An exact game-theoretic variable importance index for generalized additive models

Amir Khorrami Chokami
ESOMAS Department, University of Turin,
and Collegio Carlo Alberto, Turin, Italy,
and
Giovanni Rabitti
Department of Actuarial Mathematics and Statistics, Heriot-Watt University,
and Maxwell Institute for Mathematical Sciences, Edinburgh, U.K.

March 4, 2024

Abstract

Generalized Additive Models (GAMs) are widely used in statistics. In this work, we aim to tackle the challenge of identifying the most influential variables in GAMs. To accomplish this, we introduce a variance allocation approach based on the Shapley value. We derive a closed-form expression for this importance index, which allows for their computation on high-dimensional datasets and with any dependence structure. We discuss the practical implication that when a variable's importance is negligible, it can be safely eliminated from the GAM, simplifying the model. Through our case studies, we demonstrate that Shapley values offer more informative insights than p-values in terms of ranking the importance of variables. All the code is available online in the supplementary material.

Keywords: Cooperative game theory, Global sensitivity analysis, Statistical significance, Variable importance, Confidence intervals.

The present manuscript contains 4379 words (Overleaf word count feature).

1 Introduction

Generalized Additive Models (GAMs) are a flexible class of statistical models used for regression and classification tasks (Hastie and Tibshirani, 1990; Wood, 2006). They extend the generalized linear models by incorporating non-linear functions of the variables. A GAM has the form

$$g(\mathbb{E}[Y|\boldsymbol{X}]) = f_0 + f_1(X_1) + f_2(X_2) + ... + f_n(X_n) \tag{1}$$

where $g(\cdot)$ is the link function, Y is the response, f_0 is a constant, the f_j is a smooth function of the variables X_j, $j = 1, 2, \ldots, n$, and $\boldsymbol{X} = (X_1, X_2, ..., X_n)$.

The research on GAMs is currently active, with increasing attention being paid to GAMs in the high-dimensional context. For instance, Wood et al. (2015) utilize penalized regression splines to represent the functions f_j's, making it possible to apply GAMs to large datasets. Wood et al. (2017) enhance this penalization approach for estimating GAMs in even larger datasets. Furthermore, Fasiolo et al. (2020) introduce a set of tools for visualizing and inspecting high-dimensional GAMs.

Various studies have delved into the importance of variables in GAMs. For example, Marra and Wood (2011) use shrinkage methods to identify active variables, while Xie and Luo (2022) consider the ratio of the average value of coefficients to its estimated standard deviation as a measure of variable importance for generalized additive models. In the classical reference of Hastie et al. (2009), variables in GAMs are deemed significant based on their p-values (see page 302). Nevertheless, it is still a subject of ongoing debate whether variables that are statistically significant necessarily have an impact on the response (Lo et al., 2015). Thus, the task of identifying and ranking the most significant variables in GAMs according to their importance remains understudied in our opinion.

Global sensitivity analysis (Wagner, 1995) encompasses a variety of statistical and probabilistic methods for ranking the variables of a model based on their impact on the response. This process is often referred to as factor prioritization in the sensitivity analysis literature (Saltelli et al., 2008; Borgonovo and Plischke, 2016). One increasingly popular technique for sensitivity analysis is the Shapley value, originally introduced in economic game theory

by Shapley (1953). The Shapley value aims to fairly distribute the total value generated by a group of players to each player. Alternative Shapley allocations can be produced by varying the total value to be attributed. By associating players with variables and the total value with a particular statistical measure, the Shapley value provides a flexible method for importance attribution in statistics. Grömping (2007) and Huettner and Sunder (2012) apply the Shapley value to allocate the R^2 goodness-of-fit index among predictors in a linear regression model. Owen (2014) use it to allocate the explained variance of a computer experiment among its inputs. Seiler et al. (2023) consider the Shapley value in the context of subject's uniqueness identification.

Closed-form expressions for the Shapley values are rare, with only a few exceptions. For example, if the model is linear and the variables are independent, Lundberg and Lee (2017) provide an analytical expression for the Shapley values of individual variables. Owen and Prieur (2017) find a close-form expression for Shapley values for linear models with two variables and a Farlie-Gumbel-Morgenstern copula. Colini-Baldeschi et al. (2018) derive an analytical expression for Shapley values that apply to any dependence structure of variables when allocating variance in a linear portfolio. Galeotti and Rabitti (2024) extend the result of Colini-Baldeschi et al. (2018) to the tail-conditional case. However, when a closed-form expression is not available, a numerical estimation of the Shapley value can be computationally demanding, even for a moderate number of variables. This happens because one needs to evaluate all possible values produced by all coalitions to construct the Shapley values, leading to a computational cost that is exponential in the number of variables (Castro et al., 2009).

In this work, we address the problem of allocating variance among the variables of a GAM by using the Shapley values as a measure of variable importance. We derive a closed-form expression for these Shapley values, which allows for quick identification of the most influential variables in GAMs, thus bypassing the computational burden. Moreover, we provide the confidence intervals for these Shapley values to quantify uncertainty on their estimates. In general, our work falls within the relatively unexplored research domain concerning the interplay between variable importance methods and statistical significance

(Lo et al., 2015). In our numerical simulation and applications, we demonstrate that the information provided by the standard theory of statistical significance can be complemented through the use of innovative tools such as Shapley values.

Lastly, we note that while Bordt and von Luxburg (2023) studied Shapley values using a generalized additive model representation, their approach is not directly applicable to our specific problem as they use a different value function and do not provide a closed-form expression for Shapley values. Moreover, the high dimensionality of variables makes computation of Shapley values infeasible for their application, whereas in our case it is still feasible.

In Section 2, we introduce and analytically characterize the Shapley values in the context of variance allocation for GAMs. In Section 3 we derive the confidence intervals. Section 4 contains a numerical simulation showing that utilizing Shapley values enhances the understanding of GAMs by providing importance rankings of the variables. Such rankings are then compared with the statistical significance (in terms of p-values) of the variables. In Section 5, we demonstrate the practical utility of our approach by applying it to well-known datasets.

2 Shapley values and variance allocation

In game theory, the Shapley value (Shapley, 1953) is a method used to fairly attribute the value generated by a team of players to each individual player. Let $\nu(J)$ be the value generated by a coalition of players $J \subseteq N = \{1, 2, \ldots, n\}$, where $\nu(\varnothing) = 0$, and the total value of the game produced by the team is $\nu(N)$. Shapley (1953) considered the following desirable properties for an allocation method:

- *Efficiency:* $\sum_{i=1}^{n} \phi_i(\nu) = \nu(N)$. The total value generated by all players is equal to the sum of the Shapley values attributed to each player.

