Shapley Curves: A Smoothing Perspective

Ratmir Miftachov

Georg Keilbar

Wolfgang Härdle

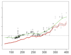
Seminar Mathematische Statistik
WIAS Berlin
May 8, 2024

How to quantify importance of a variable?

- Many well-performing estimation methods such as random forests are not intrinsically interpretable

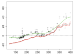
- Variable importance measures try to increase interpretability

- Statistical understanding is still limited
  - Scornet (2022), mean decrease impurity (MPI) is ill-defined in the presence of feature dependence or interactions
  - Benard et al. (2022), mean decrease accuracy (MDA) does not target a meaningful quantity under dependence

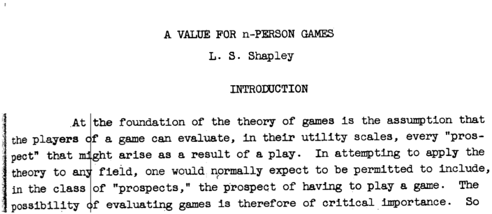


Statistical Modeling: The Two Cultures

- Breiman (2000, Statistical Science) describes two distinct paradigms of statistical modeling
- **Data modeling** "*[...] starts with assuming a stochastic data model for the inside of the black box.*"
- **Algorithmic modeling** "*[...] find a function $f(x)$ – an algorithm that operates on x to predict the responses y.*"



Shapley Value



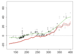
A VALUE FOR n-PERSON GAMES

L. S. Shapley

INTRODUCTION

At the foundation of the theory of games is the assumption that the players of a game can evaluate, in their utility scales, every "prospect" that might arise as a result of a play. In attempting to apply the theory to any field, one would normally expect to be permitted to include, in the class of "prospects," the prospect of having to play a game. The possibility of evaluating games is therefore of critical importance. So

- Origins in cooperative game theory (Shapley, 1953)
- Offers a unique distribution among players of surplus generated by them
- Satisfies a number of favorable properties

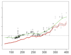


Shapley and Variable Importance

- Similarly, a variable can contribute to some value function which can be a predictive measure (MSE, R^2) or the output of a prediction algorithm

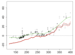
- Global vs. local measures

- Main advantage of Shapley is the additivity of contributions



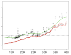
Some Related Literature

- Owen (2014) applies Shapley values to analyze explained variance
- Lundberg and Lee (2017) introduce approximation method KernelSHAP
- Aas et al. (2021) consider the problem of Shapley value estimation under feature dependence
- Benard et al. (2022) study asymptotic properties of Shapley effects based on random forests
- Williamson and Feng (2020): efficient nonparametric statistical inference on population feature importance using Shapley values



Perspective of this Paper

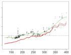
- Shapley formulation based on the conditional mean
- Definition and estimation of population-level Shapley curves as measure of true variable importance
- Shapley curves are *local* measures, $\phi_j(x) : \mathbb{R}^d \rightarrow \mathbb{R}$
- Nonparametric perspective



Model Setup and Notation

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

- $\mathrm{E}(\varepsilon_i | X_i) = 0$ and $m(x) = \mathrm{E}(\mathrm{Y_i} | \mathrm{X_i} = \mathrm{x})$ is twice continuously differentiable
- $X_i \sim F$ iid, with continuous density f
- $X_i \in \mathbb{R}^d$
- $N \overset{\text{def}}{=} \{1, \ldots, d\}$ and let $\mathcal{S}$ denote the power set of N
- For $s \in \mathcal{S}$, we write $f_{X_{-s}|X_s}(x_{-s}|x_s)$ for the conditional density of X_{-s} given $X_s = x_s$



Population Shapley Curves

- Population Shapley curve of variable X_j is a function, $\phi_j(\cdot) : \mathbb{R}^d \to \mathbb{R}$,

$$\phi_j(x) = \sum_{s \subseteq N \setminus j} \frac{1}{d} \binom{d-1}{|s|}^{-1} \left\{ m_{s \bigcup j}(x_{s \cup j}) - m_s(x_s) \right\} \qquad (2)$$

