity to the underlying payoff scale, whereas SISR remains robust.

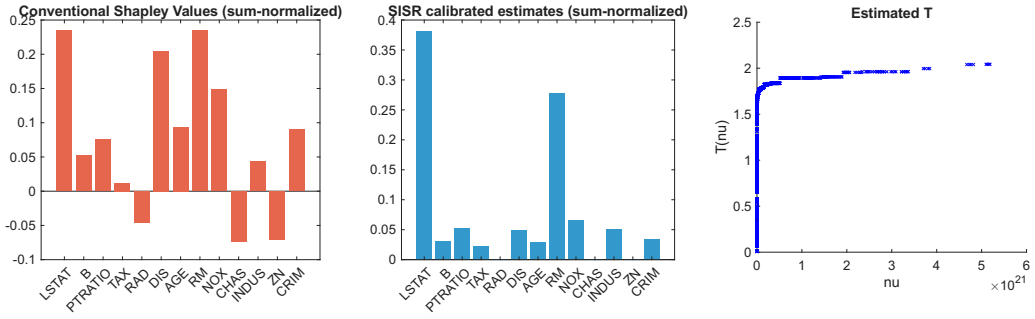
4.5 German GDP

The quarterly German economic dataset spans January 1991 to December 2023 and includes indicators for production, labor, trade, and other sectors, all of which have been stationarized (Jung, 2025). To model GDP growth, we employ a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997). As a type of recurrent neural network, the LSTM is well-suited for capturing long-range temporal dependencies in economic time series. The payoff for any subset of features is evaluated using the SAGE algorithm (Covert et al., 2020). Figure 7 contrasts the conventional Shapley values against SISR-calibrated attributions.

According to the figure, both conventional Shapley values and the SISR method



(a) Negative-MSE payoff



(b) Exponential payoff

Figure 6: Boston housing: feature attributions computed with conventional Shapley and SISR-calibrated Shapley values for the negative-MSE payoff (top) and the exponential payoff (bottom), along with the corresponding estimated monotone transformations.

agree that manufacturing (`Manuf`) is a dominant factor influencing GDP. A notable divergence appears with the German stock index (`GDAXI`), which represents the 40 largest and most actively traded companies on the Frankfurt Stock Exchange—a key indicator for the German economy. While the conventional Shapley approach assigns little importance to the stock index, SISR attributes nearly 50% of the total importance to it. The estimated transformation has two significant features: it is highly nonlinear, challenging the standard additivity assumption, and broken into disjoint segments, indicating underlying sparsity of the Shapley surrogate model (see Section 4.3). Across varying sparsity levels, SISR calibration consistently supports the significance of `GDAXI` for GDP. The finding is supported by independent statistical checks (selection by stepwise AIC and BIC, and a p-value below 0.01), as well as by economic literature identifying the German stock market as a significant indicator of