- *Symmetry:* If $\nu(J \cup \{i\}) = \nu(J \cup \{j\})$ for all $J \subseteq N \setminus \{i, j\}$, then $\phi_i(\nu) = \phi_j(\nu)$. Equal shares must be paid to players who make the same contributions to any coalition.

- *Dummy player:* If $\nu(J \cup \{i\}) = \nu(J)$ for all $J \subseteq N$, then $\phi_i(\nu) = 0$. A player who makes no contribution to any coalition, regardless of which coalition they join, will receive zero payment.

- *Linearity:* If ν and ν' are two games, then $\phi_i(\nu + \nu') = \phi_i(\nu) + \phi_i(\nu')$, where $\nu + \nu'$ is the game obtained by summing the two value functions. When two games are merged, each player's total share must be the sum of their shares from each individual game.

Shapley (1953) showed that the value given by

$$\phi_i(\nu) = \sum_{J \subseteq N \setminus \{i\}} \frac{(n - |J| - 1)!|J|!}{n!} \left[\nu(J \cup \{i\}) - \nu(J)\right] \tag{2}$$

is the unique attribution method satisfying these four properties. Note that the Shapley value for the i-th player is based on the marginal increase in the value $\nu(J \cup \{i\}) - \nu(J)$ when they join a coalition J, and such marginal increase is then averaged over all possible coalitions J. It is worth noting that the computation of Shapley value is computationally demanding since it requires estimating 2^n coalitional values.

In addition to Equation (2), the Shapley value can also be expressed as

$$\phi_i(\nu) = \frac{1}{n!} \sum_{\psi \in \mathcal{P}(N)} \left(\nu(P^\psi(i) \cup \{i\}) - \nu(P^\psi(i))\right) \tag{3}$$

where $P^\psi(i)$ is the set of all players who precede i in the order directed by the permutation ψ and $\mathcal{P}(N)$ is the set of all permutations of N.

We will now apply the Shapley value method to the problem of allocating variance among the variables of a GAM. Let $S_J = \sum_{i \in J} f_i(X_i)$ denote the partial sum obtained by considering only those functions indexed by J. We define the value function as the variance of this partial sum, given by

$$\nu(J) = \mathbb{V}[S_J]. \tag{4}$$

The rationale for this choice is that each set of f_i's should be attributed the amount of variance of the output it explains. Since the total value to be allocated is $\mathbb{V}[g(\mathbb{E}[Y|\boldsymbol{X}])]$, we aim to distribute this value fairly among the variables. For the particular case of a linear portfolio, this value function coincides with that of Colini-Baldeschi et al. (2018).

With this choice, we can obtain the following result.

Proposition 1. *For the value function defined in Eq.* (4) *we have*

$$\phi_i(\nu) = Cov\left[f_i(X_i), g(\mathbb{E}[Y|\boldsymbol{X}])\right]. \tag{5}$$

Proof. Using Equation (4) for $i \notin J$ we have

$$\begin{aligned}
\nu(J \cup \{i\}) - \nu(J) &= \mathbb{V}\left[\sum_{j \in J \cup \{i\}} f_j(X_j)\right] - \mathbb{V}\left[\sum_{j \in J} f_j(X_j)\right] \\
&= \sum_{j,l \in J \cup \{i\}} \mathrm{Cov}\left[f_j(X_j), f_l(X_l)\right] - \sum_{j,l \in J} \mathrm{Cov}\left[f_j(X_j), f_l(X_l)\right] \\
&= \mathbb{V}\left[f_i(X_i)\right] + 2\sum_{j \in J} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right].
\end{aligned}$$

From the representation (3) of the Shapley value it follows that

$$\begin{aligned}
\phi_i(\nu) &= \frac{1}{n!} \sum_{\psi \in \mathcal{P}(N)} \left(\mathbb{V}\left[f_i(X_i)\right] + 2\sum_{j \in J} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right] \right) \\
&= \mathbb{V}\left[f_i(X_i)\right] + \frac{2}{n!} \sum_{j \in N \setminus \{i\}} \sum_{\psi : j \in \mathcal{P}^\psi(i)} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right].
\end{aligned}$$

Let us show now that

$$\sum_{\psi : j \in \mathcal{P}^\psi(i)} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right] = \frac{n!}{2} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right].$$

In fact, we have to consider the position of i and j in a permutation ψ. Clearly, the permutations where j precedes i are as many as those where i precedes j: hence, their number is $\frac{n!}{2}$. Thus,

$$\begin{aligned}
\phi_i(\nu) &= \mathbb{V}\left[f_i(X_i)\right] + \sum_{j \in N \setminus \{i\}} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right] \\
&= \sum_{j=1}^{n} \mathrm{Cov}\left[f_i(X_i), f_j(X_j)\right] \\
&= \mathrm{Cov}\left[f_i(X_i), g(\mathbb{E}[Y|\boldsymbol{X}])\right].
\end{aligned}$$

□

In this case, the interpretation of the Shapley value is intuitive: the larger the magnitude of ϕ_i, the greater the variation in the output as the i-th variable changes. Therefore, it is possible to rank the component functions based on the magnitude of their corresponding Shapley values. The closed-form expression in Equation (5) provides a remarkable computational shortcut for scaling the dimensional burden of Shapley values, eliminating the need to estimate all 2^n coalitional values. As a result, it is particularly well-suited for GAMs with high-dimensional input variables (e.g., Wood et al. (2015, 2017); Fasiolo et al. (2020)). We remark that other value functions could be adopted. For example, Owen (2014) and Song et al. (2016) consider Shapley values with a value function $\nu'(J) = \mathbb{V}[\mathbb{E}(Y|X_J)]$. This value function allocates the variance $\mathbb{V}[Y]$ among the variables, it can be defined for any metamodel and is not limited to GAMs. However, these Shapley values are generally very hard to compute and no analytical expression for $n > 2$ is available even for linear models (Owen and Prieur, 2017). On the other hand, the Shapley values proposed in Lundberg and Lee (2017) (the SHAP method) assume a linear model with independent variables, making them less suitable for GAMs. In contrast, the Shapley values in Equation (5) remain well-defined under any variable dependence structure. These two features make our Shapley value formulation an ideal variable importance index for GAMs, as it is always well-defined and computationally fast.