- for $j \in N$, where the components are defined as

$$\begin{aligned} m_s(x_s) &= \int m(x) f_{X_{-s}|X_s}(x_{-s}|x_s)\, dx_{-s} \\ &= \mathsf{E}\left(Y|X_s = x_s\right) \end{aligned} \qquad (3)$$

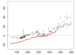
- for $s = N$, note that $m_N(x) = m(x)$ and $m_\emptyset = \mathsf{E}(Y)$
- Additivity property: $m(x) - \mathsf{E}(Y) = \sum_{j=1}^{d} \phi_j(x)$

Example 1: Additive Interaction Model

$$m(x_1, x_2) = g_1(x_1) + g_2(x_2) + g_{12}(x_1, x_2)$$

⊡ $\phi_1(x)$ depends on direct effect, its share of the interaction effect and on dependence among features

$$\begin{aligned}\phi_1(x) = g_1(x_1) &+ \frac{1}{2}\left[\mathsf{E}\left\{g_2\left(X_2\right)|X_1 = x_1\right\} - \mathsf{E}\left\{g_1\left(X_1\right)|X_2 = x_2\right\}\right] \\ &+ \frac{1}{2}\left[g_{12}(x_1, x_2) - \mathsf{E}\left\{g_{12}\left(X_1, x_2\right)|X_2 = x_2\right\}\right. \\ &+ \left.\mathsf{E}\left\{g_{12}\left(x_1, X_2\right)|X_1 = x_1\right\}\right] - \frac{1}{2}\mathsf{E}(Y)\end{aligned}$$



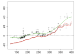
Example 2: Threshold Regression

$$m(x) = \{\psi + \theta \mathbf{I}(x_2 \leq C)\}\, x_1$$

- C, ψ, θ are scalar parameters
- Dagenais (1969), Hansen (2000)
- Under feature mean independence we have

$$\phi_1(x) = \left\{\psi + \frac{1}{2}\theta \mathbf{I}(x_2 \leq C) + \frac{1}{2}\theta F_{X_2}(C)\right\} \{x_1 - \mathsf{E}(X_1)\},$$

$$\phi_2(x) = \frac{1}{2}\theta \{\mathbf{I}(x_2 \leq C) - F_{X_2}(C)\} \{x_1 + \mathsf{E}(X_1)\}$$

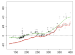


Estimation of Shapley Curves

- Shapley curves as weighted sum of conditional mean functions

$$\begin{aligned}\phi_j(x) &= \sum_{s \subseteq N} \operatorname{sgn}\{j \in s\} \frac{1}{d} \binom{d-1}{|s| - \mathbf{I}\{j \in s\}}^{-1} m_s(x_s) \qquad (4)\\ &= \sum_{s \subseteq N} \omega_{j,s} m_s(x_s),\end{aligned}$$

- $\mathbf{I}(\cdot)$ is the indicator function, $\operatorname{sgn}(\cdot)$ is the sign function



Two Estimation Approaches

- Plug-in estimation using weighted sum (4)
- *Component-based approach*
  - separate estimation of the conditional mean functions, $m_s(x_s)$, $s \in S$
- *Integration-based approach*
  - only one estimation of the full conditional mean function, $m(x)$
  - estimation of subset-specific conditional mean functions by integrating out variables not in subset using pilot estimator
  - related to marginal integration literature: Linton and Nielsen (1995), Tjøstheim and Auestad (1994), Fan, Härdle and Mammen (1998)

Reducing Computational Complexity

- Number of subsets is 2^d, which is computationally infeasible in high dimensions
- Computational cost can be reduced by using a weighted least squares procedure (Lundberg and Lee, 2017)
- Idea: sample subsets from a distribution Q_0 implied by the Shapley weights

Weighted Least Squares Formulation

$$\Phi(x) = \arg\min_{\beta\in\mathbb{R}^{d+1}} \mathrm{E}_{\mathrm{Q}_0}\left\{\mathrm{z}(\mathrm{s})^{\top}\beta - \mathrm{m}_{\mathrm{s}}(\mathrm{x}_{\mathrm{s}})\right\}^2$$