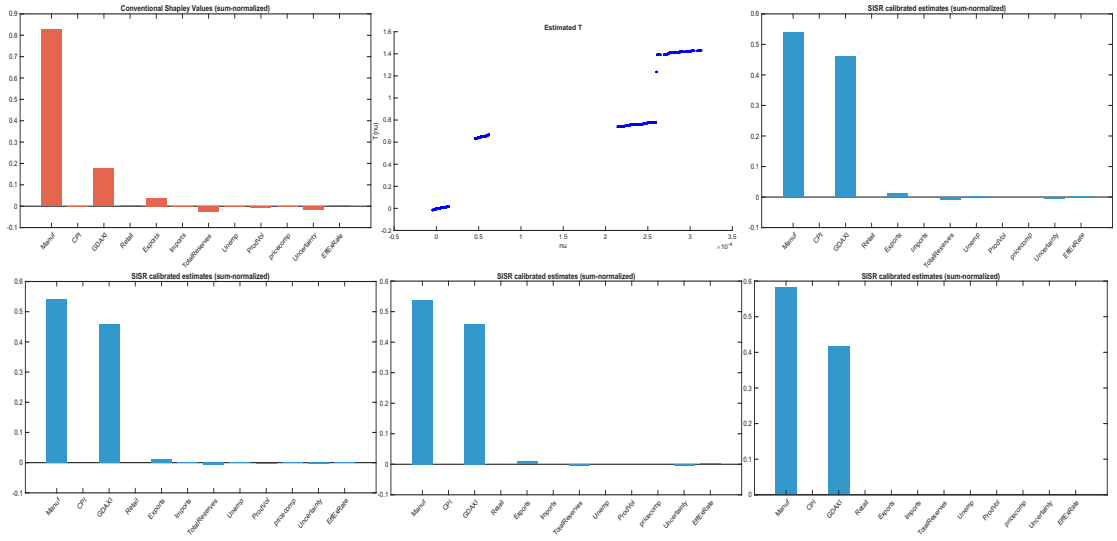


Figure 7: GDP data. Top row (left to right): conventional Shapley values, SISR-estimated transformation and calibrated Shapley values with no sparsity. Bottom row (left to right): calibrated Shapley values at sparsity levels $s = 10, 6, 2$. The RIC criterion identifies $s = 2$ as optimal.

GDP (Drechsel and Scheufele, 2011).

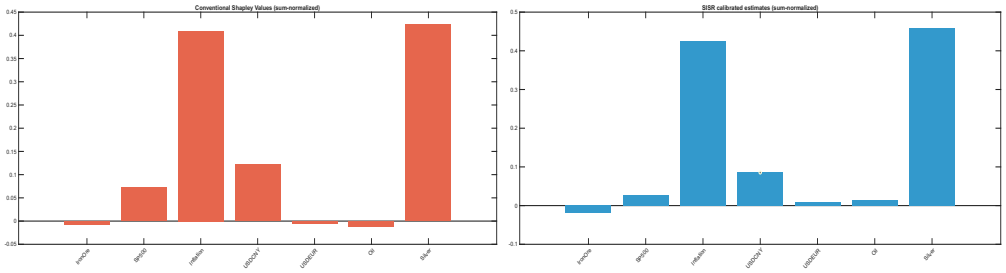
This example cautions against applying *off-the-shelf* Shapley formulas to raw coalition values, as they may mask or distort true importance. By jointly learning a monotone correction and enforcing sparsity, SISR yields importance scores consistent with independent checks.

4.6 Gold Price

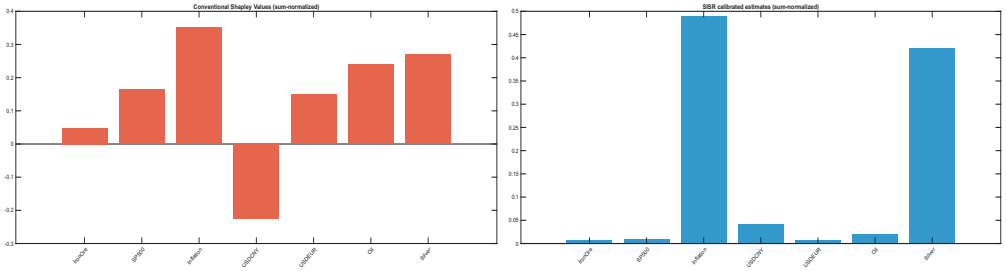
The monthly time-series dataset from Jabeur et al. (2024) contains gold prices and related economic indicators from January 1986 to December 2019. The features are a diverse set of financial and macroeconomic variables, including key commodity prices (silver, oil, iron ore), the S&P 500 index, and US inflation. For this analysis, the response variable is constructed as a binary indicator of monthly price fluctuation, coded as 1 for an increase and 0 otherwise.

We trained an XGBoost classifier (Chen and Guestrin, 2016) and tuned its hyper-parameters with `GridSearchCV` in `scikit-learn` (Pedregosa et al., 2011). Feature-subset worth was measured by two payoff functions: classification accuracy and the negative cross-entropy, via the TreeSHAP algorithm (Lundberg et al., 2018)

for our XGBoost model.



(a) Classification accuracy payoff



(b) Cross-entropy payoff

Figure 8: Gold price: feature attributions computed with conventional Shapley and SISR-calibrated Shapley values for the classification payoff (top) and the cross-entropy payoff (bottom).