Moreover, for the Shapley value in Equation (5) it holds that $0 \leq |\phi_i(\nu)| \leq \sqrt{\mathbb{V}[f_i(X_i)] \cdot \mathbb{V}[g(\mathbb{E}[Y|\boldsymbol{X}])]}$ by the Cauchy-Schwartz inequality. When $\phi_i = 0$, then

$$\mathbb{V}[g(\mathbb{E}[Y|\boldsymbol{X}])] = \sum_{j \neq i} \mathrm{Cov}(f_j(X_j), g(\mathbb{E}[Y|\boldsymbol{X}])) = \mathbb{V}[S_{N\setminus\{i\}}] + \mathrm{Cov}(S_{N\setminus\{i\}}, f_i(X_i)), \qquad (6)$$

where $S_{N\setminus\{i\}} = \sum_{j \in N\setminus\{i\}} f_j$ is the reduced GAM without the f_i component. Equation (6) asserts that by removing a component function with a null Shapley value, the variance of the quantity of interest does not change, except for possible spurious effects with the remaining component functions.

Regarding the interpretation of a negative Shapley value, as Colini-Baldeschi et al. (2018, page 925) write for a variance allocation in a linear portfolio problem, a negative Shapley value is the reward for a random variable which *contributes to hedge a risk*. In the GAM framework, a negative Shapley value similarly indicates that the function $f_i(X_i)$ is

negatively correlated with $g(\mathbb{E}[Y|\boldsymbol{X}])$. This means that $f_i(X_i)$ offsets the increase in the right-hand side of Equation (1) caused by the other terms. The larger the magnitude of this negative Shapley value, the more significant the dampening effect on the sum will be.

In general, a Shapley value might be small in absolute terms, but it is essential to consider it in relation to the magnitude of the total variance. By normalizing the Shapley value, we can consider the screening strategy adopted in the sensitivity analysis literature, where variables explaining a smaller fraction of the variance than a fixed threshold ε are deemed as unifluential (see Liu and Owen (2006) and Borgonovo and Rabitti (2023)). Hence, we consider a component function f_i to be uninfluential whenever $|\phi_i|/\mathbb{V}[g(\mathbb{E}[Y|\boldsymbol{X}])] < \varepsilon$. We remark that there is no automatic criterion to select this parameter in the sensitivity analysis literature, where this value is typically set to 1% (Borgonovo and Rabitti, 2023). Nonetheless, by the efficiency property, we can compute the exact error when excluding the uninfluential component functions, namely $\text{err}_\varepsilon \coloneqq 1 - \sum_{j\in\mathcal{A}_\varepsilon} \phi_j/\mathbb{V}[g(\mathbb{E}[Y|\boldsymbol{X}])]$, where $\mathcal{A}_\varepsilon$ is the set of the indices of all the influential component functions. As correctly pointed out by a reviewer, we can not adopt an analogous procedure when considering the p-values, because they do not have an efficiency-like property.

3 Confidence Intervals

In the previous section, Shapley values for the selected GAM were presented without uncertainty quantification. In this section, we propose a criterion for the uncertainty quantification of these estimates. We consider a measure of uncertainty based on confidence intervals. Huettner and Sunder (2012) employ the same approach as a measure of uncertainty of the estimated Shapley values used to allocate the R^2 of a linear model. The authors use these confidence intervals to identify the most important variables: if two intervals corresponding to two Shapley values do not overlap, then the variable with the higher Shapley value is considered more important.

We remark that in our setting the uncertainty quantification is performed after the GAM has been fixed, since it provides a *post-hoc* explanation approach of the fitted model.

Assume we have a dataset $(x_{i,k}, y_k), k = 1, \ldots, m,\ i = 1, \ldots, n$, where m is the sample size.

In the following, we use the notation

$$\hat{\phi}_{i,m} = \frac{1}{m}\sum_{k=1}^{m} \underbrace{\left(f_i\left(x_{i,k}\right) - \bar{f}_i\left(x_i\right)\right)\left(g\left(x_{N,k}\right) - \bar{g}\left(x_N\right)\right)}_{=:\, w_k}$$

for the estimate of $\mathrm{Cov}\left(f_i\left(X_i\right), g(\mathbb{E}[Y|\boldsymbol{X}])\right)$, where $x_{N,k}$ indicates the k-th observation of the vector $\boldsymbol{X}$, $g\left(x_{N,k}\right) \coloneqq g\left(\mathbb{E}\left[Y|\boldsymbol{X} = x_{N,k}\right]\right)$ is the GAM output value corresponding to the k-th observation, $\bar{g}\left(x_N\right) = \sum_{k=1}^{m} g(x_{N,k})/m$ is the empirical mean, $f_i\left(x_{i,k}\right)$ is the value of the i-th component function for the k-th observation and $\bar{f}_i\left(x_i\right) = \sum_{k=1}^{m} f_i(x_{i,k})/m$ is the empirical mean of the component f_i.

Proposition 2. *Consider the GAM as in Equation* (1)*. Then, the* $(1-\alpha)$*-level confidence interval for the* i*-th Shapley value is given by*

$$\left(\hat{\phi}_{i,m} - z_{\alpha/2}\hat{\sigma}_{i,m},\, \hat{\phi}_{i,m} + z_{\alpha/2}\hat{\sigma}_{i,m}\right), \tag{7}$$

where $\hat{\sigma}_{i,m} \coloneqq \sqrt{\left(\sum_{k=1}^{m} w_k^2/m - \hat{\phi}_{i,m}^2\right)/m}$*, and* $z_{\alpha/2}$ *is the value from the standard normal distribution such that the upper tail probability is* $\alpha/2$*.*

Proof. An estimate of $\mathrm{Cov}\left(f_i\left(X_i\right), g(\mathbb{E}[Y|\boldsymbol{X}])\right)$ is given by

$$\hat{\phi}_{i,m} = \frac{1}{m}\sum_{k=1}^{m} \left(f_i\left(x_{i,k}\right) - \bar{f}_i\left(x_i\right)\right)\left(g\left(x_{N,k}\right) - \bar{g}\left(x_N\right)\right).$$

By the Law of Large Numbers, we have that $\hat{\phi}_{i,m} \to \phi_i$. To construct the confidence interval, we only need to compute the asymptotic variance, as $m \to \infty$. Recalling that $w_k \coloneqq \left(f_i\left(x_{ik}\right) - \bar{f}_i\left(x_i\right)\right)\left(g\left(x_{N,k}\right) - \bar{g}\left(x_N\right)\right)$, it holds that

$$\begin{aligned}
\mathbb{V}\left(\frac{1}{m}\sum_{k=1}^{m} w_k\right) &= \frac{1}{m^2}\left(\mathbb{E}\left[\left(\sum_{k=1}^{m} w_k\right)^2\right] - \mathbb{E}^2\left[\sum_{k=1}^{m} w_k\right]\right) = \\
&= \frac{1}{m^2}\left(\mathbb{E}\left[\sum_{k=1}^{m} w_k^2\right] + 2\sum_{k=1}^{m}\sum_{h=1}^{k-1}\mathbb{E}\left[w_k\right]\mathbb{E}\left[w_h\right] - \mathbb{E}^2\left[\sum_{i=1}^{m} w_k\right]\right) \\
&= \frac{1}{m^2}\left(\mathbb{E}\left[\sum_{k=1}^{m} w_k^2\right] + m(m-1)\phi_i^2 - m^2\phi_i^2\right),
\end{aligned}$$