- $\Phi(x) = (\phi_0(x), \phi_1(x), \ldots, \phi_d(x))^{\top} \in \mathbb{R}^{d+1}$
- $z(s) \in \{0,1\}^{d+1}$ with elements $z_j(s) = \mathbf{I}(j \in s)$
- Q_0 assigns probability mass $\binom{d-2}{|s|-1}^{-1}$ for $s \in S \setminus \{\emptyset, N\}$ and 1 for $s \in \{\emptyset, N\}$ (properly normalized)
- In practice: replace $\mathrm{E}_{\mathrm{Q}_0}$ by the empirical distribution of m sampled subsets, replace $m_s(x_s)$ by an estimate
- Introduces additional estimation error of order $\mathcal{O}_p(m^{-1/2})$

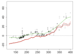
Component-Based Approach

- Regress Y_i on $X_{s,i}$ for all subsets of variables (2^d regression equations)

$$Y_i = m_s(X_{s,i}) + \varepsilon_{s,i}, \quad i = 1, \ldots, n; \quad s \in \mathcal{S}$$

- with $\mathrm{E}(\varepsilon_{s,i} | X_{s,i}) = 0$
- Component-based estimator of Shapley curve of variable X_j:

$$\widehat{\phi}_j(x) = \sum_{s \subseteq N \setminus j} \frac{1}{d} \binom{d-1}{|s|}^{-1} \left\{ \widehat{m}_{s \bigcup j}(x_{s \cup j}) - \widehat{m}_s(x_s) \right\}$$

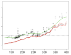


Component-Based Approach

- Local linear method (Tsybakov, 1986; Fan, 1993)
- Let $Y = (Y_1, \dots, Y_n)^\top$, $Z_s = (z_{s,1}, \dots, z_{s,n})^\top$, $z_{s,i} = (1, X_{s,i}^\top)$
- $K_s = \operatorname{diag}[\{h_s^{-|s|} \prod_{l=1}^{|s|} k(h_s^{-1}(X_{s,il} - x_{s,l}))\}_{i=1}^{n}]$, where k is a one-dimensional kernel function
- h_s are bandwidth parameters

$$\widehat{\beta}_s(x_s) = \begin{pmatrix} \widehat{\beta}_{s,0}(x_s) \\ \widehat{\beta}_{s,1}(x_s) \end{pmatrix} = \left(Z_s^\top K_s Z_s\right)^{-1} Z_s^\top K_s Y,$$

- Local linear estimator is $\widehat{m}_s(x_s) = \widehat{\beta}_{s,0}(x_s)$



Global Consistency of Component-based Approach

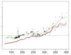
Proposition (1)

Let $\widehat{\phi}_j(x)$ be the component-based estimator with components estimated via the local linear method with bandwidths $h_s \sim n^{-\frac{1}{4+|s|}}$. Then we have under Assumptions A1 – A3, as n goes to infinity,

$$\text{MISE}\left\{\widehat{\phi}_j(x), \phi_j(x)\right\} = \mathcal{O}\left(n^{-\frac{4}{4+d}}\right).$$

⊡ MISE is dominated by MISE of the full model:

$$\text{MISE}\left\{\widehat{\phi}_j(x), \phi_j(x)\right\} = \frac{1}{d^2}\,\text{MISE}\left\{\widehat{m}(x), m(x)\right\} + \mathcal{O}(n^{-\frac{4}{4+d}})$$



Asymptotic Normality

Theorem (2)

Let the conditions of Proposition 1 hold and let $h_m \sim n^{-\frac{1}{4+d}}$ denote the optimal bandwidth of the full model. Then we have, for a point x in the interior of $\mathcal{X}$, as n goes to infinity,

$$\sqrt{nh_m^d}\left\{\widehat{\phi}_j(x) - \phi_j(x)\right\} = \sqrt{nh_m^d}\,\frac{1}{d}\left\{\widehat{m}(x) - m(x)\right\} + o_p(1) \xrightarrow{\mathcal{L}} N\left(B(x), V(x)\right),$$