According to Figure 8, when payoffs are based on classification accuracy, both conventional Shapley and SISR attributions yield nearly identical results, consistently identifying `silver` and `inflation` as the drivers for predicting monthly gold price rises and falls. In contrast, for the cross-entropy payoff, naive Shapley values become severely distorted: numerous features appear almost as influential as `silver`, and the importance of `USDCNY` even turns negative. SISR attributions, however, remain stable by learning the appropriate transformation. This illustrates the high sensitivity of the standard Shapley procedure to the choice of payoff function, in contrast to SISR, which provides stable and robust attributions across different payoff schemes.

5 Conclusion

Modern economic and financial analysis increasingly relies on complex models—including neural networks and ensemble methods—to capture nonlinear, sequential,

and high-dimensional relationships in macroeconomic and market data. However, these advances have also created a gap in model explainability, limiting the transparency and trust required for decision-making and policy.

This paper proposed *Sparse Isotonic Shapley Regression* (SISR) to address two central flaws of conventional Shapley values in explainable AI: the failure of additive attributions for inherently nonlinear payoffs, and the absence of native sparsity control in high-dimensional settings.

By jointly learning a monotonic transformation via weighted isotonic regression and enforcing sparsity through normalized hard-thresholding, each step of SISR admits a closed-form update and enjoys theoretical guarantees of global convergence. SISR eliminates the need for closed-form functional specification, reduces computational cost, and is well-suited for explaining large, modern economic models.

Our empirical studies demonstrate that SISR yields consistent, sparse, and economically meaningful attributions, even in the presence of irrelevant features and inter-feature dependence—situations where standard Shapley values suffer from severe rank and sign distortions due to their inability to handle nonlinear relationships. SISR stabilizes attributions across varying payoff scales, filters out spurious features, and yields results consistent with economic insights, all of which are essential for model validation, policy assessment, and stakeholder trust.

In summary, SISR provides robust and interpretable attributions for black-box forecasting or risk models and enables the construction of simplified surrogate models that preserve the key predictive and explanatory relationships of the original model, advancing interpretable machine learning for modern economic and financial applications.

References

Algaba, E., Fragnelli, V., and Sánchez-Soriano, J. (2019). *Handbook of the Shapley Value*. Chapman & Hall/CRC Series in Operations Research. CRC Press.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2019). Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. *Proceedings of the 36th International Conference on Machine Learning*.

Au, Q., Herbinger, J., Stachl, C., Bischl, B., and Casalicchio, G. (2022). Grouped feature importance and combined features effect plot. *Data Min. Knowl. Discov.*, 36(4):1401–1450.

Charnes, A., Golany, B., Keane, M., and Rousseau, J. (1988). *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, pages 123–133. Springer Netherlands, Dordrecht.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Cohen, S., Dror, G., and Ruppin, E. (2007). Feature selection via coalitional game theory. *Neural Computation*, 19(7):1939–1961.

Covert, I., Lundberg, S., and Lee, S.-I. (2020). Explaining by removing: A unified framework for model explanation. In *Advances in Neural Information Processing Systems*.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24.

Drechsel, K. and Scheufele, R. (2011). Forecasting german gdp using unrestricted mixed-data sampling. *Journal of Forecasting*, 31(5):412–433.

Duan, H. and Okten, G. (2025). Derivative-based shapley value for global sensitivity analysis and machine learning explainability. *International Journal for Uncertainty Quantification*, 15(1):1–16.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Fryer, D., Strümke, I., and Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*, 9:144352–144360.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jabeur, S. B., Mefteh-Wali, S., and Viviani, J.-L. (2024). Forecasting gold price with the xgboost algorithm and shap interaction values. *Annals of Operations Research*, 334(1):679–699.

Jothi, N., Husain, W., and Rashid, N. A. (2021). Predicting generalized anxiety disorder among women using shapley value. *Journal of Infection and Public Health*, 14(1):103–108.

Jung, Y. (2025). Machine learning-based estimation of monthly gdp.

Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ma, S. and Tourani, R. (2020). Predictive and causal implications of using shapley value for model interpretation. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, volume 127 of *Proceedings of Machine Learning Research*, pages 23–38. PMLR.

Maniar, A. (2023). Xgboost model optimization – boston housing. https://www.kaggle.com/code/advikmaniar/xgboost-model-optimization-94-boston-housing/notebook.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Strumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665.