where we used the fact that

$$
\begin{aligned}
&2\sum_{k=1}^{m}\sum_{h=1}^{k-1}\mathbb{E}\left[w_h\right]\mathbb{E}\left[w_k\right] = 2\sum_{k=1}^{m}\mathbb{E}\left[w_k\sum_{k=1}^{k-1}\mathbb{E}\left[w_h\right]\right] = \\
&= 2\sum_{k=1}^{m}(k-1)\mathbb{E}\left[w_k\right]\mathbb{E}\left[w_1\right] = 2\phi_i^2\sum_{k=1}^{m}(k-1) = \\
&= 2\phi_i^2\left(\sum_{k=1}^{m}k - m\right) = 2\phi_i^2\left(\frac{m(m+1)}{2} - m\right) = \\
&= \phi_i^2\left(m^2 + m - 2m\right) = \phi_i^2 m(m-1).
\end{aligned}
$$

□

This result provides us with a criterion for determining the most important component functions, taking into account the uncertainty in the estimation of Shapley values. Indeed, a component function is more important than another if the corresponding Shapley value is higher in absolute value, and their confidence intervals do not intersect, as done in Huettner and Sunder (2012) in the case of linear regression. This constitutes another advantage with respect to the use of the p-values to identify the most influential variables. However, in contrast to Huettner and Sunder (2012), who constructed the confidence interval using a bootstrap procedure, we have derived an analytical expression.

In the next two sections, we will illustrate some of the meaningful insights that can be derived from our Shapley values for GAMs.

4 Numerical Simulation

In this section we consider the model considered in Meier et al. (2009), which is given by

$$
Y = f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + \zeta, \tag{8}
$$

where $\zeta \sim N(0,1)$, $f_1(x) = -\sin(2x)$, $f_2(x) = x^2 - \frac{25}{12}$, $f_3(x) = x$ and $f_4(x) = e^{-x} - \frac{2}{5}\sinh\left(\frac{5}{2}\right)$. We consider a sample of $m = 1500$ observations from the random vector $\boldsymbol{X}$ of dimension $n = 150$, where all the other functions $f_5(x), \ldots, f_{150}(x)$ are identically zero to induce sparsity. The marginal distribution of the X_i, $i = 1, \ldots, n$, is a Uniform$(-2.5, 2.5)$. We

consider the four dependence settings among the covariates proposed in Meier et al. (2009), specified by the correlation matrices $\Sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \ldots, n$, with $\rho = 0, 0.1, 0.3, 0.5$. The signal-to-noise ratios are approximately 15.05, 14.34, 13.08 and 12.17, respectively. The R function to simulate from the model (8) is available from the `spaddinf.R` package. Here and in the three applications in the following section, we estimate the GAMs using the `mgcv` R-package, then we compute the Shapley values using Equation (5) and their confidence intervals as in Equation (7). Results are shown in Figure 1.

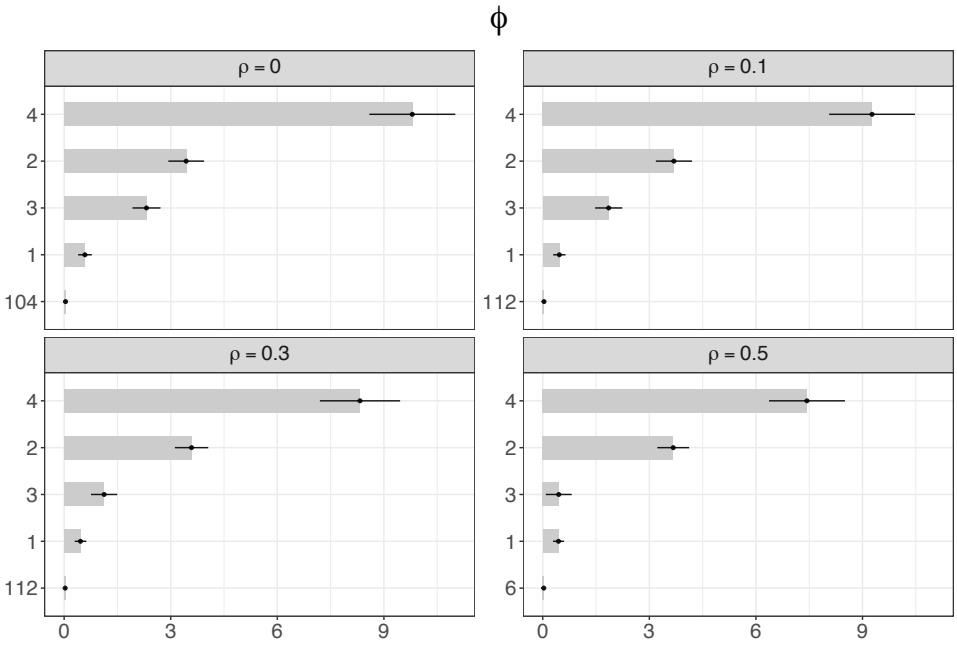

Figure 1: Shapley values for the fitted GAMs applied to the simulated data in the four dependence scenarios. Every panel displays the five highest Shapley values with the corresponding confidence interval at the 95%-level. Their computational time is around 5.65 seconds on a standard laptop, for each dependence case. For the sake of simplicity, the values on the y-axis correspond to the indices of the component functions.

From Figure 1 we note that in all dependence scenarios, the highest Shapley values correspond to the four active variables, even accounting for their uncertainty. In particular, it is evident that the fifth component function selected for importance is already negligible. Additionally, it is possible to rank the most important variables. In parallel, we record the p-

	Shapley values		p-values		
	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$***$	$**$	$*$
$\rho = 0$	$1, 2, 3, 4$	$2, 3, 4$	$1, 2, 3,$ $4, 41, 106$	$16, 104, 127$	$27, 28, 81, 82, 83, 87, 94, 97,$ $103, 105, 122, 124, 128, 130, 143$
$\rho = 0.1$	$1, 2, 3, 4$	$2, 3, 4$	$1, 2, 3, 4, 29$	$57, 83, 111,$ $112, 139, 143$	$6, 14, 20, 34, 67,$ $84, 89, 115, 123, 124, 127$
$\rho = 0.3$	$1, 2, 3, 4$	$2, 3, 4$	$1, 2, 3, 4$	14	$11, 59, 67,$ $95, 112, 124, 139, 143$
$\rho = 0.5$	$1, 2, 3, 4$	$2, 4$	$1, 2, 3, 4$	$14, 144$	$11, 27, 46, 60, 67, 83,$ $95, 117, 124, 139, 140$