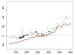
where the asymptotic bias and the asymptotic variance are given as

$$B(x) = \frac{1}{d}\,\frac{\mu_2(k)}{2}\sum_{l=1}^{d}\frac{\partial^2 m(x)}{\partial x_l^2} \quad \text{and} \quad V(x) = \frac{1}{d^2}\|k\|_2^2\,\frac{\sigma^2(x)}{f(x)},$$

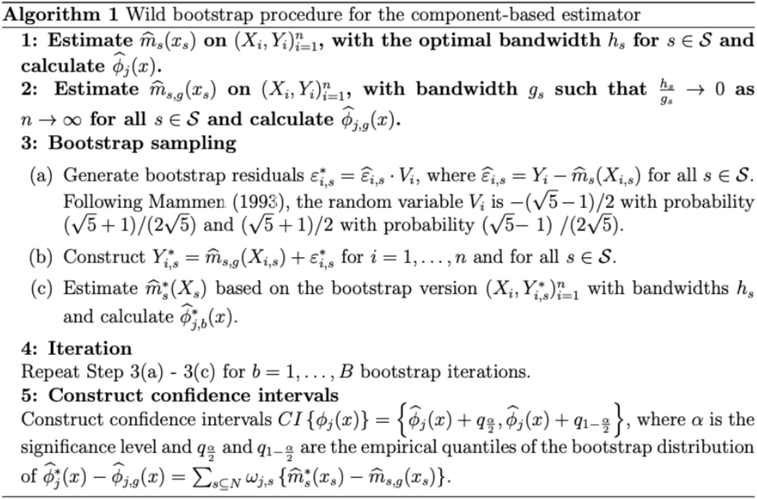
respectively and $\|k\|_2^2 = \int k^2(s)ds$ denotes the squared L_2 norm of k.

Wild Bootstrap Procedure

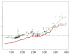
- Inference based on asymptotic distribution unreliable in finite samples (Härdle and Marron, 1991)
- Even worse than usual in our case (lower order terms)
- Instead rely on bootstrap confidence intervals
- Wild bootstrap (Wu, 1986; Mammen, 1993; Härdle and Mammen, 1993)



Wild Bootstrap Procedure



```
Algorithm 1 Wild bootstrap procedure for the component-based estimator
  1: Estimate $\widehat{m}_s(x_s)$ on $(X_i, Y_i)_{i=1}^n$, with the optimal bandwidth $h_s$ for $s \in \mathcal{S}$ and calculate $\widehat{\phi}_j(x)$.
  2: Estimate $\widehat{m}_{s,g}(x_s)$ on $(X_i, Y_i)_{i=1}^n$, with bandwidth $g_s$ such that $\frac{h_s}{g_s} \to 0$ as $n \to \infty$ for all $s \in \mathcal{S}$ and calculate $\widehat{\phi}_{j,g}(x)$.
  3: Bootstrap sampling
    (a) Generate bootstrap residuals $\varepsilon_{i,s}^* = \widehat{\varepsilon}_{i,s} \cdot V_i$, where $\widehat{\varepsilon}_{i,s} = Y_i - \widehat{m}_s(X_{i,s})$ for all $s \in \mathcal{S}$. Following Mammen (1993), the random variable $V_i$ is $-(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}- 1) /(2\sqrt{5})$.
    (b) Construct $Y_{i,s}^* = \widehat{m}_{s,g}(X_{i,s}) + \varepsilon_{i,s}^*$ for $i = 1, \dots, n$ and for all $s \in \mathcal{S}$.
    (c) Estimate $\widehat{m}_s^*(X_s)$ based on the bootstrap version $(X_i, Y_{i,s}^*)_{i=1}^n$ with bandwidths $h_s$ and calculate $\widehat{\phi}_{j,b}^*(x)$.
  4: Iteration
  Repeat Step 3(a) - 3(c) for $b = 1, \dots, B$ bootstrap iterations.
  5: Construct confidence intervals
  Construct confidence intervals $CI\{\phi_j(x)\} = \left\{\widehat{\phi}_j(x) + q_{\frac{\alpha}{2}}, \widehat{\phi}_j(x) + q_{1-\frac{\alpha}{2}}\right\}$, where $\alpha$ is the significance level and $q_{\frac{\alpha}{2}}$ and $q_{1-\frac{\alpha}{2}}$ are the empirical quantiles of the bootstrap distribution of $\widehat{\phi}_j^*(x) - \widehat{\phi}_{j,g}(x) = \sum_{s \subseteq N} \omega_{j,s} \{\widehat{m}_s^*(x_s) - \widehat{m}_{s,g}(x_s)\}$.
```