Table 1: Component functions selected by the Shapley values and the p-values in the fitted GAMs for different correlation matrices depending on ρ. For any ρ, the table displays in the left columns the Shapley values selected with different thresholds, and in the right columns the indices of the component functions whose p-values are below 0.001 (***), between 0.001 and 0.01 (**) and between 0.01 and 0.05 (*).

values and report the indices of statistically significant components in Table 1. The p-values in Table 1 do identify the four active components, but, at the same time, many inactive component functions in the GAM are considered (even strongly) statistically significant. Moreover, p-values do not provide insights into the importance ranking. On the other hand, considering the Shapley values, we illustrate how the identification of active components varies for two values of ε (the standard value of 1% and a value of 5%, aiming for a more stringent selection). We observe that for the case $\varepsilon = 1\%$, all and only the active component functions are identified. The choice of $\varepsilon = 5\%$ does not select all the active component functions, as it excludes those with low covariance. This is evidenced by the Q-Q plots in Figure 2. These Q-Q plots compare models with all 150 variables (complete models) to reduced models constructed using the selected component functions, both through Shapley values and significant p-values (see Table 1). As seen in Figure 2, the fit remains completely unchanged when reducing the models based on approaches using p-values and Shapley values with $\varepsilon = 1\%$. We emphasize that the procedure using Shapley values selects only the four non-zero component functions, while p-values always include these four component functions and additionally select many others, which, by construction, are actually inactive.

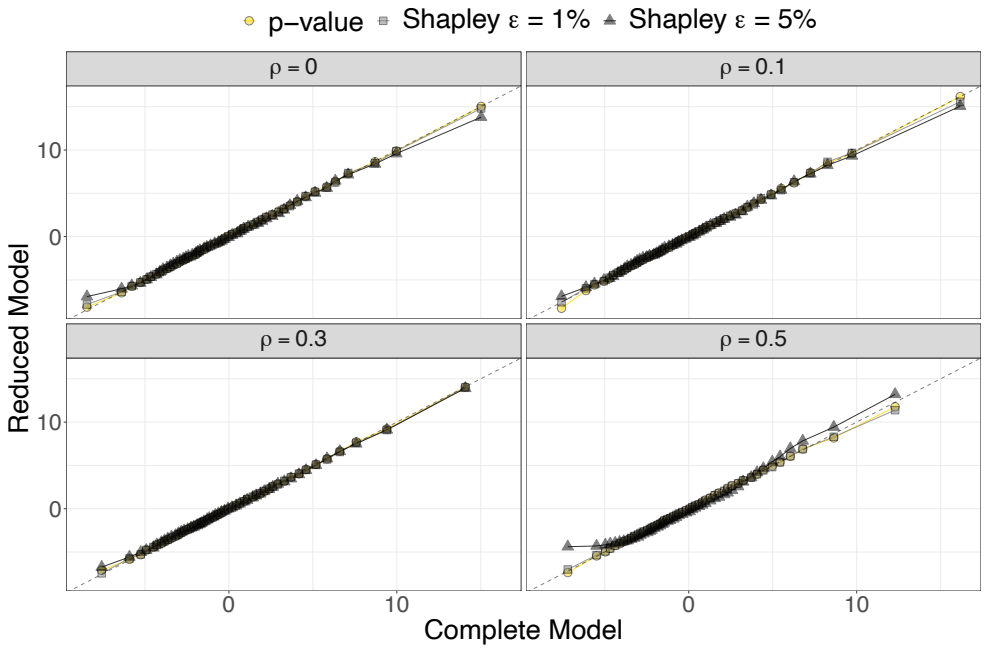

Figure 2: Q-Q plots comparing predictions obtained using the complete model with those obtained by the reduced models constructed based on Shapley values and p-values, in the four considered dependence cases denoted by ρ.

Note that the threshold $\varepsilon = 5\%$ eliminates one active component function in the dependence cases $\rho = 0$, 0.1 and 0.3, while it removes two in the dependence case $\rho = 0.5$. This is evidenced by the poorer performance of the corresponding reduced models. In this latter dependence case, the errors are $\text{err}_{1\%} = 0.023$ and $\text{err}_{5\%} = 0.096$, which are coherent with our previous discussion.

Lastly, we note that including variables with a significant p-value but a small Shapley value has a negligible effect on the prediction.

5 Applications

Our objective is to explore whether the Shapley values can offer supplementary insights regarding ranking and the selection of variables based on their statistical significance using p-values. The key findings of our analysis are summarized in Table 2.

Case Study	Response type	Insights of the analysis
Diabetes in Pima Indian Women	Boolean	Ranking based on the Shapley values reflects the level of statistical significance.
Concrete Compressive Strength	Continuous	Statistical significance does not imply high Shapley values.
Elevator	Continuous	No statistical significance does not imply small Shapley values.

Table 2: Findings from the case studies.

5.1 Diabetes in Pima Indian Women dataset

This dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases and focuses on diagnosing diabetes in female patients of Pima Indian heritage who are at least 21 years old. The dataset, available in the R-library `MASS`, is composed by 532 observations and includes specific diagnostic measurements: number of pregnancies (`npreg`), plasma glucose concentration (`glu`), diastolic blood pressure (`bp`, in mm Hg), triceps skin fold thickness (`skin`, in mm), body mass index (`bmi`), diabetes pedigree function as an index of familiarity (`ped`), `age` and `type`, indicating the presence/absence of diabetes. A previous analysis of this dataset using logistic GAM was conducted by Marra and Wood (2012). The Shapley values are illustrated in Figure 3, from which we can note several aspects.

The analysis reveals a notable agreement between the statistical significance measured by p-values and the variable importance measured by the Shapley values.
First and foremost, the variables of glucose and age exhibit the highest level of statistical significance and importance. They are followed by BMI and the pedigree function. Remarkably, by utilizing the Shapley values, it becomes possible to rank variables with similar statistical significance. Specifically, glucose emerges as the most influential variable in determining the presence of diabetes, followed by age and BMI. Interestingly, the importance

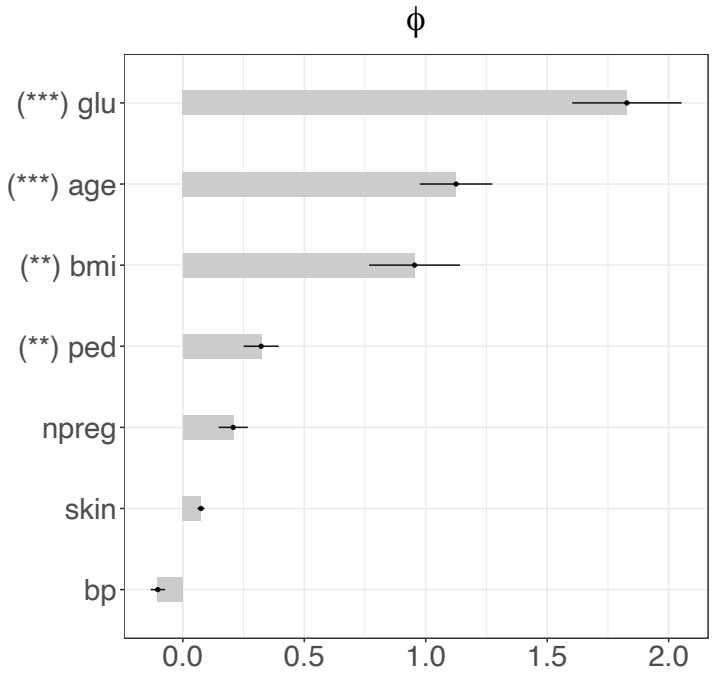

Figure 3: Shapley values with corresponding confidence intervals at the 95%-level for the logistic GAM applied to the Pima Indian women dataset for diabetes prediction. Each variable is accompanied by its corresponding p-value significance displayed to the left of its name.

of age and BMI shows a relatively minor discrepancy, confirmed by the overlapping of their confidence intervals. Hence, taking into account the uncertainty of the estimates of their Shapley values, we can not safely conclude which of the two variables is more important. Additionally, the variance allocated to the pedigree function is comparable to that of the number of pregnancies, despite only the former being statistically significant.
Finally, the variables of triceps skin thickness and blood pressure demonstrate the lowest level of significance, which aligns with the findings reported in Marra and Wood (2012).

5.2 Concrete Compressive Strength dataset

This dataset, available from the UCI repository, is used in Yeh (1998) to predict the compressive strength of high-performance concrete with respect to its age and to some

ingredients. The dataset is composed by 1030 observations and 9 quantitative variables, namely cement, blast furnace slag (`BFS`), fly ash (`FA`), water, superplasticizer, coarse and fine aggregates (all of the above measured in kg per m^3 of mixture), `age` (in days) and concrete compressive strength (in MPa). We fitted a generalized additive model and we estimated the corresponding Shapley values, which are depicted in Figure 4.

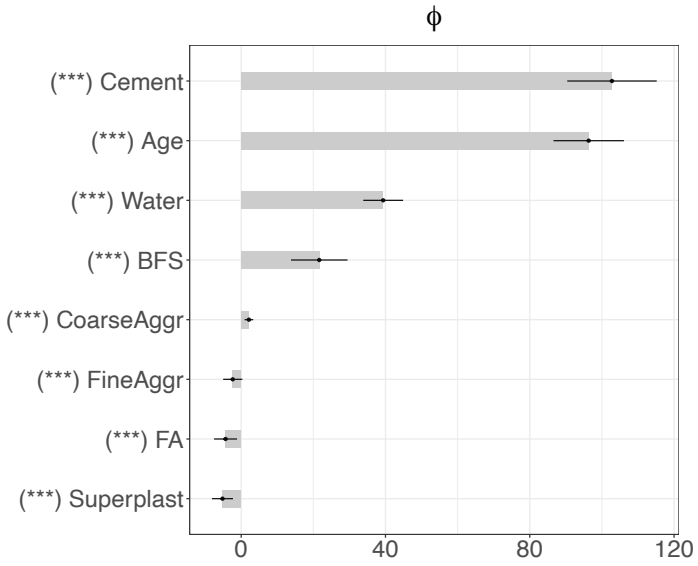

Figure 4: Shapley values with corresponding confidence intervals at the 95%-level for the GAM applied to the Cement compressive strength dataset. Each variable is displayed with its corresponding p-value significance to the left of its name.

Figure 4 shows that all variables are highly statistically significant according to their p-values. However, the Shapley values are very different. Out of the eight variables, only four are truly important, while the other four have a negligible impact on the variability of the quantity of interest. This aspect highlights that the set of highly significant variables might contain variables that are not important. To further investigate this point, we compare the predicted compressive strength obtained from both the complete GAM (including all f_i's) and a series of reduced GAMs. These reduced models are constructed by systematically removing the least important variables. In this case, we exclude the least important variables from the complete model in a sequential way:

- M_1: `CoarseAggr` removed from the complete model;

- M_2: `CoarseAggr` and `FineAggr` removed from the complete model;
- M_3: `CoarseAggr`, `FineAggr` and `FA` removed from the complete model;
- M_4: `CoarseAggr`, `FineAggr`, `FA` and `Superplast` removed from the complete model.

Models M_2 and M_4 correspond to the selections with $\varepsilon = 1\%$ ($\text{err}_{1\%} = -0.0008$) and $\varepsilon = 5\%$ ($\text{err}_{5\%} = 0.039$), respectively. We did not proceed any further since the remaining four variables exhibit non-negligible Shapley values. The results are presented in Figure 5. Figure 5 comprises three panels. The top panel presents a Q-Q plot that compares the

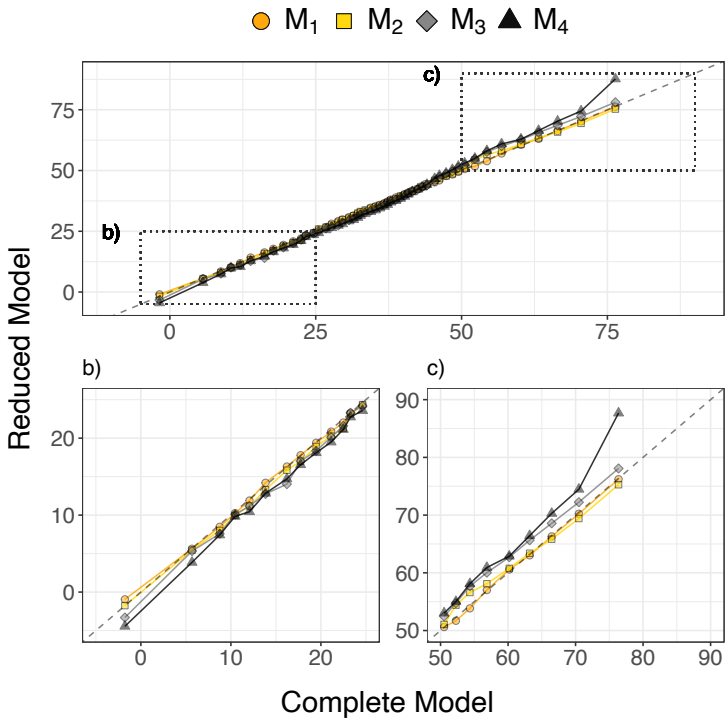

Figure 5: Q-Q plot used to compare predictions between the complete model (including all variables) and reduced models M_1, M_2, M_3, M_4.

predictions of the complete model with those of the reduced models. The subsequent panels below provide an enlarged view of the Q-Q plot at the lower and higher extremes.