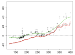


Some Background on Wild Bootstrap

- Real world: original sample $\{X_i, Y_i\}_{i=1}^n$, interested in conditional distribution $P^{Y|X}$
- Bootstrap world: bootstrap sample $\{X_i, Y_i^*\}_{i=1}^n$, interested in bootstrap distribution P^*, conditional on data $\{X_i, Y_i\}_{i=1}^n$
- $\mathrm{E}(\varepsilon_\mathrm{i}^*) = 0$, $\mathrm{E}(\varepsilon_\mathrm{i}^*)^2 = \widehat{\varepsilon}_\mathrm{i}^2$, $\mathrm{E}(\varepsilon_\mathrm{i}^*)^3 = \widehat{\varepsilon}_\mathrm{i}^3$
- Oversmoothing g such that $h/g \to 0$ since

$$\mathrm{E}^{\mathrm{Y|X}}\{\widehat{\mathrm{m}}_\mathrm{h}(\mathrm{x} - \mathrm{m}(\mathrm{x})\} \approx h^2 \frac{\mu_2(k)}{2} \sum_{j=1}^{d} \frac{\partial^2 m(x)}{\partial x_j^2}$$

$$\mathrm{E}^*\{\widehat{\mathrm{m}}_\mathrm{h}^*(\mathrm{x}) - \widehat{\mathrm{m}}_\mathrm{g}(\mathrm{x})\} \approx h^2 \frac{\mu_2(k)}{2} \sum_{j=1}^{d} \frac{\partial^2 \widehat{m}_g(x)}{\partial x_j^2}$$



Bootstrap Consistency

Proposition (3)

Under Assumptions A1 – A4 *and let $P^{Y|X}$ denote the conditional distribution and P^* denote the bootstrap distribution. Then we have, for a point x in the interior of $\mathcal{X}$ and $z \in \mathbb{R}$, as n goes to infinity*

$$\left| P^{Y|X}\left[\sqrt{nh_m^d}\left\{\widehat{\phi}_j(x) - \phi_j(x)\right\} < z\right] - P^*\left[\sqrt{nh_m^d}\left\{\widehat{\phi}_j^*(x) - \widehat{\phi}_{g,j}(x)\right\} < z\right]\right| \to 0.$$

Integration-Based Approach

- Involves a single regression equation
- $m_s(x)$ is obtained by integrating out variables not in s
- Similar to marginal integration estimator of Linton and Nielson (1995)
- Under feature independence (Lundberg and Lee, 2017):

$$\widetilde{m}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(X_s = x_s, X_{-s,i})$$

- Under feature dependence:

$$\widetilde{m}_s(x_s) = \int_{\mathbb{R}^{d-|s|}} \widehat{m}(x)\widehat{f}_{X_{-s}|X_s}(x_{-s}|x_s)dx_{-s}$$

- $\widehat{f}_{X_{-s}|X_s}(x_{-s}|x_s)dx_{-s}$ is an estimator of the conditional density

Asymptotic Normality

Theorem (5)

Under regularity conditions, let $\widetilde{\phi}_j(x)$ be the integration-based estimator with known density and a pilot estimator based on local linear estimation with bandwidth $h_m \sim n^{-\frac{1}{4+d}}$. Then we have for a point x in the interior of $\mathcal{X}$ as n goes to infinity,

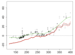
$$\sqrt{nh_m^d}\left\{\widetilde{\phi}_j(x) - \phi_j(x)\right\} = \sqrt{nh_m^d}\sum_{s\subseteq N}\omega_{j,s}\left\{\widetilde{m}_s(x_s) - m_s(x_s)\right\} \xrightarrow{\mathcal{L}} N\left(B_{int}(x), V(x)\right),$$

where the asymptotic bias term is

$$B_{int}(x) = \frac{\mu_2(k)}{2}\sum_{s\subseteq N}\omega_{j,s}\left\{\sum_{l=1}^{d}\int_{X_{-s}}\frac{\partial^2 m(x)}{\partial x_l^2} f_{X_{-s}|X_s}(x_{-s}|x_s)dx_{-s}\right\} \qquad (5)$$

and the asymptotic variance term is

$$V(x) = \frac{1}{d^2}\|k\|_2^2\frac{\sigma^2(x)}{f(x)}.$$



Bias of Integration-Based Approach

- Bandwidth of the pilot estimator, $h_m \sim n^{-\frac{1}{4+d}}$, balances the squared bias and variance of the full model
- However, all components are based on this pilot estimator and thus rely on the same bandwidth
- For all other subsets s, we have $|s| < d$ which leads to oversmoothing

$$\text{Bias}^2\left\{\widetilde{m}_s(x_s)\right\} = \mathcal{O}(n^{-\frac{4}{4+d}})$$
$$\text{Var}\left\{\widetilde{m}_s(x_s)\right\} = \mathcal{O}(n^{-\frac{(4+d-|s|)}{4+d}})$$

- Implications for other estimation approaches (e.g., tuning parameters in random forests)

Remark on Theorem 5

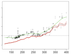
- Derivation hinges on comparison of local linear and 'internal estimator' based on known density

$$\widehat{m}_I(x) = n^{-1} \sum_{i=1}^{n} h_m^{-d} \prod_{j=1}^{d} k\{h_m^{-1}(x_j - X_{i,j})\}/f(X_i)$$

- Integrated difference is

$$\int \{\widehat{m}(x) - \widehat{m}_I(x)\} dF_{X_{-s}|X_s}(x_{-s}|x_s) = \mathcal{O}_p(n^{-1} h_m^{-d})$$

- Term is lower order even if $d > 0$ since $h_m \sim n^{-1/(4+d)}$
- No problems as in Linton and Nielsen (1995)



Simulation: Estimation Accuracy

sim_consistency

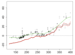
DGP 1: Additive $m(x) = -\sin(2x_1) + \cos(2x_2) + x_3$

DGP 2: Interaction $m(x) = -\sin(2x_1) + \cos(3x_2) + 0.5x_3 + 2\cos(x_1)\sin(2x_2)$

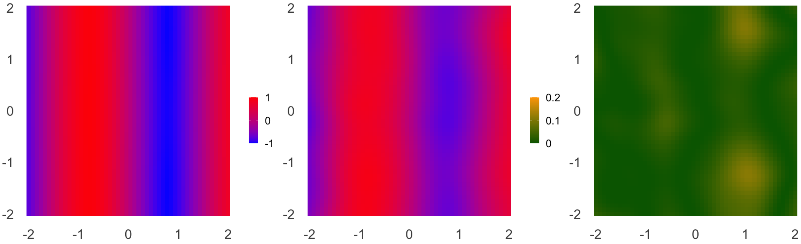
- Regressors are zero mean Gaussian with variance $\sigma^2 = 4$ and correlation ρ
- ε is standard normal or $t(5)$
- Compare component-based and integration-based approach (MISE)
- Bandwidth selected via cross-validation