The results depicted in Figure 5 indicate that eliminating the four least important variables, all of which demonstrate high statistical significance (with p-values below 10^{-3}), has a minimal impact on the prediction performance of the complete GAM model. Notably,

both M_1 and M_2 exhibit nearly identical performance to the complete model, indicating that they are unaffected by the exclusion of `CoarseAggr` and `FineAggr`, whose Shapley values are approximately zero. Additionally, the reduced model M_4, which comprises only the four most important variables, exhibits the poorest performance among M_1, M_2, and M_3. These observations are consistent with the dummy axiom and with the fact that as long as we sequentially exclude variables with non-negligible increasing importance, the accuracy of the selected models decreases.

5.3 Elevator Dataset

This dataset is available from a data repository at the University of Porto (Torgo, 2014) and has been analyzed with a linear regression model by Bertsimas and King (2016). This dataset is related to an action to control the elevators of an F16 aircraft. There are 16599 observations of the target action (`Goal`) as function of 12 continuous variables. The Shapley values and the p-value significances for these variables are illustrated in Figure 6.

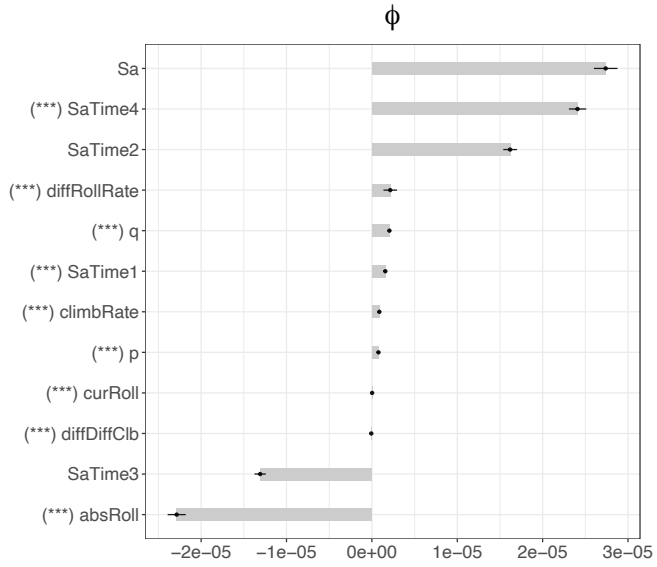

Figure 6: Shapley values with corresponding confidence intervals at the 95%-level for the GAM applied to the Elevator dataset. Each variable is displayed with its corresponding p-value significance to the left of its name.

Figure 6 shows that the most important variable (`Sa`) is not statistically significant. This also holds for `SaTime2` and `SaTime4`. In contrast, variables such as `curRoll` and `diffDiffClb`, among others, demonstrate statistical significance despite their nearly negligible Shapley values (both variables are excluded by the thresholds $\varepsilon = 1\%$ and $\varepsilon = 5\%$ - this latter case excludes also `SaTime1`, `climbRate` and `p`). In this case study, the errors are $\text{err}_{1\%} = -0.001$ and $\text{err}_{5\%} = 0.082$. In addition, we observe that the confidence intervals are much narrower compared to the previous case studies. This is not unexpected given the larger number of observations.

To acquire further insights, we examine the Q-Q plot, which compares the predictions generated by the complete model against those generated by the following reduced models:

- R_1: `Sa` (with the highest Shapley value but non-statistically significant) removed from the complete model;
- R_2: `SaTime2` (with a high positive Shapley value but non-statistically significant) removed from the complete model;
- R_3: `SaTime3` (with a highly negative Shapley value but non-statistically significant) removed from the complete model;
- R_4: `curRoll` (with an almost zero Shapley value but statistically significant) removed from the complete model.

Results are shown in Figure 7. The Q-Q plot provides evidence that removing variables with high Shapley values significantly impacts the behavior of the models compared to the complete model. Specifically, when removing a variable (`SaTime3`) with a highly negative Shapley value, we obtain model R_3, whose quantiles tend to be higher than those of the complete model. This result is intuitive because the excluded variable was counterbalancing the GAM, leading to higher predicted values by model R_3. Similarly, this phenomenon occurs for models R_1 and R_2, where the omitted variables possess highly positive Shapley values. It is worth stressing that the removed variables mentioned above are not statistically significant. Furthermore, we consider the variable `curRoll`, which is statistically significant despite having an almost zero Shapley value. We excluded it from the construction of model

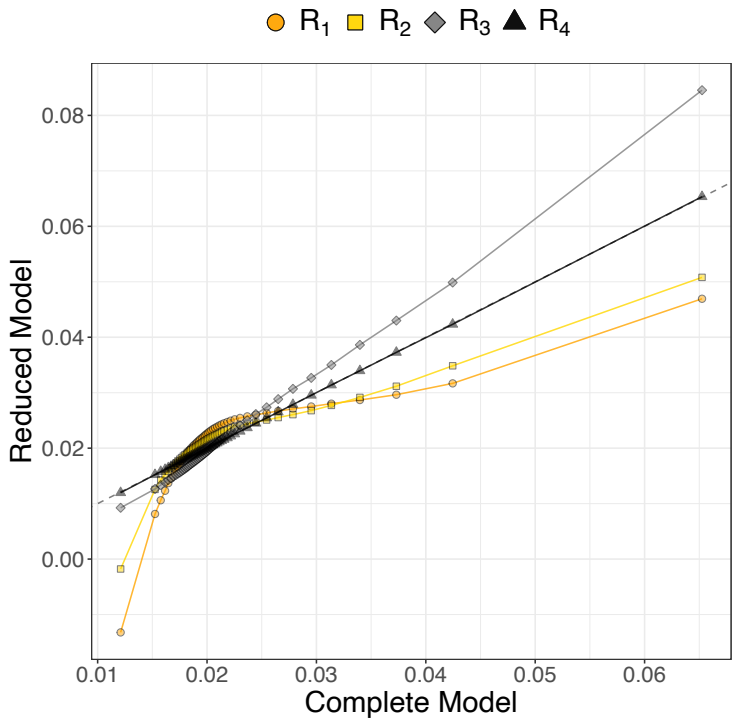

Figure 7: Q-Q plot used to compare predictions between the complete model (including all variables) and reduced models R_1, R_2, R_3, R_4.