Simulation: Estimation Accuracy

DGP	ρ	n	$\widehat{\phi}_1$	$\widetilde{\phi}_1$	$\widehat{\phi}_2$	$\widetilde{\phi}_2$	$\widehat{\phi}_3$	$\widetilde{\phi}_3$
Additive	0	300	8.84	14.79	8.53	8.72	3.05	0.34
		500	4.86	5.37	5.47	6.30	1.91	0.22
		1000	3.04	3.29	3.48	3.87	1.09	0.11
		2000	1.83	1.91	2.13	2.43	0.66	0.07
	0.8	300	7.29	7.81	7.99	8.87	2.69	1.27
		500	5.09	5.41	5.65	6.05	1.86	0.83
		1000	3.37	3.38	3.46	3.79	1.27	0.52
		2000	2.24	2.10	2.22	2.37	0.87	0.34
Interactive	0	300	9.31	12.76	11.45	14.15	1.74	0.40
		500	5.68	7.51	6.74	7.86	0.87	0.23
		1000	3.20	4.20	3.92	4.56	0.51	0.13
		2000	1.90	2.45	2.22	2.63	0.30	0.08
	0.8	300	10.85	12.04	11.89	14.34	3.16	1.81
		500	7.59	7.82	7.98	9.79	2.20	1.27
		1000	4.99	4.70	4.79	5.94	1.50	0.79
		2000	3.27	2.86	3.00	3.57	1.05	0.49



Simulation: Estimation Accuracy



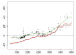
Heatmaps for $m(x) = -\sin(2x_1) + \cos(2x_2) + x_3$ with Gaussian error terms with $n = 2000$ for the first variable at $x_3 = 0$. Left: Population Shapley Curve. Centre: Componend-based estimated Shapley Curve. Right: Squared residuals between estimated and population curve.

Simulation: Bootstrap Coverage

coverage

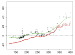
DGP 3: $m(x) = -\sin(2x_1) + 0.1x_2 + 2\cos(x_1)\sin(x_2)$

	Variable 1					
n	$N(0,1)$			$t(5)$		
	0.15	0.1	0.05	0.15	0.1	0.05
100	0.79	0.85	0.90	0.79	0.86	0.92
250	0.82	0.87	0.93	0.83	0.87	0.92
500	0.82	0.86	0.93	0.83	0.87	0.92
1000	0.82	0.87	0.93	0.83	0.87	0.91
2000	0.82	0.87	0.93	0.82	0.87	0.94
4000	0.85	0.90	0.95	0.82	0.88	0.94

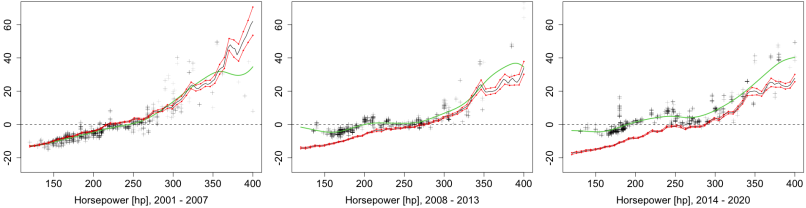


Empirical Application

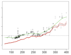
- Extensive vehicle price data for the U.S (2001 – 2020)
- Collected by the Argonne National Laboratory provided by Moawad et al. (2021)
- 38,435 vehicle prices and characteristics
- Focus on length, weight and horsepower



Empirical Application

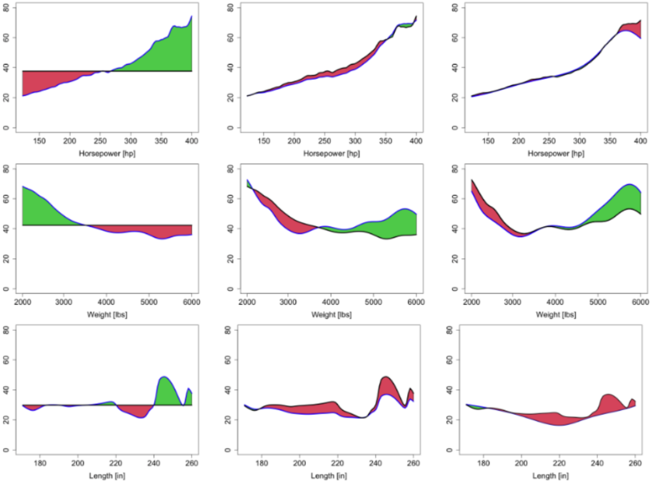


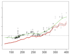
Estimated Shapley curves (black line), confidence intervals (red lines), estimated SHAP values (black crosses) and smoothed curve based on these values (green line). Vehicle length fixed at 190 inches and vehicle weight at $3,500$. application



Empirical Application

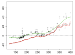
seq_plots





Generalized Shapley Curves

- Shapley values often applied to other problems than mean regression
- E.g. quantile regression, treatment effect estimation
- Introduce generalization of the notion of Shapley curves
- Why relevant? Questionable approaches in practice
- Local estimating equations (Carroll et al., 1998; Athey et al., 2019)
- Local likelihood methods (Tibshirani and Hastie, 1986; Fan et al., 1998; Spokoiny and Polzehl, 2006)

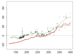


Generalized Shapley Curves

$$\mathsf{E}\left\{\psi_{\theta(\mathsf{x})}(\mathsf{Y_i})|\mathsf{X_i}=\mathsf{x}\right\}=0$$

- Parameter of interest $\theta(x)$ defined as a solution to a local estimation equation
- ψ is some score function
- Example ML estimation: $\psi_{\theta(x)}(Y_i)=\nabla\log\{f_{\theta(x)}(Y_i)\}$
- Problem of naively applying the integration-based approach

$$\theta_s(x_s)\neq \mathsf{E}\left\{\theta(\mathsf{x})|\mathsf{X_{-s}}=\mathsf{x_{-s}}\right\}$$

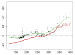


Generalized Shapley Curves

- Generalized Shapley curves $\theta_j^\theta(x) : \mathbb{R}^d \to \mathbb{R}$,

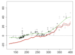
$$\phi_j^\theta(x) = \sum_{s \subseteq N \setminus j} \frac{1}{d} \binom{d-1}{|s|}^{-1} \left\{\theta_{s \bigcup j}(x_{s \cup j}) - \theta_s(x_s)\right\}$$

- with $\theta_s(x_s)$ being a solution to $\mathrm{E}\left\{\psi_{\theta_{\mathrm{s}}(\mathrm{x})}(\mathrm{Y}_\mathrm{i}) | \mathrm{X}_{\mathrm{s,i}} = \mathrm{x}_\mathrm{s}\right\} = 0$ for $s \in \mathcal{S}$



Conclusion

- Definition of Shapley curves as measure of ‘true’ variable importance on population level
- Study of asymptotic properties of both the component- and integration-based approach in a nonparametric setting
- Taylored wild bootstrap procedure for inference
- Preprint available on arXiv: https://arxiv.org/abs/2211.13289



Shapley Curves: A Smoothing Perspective

Ratmir Miftachov
Georg Keilbar
Wolfgang Härdle

Seminar Mathematische Statistik
WIAS Berlin
May 8, 2024

Assumptions

Proposition 1 Proposition 3 Theorem 5

Assumption (A1)

(i) The support of X is $\mathcal{X}$.
(ii) The density f of X is bounded, bounded away from zero, and twice continuously differentiable on $\mathcal{X}$.
(iii) $\mathrm{Var}(\varepsilon|x) = \sigma^2(x) < \infty$ *for all* $x \in \mathcal{X}$.
(iv) $\mathrm{E}(|Y|^{2+\delta}|X = x) < \infty$ *for some* $\delta > 0$.

Assumption (A2)

Assume $m(x)$ belongs to $\mathcal{M}_d$, the space of d-dimensional twice continuously differentiable functions.

Assumptions

Proposition 1 Proposition 3 Theorem 5

Assumption (A3)

Assume $k(\cdot)$ *is a univariate twice continuously differentiable probability density function symmetric about zero and* $\int s^2 k(s) ds = \mu_2(k) < \infty$ *and* $\int k^{2+\delta}(s) ds < \infty$ *for some* $\delta > 0$*.*

Assumption (A4)

Assume that the conditional variance $\sigma^2(x)$ *is twice continuously differentiable and* $\sup_x \mathsf{E}(\varepsilon^3 | X = x) < \infty$*.*