R_4. Interestingly, this model demonstrates no difference in predicted values compared to the complete model. Consequently, in this particular application, the most important variable lacks statistical significance, while a statistically significant variable has no impact on the model's behavior.

6 Conclusions

In this work, we addressed the issue of identifying the most significant variables in a generalized additive model. To tackle this problem, we have employed a tool derived from game theory, specifically the Shapley value method, which enables us to quantify the variable importance, also considering its uncertainty. By obtaining an analytical expression of the Shapley value, we derived a global sensitivity index specifically designed for GAMs. This approach offers a computational advantage over the Shapley value formulation proposed by Owen and Prieur (2017), which does not admit a closed form expression for GAMs and

hence its computational cost grows exponentially with n. This makes our proposed Shapley values suitable for high-dimensional GAMs compared to alternative approaches based on other Shapley values in the literature.

In our study, we have found that relying solely on p-values does not ensure that variables considered statistically significant in a GAM will have a relevant impact on the response variable. Conversely, it is also possible for important variables to lack statistical significance. This is in line with what has been highlighted by Lo et al. (2015). Both our theoretical and empirical findings support the exclusion of an irrelevant variable with a null Shapley value, which acts as a dummy player.

We believe that further research is necessary to delve into the divergence between variable importance measures, such as the Shapley values in our case, and the statistical significance determined by traditional statistical tests. Moreover, an open area of research is finding closed forms and computationally-efficient approximations for Shapley values when interactions are present in the selected GAMs.

7 Supplementary Materials

Readme file: Text file containing information about the files (R-scripts and data) included in the supplementary material (`Readme.rtf`);

R-script for simulation and examples: R-script to reproduce the analyses of Section 4 and Section 5 (`Shapley_values_for_GAM.R`);

R-script containing custom R-functions: a function to perform the simulation of Section 4 (from the `spaddinf.R` package), a function to compute the margin of error (i.e. half the confidence interval) to construct the confidence intervals for the Shapley values and a function to create the bar plots of the Shapley values (`Shapley_values_for_GAM_R_functions.R`);

Data for Section 4: R-workspace (`240109_GAM.RData`) containing

- an R-list of the simulated values in all the considered dependence cases;

- an R-list containing the four GAMs estimated for each dependence case;
- an R-list of the predictions of the `s()` functions for each dependence case;
- a matrix containing the Shapley values. Each column refers to one dependence case.

Data for Subsection 5.2: An Excel table containing the data to reproduce the application in Subsection 5.2 (`Concrete_Data.xls`);

Data for Subsection 5.3: Two files containing the data to reproduce the application in Subsection 5.3. The final dataset used in our work merges the two datasets to obtain a unique dataset (`elevators.data` and `elevators.test`).

Acknowledgements

Dr. Amir Khorrami Chokami acknowledges support of MUR - Prin 2022 - Grant no. 2022CLTYP4, funded by the European Union – Next Generation EU.

Competing interests

The authors report there are no competing interests to declare.

References

Bertsimas, D. and A. King (2016). OR Forum - An Algorithmic Approach to Linear Regression. *Operations Research 64*(1), 2–16.

Bordt, S. and U. von Luxburg (2023). From Shapley Values to Generalized Additive Models and back. In F. Ruiz, J. Dy, and J.-W. van de Meent (Eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Volume 206 of *Proceedings of Machine Learning Research*, pp. 709–745. PMLR.

Borgonovo, E. and E. Plischke (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research 248*(3), 869–887.

Borgonovo, E. and G. Rabitti (2023). Screening: From tornado diagrams to effective dimensions. *European Journal of Operational Research 304*(3), 1200–1211.

Castro, J., D. Gómez, and J. Tejada (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research 36*(5), 1726–1730.

Colini-Baldeschi, R., M. Scarsini, and S. Vaccari (2018). Variance Allocation and Shapley Value. *Methodology and Computing in Applied Probability 20*(3), 919–933.

Fasiolo, M., R. Nedellec, Y. Goude, and S. N. Wood (2020). Scalable Visualization Methods for Modern Generalized Additive Models. *Journal of Computational and Graphical Statistics 29*(1), 78–86.

Galeotti, M. and G. Rabitti (2024). Tail variance allocation, Shapley value and the majorization problem. *Journal of Applied Probability 61*(1).

Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician 61*(2), 139–147.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*, Volume 1. Boca Raton: Chapman and Hall/CRC.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: data mining, inference and prediction* (2 ed.). Springer.

Huettner, F. and M. Sunder (2012). Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. *Electronic Journal of Statistics 6*, 1239–1250.

Liu, R. and A. B. Owen (2006). Estimating mean dimensionality of analysis of variance decompositions. *Journal of the American Statistical Association 101*(474), 712–721.

Lo, A., H. Chernoff, T. Zheng, and S.-H. Lo (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences 112*(45), 13892–13897.

Lundberg, S. M. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc.

Marra, G. and S. N. Wood (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis 55*(7), 2372–2387.

Marra, G. and S. N. Wood (2012). Coverage Properties of Confidence Intervals for Generalized Additive Model Components. *Scandinavian Journal of Statistics 39*(1), 53–74.

Meier, L., S. van de Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *The Annals of Statistics 37*(6B), 3779 – 3821.

Owen, A. B. (2014). Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification 2*(1), 245–251.

Owen, A. B. and C. Prieur (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification 5*(1), 986–1002.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, M. Gatelli, D. Saisana, and S. Tarantola (2008). *Global Sensitivity Analysis. The Primer*. John Wiley & Sons.

Seiler, B. B., M. Mase, and A. B. Owen (2023). What makes you unique? *Electronic Journal of Statistics 17*, 1–18.

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

Song, E., B. L. Nelson, and J. Staum (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification 4*(1), 1060–1083.

Torgo, L. (2014). Regression data sets. http:// www.dcc.fc.up.pt/ ltorgo/Regression/DataSets.html. Accessed August, 20, 2014.

Wagner, H. M. (1995). Global Sensitivity Analysis. *Operations Research 43*(6), 948–969.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall/CRC Press.

Wood, S. N., Y. Goude, and S. Shaw (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 64*(1), 139–155.

Wood, S. N., Z. Li, G. Shaddick, and N. H. Augustin (2017). Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data. *Journal of the American Statistical Association 112*(519), 1199–1210.

Xie, S. and R. Luo (2022). Measuring Variable Importance in Generalized Linear Models for Modeling Size of Loss Distributions. *Mathematics 10*(10).

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research 28*(12), 1797–1808.