On Shapley value for measuring importance of dependent inputs

Art Owen, Clémentine Prieur

On Shapley value for measuring importance of dependent inputs

Art B. Owen
Stanford University

Clémentine Prieur
Université Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France
Inria project/team AIRSEA

Orig: October 2016

This: March 2017

Abstract

This paper makes the case for using Shapley value to quantify the importance of random input variables to a function. Alternatives based on the ANOVA decomposition can run into conceptual and computational problems when the input variables are dependent. Our main goal here is to show that Shapley value removes the conceptual problems. We do this with some simple examples where Shapley value leads to intuitively reasonable nearly closed form answers.

1 Introduction

The importance of inputs to a function is commonly measured via Sobol' indices. Those are defined in terms of the functional analysis of variance (ANOVA) decomposition, which is conventionally defined with respect to statistically independent inputs. In applications to computer experiments, it is common that the input space is constrained to a non-rectangular region, or that the input variables have some other known form of dependence, such as a general Gaussian distribution. When the inputs are described by an empirical distribution on observational data it is extremely rare that the variables are statistically independent. Even designed experiments avoid having independent inputs (i.e., a Cartesian product of input levels) when the dimension is moderately large (Wu and Hamada, 2011).

A common way to address dependence is to build on work by Stone (1994) and Hooker (2012) who define an ANOVA for dependent inputs and then define variable importance through that generalization of ANOVA. This is the method taken by Chastaing et al. (2012) for computer experiments.

The dependent-variable ANOVA leads to importance measures with two conceptual problems:

1) the needed ANOVA is only defined when the random $\boldsymbol{x}$ has a distribution with a density (or mass function) uniformly bounded below by a positive constant times another density/mass function that has independent margins, and
2) the resulting importance of a variable can be negative (Chastaing et al., 2015).

The first condition is very problematic. It fails even for Gaussian $\boldsymbol{x}$ with nonzero correlation. It fails for inputs constrained to a simplex. It fails when the empirical distribution of say (x_{i1}, x_{i2}) is such that some input combinations are never observed or, by definition, cannot possibly be observed.

The second condition is also conceptually problematic. A variable on which the function does not depend at all will get importance zero and thus be more important than one that the function truly does depend on in a way that gave it negative importance.

The Shapley value, from economics, provides an alternative way to define variable importance. As we describe below, Shapley value provides a way to attribute the value created by a team to its individual members. In our context the members are individual input variables. Owen (2014) derived Shapley value importance for independent inputs where the value is variance explained. The Shapley value of a variable turns out to be bracketed between two different Sobol' indices. Song et al. (2016) recently advocated the use of Shapley value for the case of dependent inputs. They report that it is more suitable than Sobol' indices for such problems. They use the term "Shapley effects" to describe variance based Shapley values.

The Shapley value provides an importance measure that avoids the two problems mentioned above: It is available for any function in L^2 of the appropriate domain and it never gives negative importance.

Although Shapley value solves the conceptual problems, computational problems remain a serious challenge (Castro et al., 2009). The Shapley value is defined in terms of $2^d - 1$ models where d is the dimension of $\boldsymbol{x}$. Song et al. (2016) presented a Monte Carlo algorithm to estimate Shapley importance and they apply it to detailed real-world problems. We address only the conceptual appropriateness of Shapley value to variable importance, not computational issues.

The outline of this paper is as follows. Section 2 introduces our notation, defines the functional ANOVA and the Sobol' indices and presents the dependent-variable ANOVA. Section 3 presents the Shapley value and its use for variable importance. From the definition there it is clear that Shapley value for variance explained will never be negative. Section 4 gives several examples of simple cases and exceptional corner cases where we can derive the Shapley value of variable importance and verify that it is reasonable. Section 5 has brief conclusions. Section 6 contains the longer proofs.

2 Notation

We consider real valued functions f defined on a space $\mathcal{X}$. The point $\boldsymbol{x} \in \mathcal{X}$ has d components, and we write $\boldsymbol{x} = (x_1, \ldots, x_d)$ where $x_j \in \mathcal{X}_j$. The individual $\mathcal{X}_j$ are ordinarily interval subsets of $\mathbb{R}$ but each of them may be much more general (regions in Euclidean space, functions on $[0,1]$, or even images, sounds, and video). What we must assume is that $\boldsymbol{x}$ follows a distribution P chosen by the user, and that $f(\boldsymbol{x})$ is then a random variable with $\mathbb{E}(f(\boldsymbol{x})^2) < \infty$.

When the components of $\boldsymbol{x}$ are independent, then Sobol' indices (Sobol', 1990, 1993) provide ways to measure the importance of individual components of $\boldsymbol{x}$ as well as sets of them. They are based on a functional ANOVA decomposition. For details and references on the functional ANOVA, see Owen (2013).

2.1 ANOVA for independent variables

Here is a brief summary of the ANOVA to introduce our notation. For simplicity we will take $f \in L^2[0,1]^d$ with the argument $\boldsymbol{x} = (x_1, \ldots, x_d)$ of f uniformly distributed on $[0,1]^d$, but the approach extends straightforwardly to $L^2(\prod_{j=1}^d \mathcal{X}_j)$ with independent not necessarily uniform $x_j \in \mathcal{X}_j$.

The set $\{1, 2, \ldots, d\}$ is written $1{:}d$. For $u \subseteq 1{:}d$, $|u|$ denotes cardinality and $-u$ is the complement $\{1 \leqslant j \leqslant d \,|\, j \notin u\}$. If $u = (j_1, j_2, \ldots, j_{|u|})$ then $\boldsymbol{x}_u = (x_{j_1}, x_{j_2}, \ldots, x_{j_{|u|}}) \in [0,1]^{|u|}$ and $\mathrm{d}\boldsymbol{x}_u = \prod_{j \in u} \mathrm{d}x_j$. We use $u + v$ as a shortcut for $u \cup v$ when $u \cap v = \varnothing$, especially in subscripts.

The ANOVA is defined via functions $f_u \in L^2[0,1]^d$. These functions satisfy $f(\boldsymbol{x}) = \sum_{u \subseteq 1:d} f_u(\boldsymbol{x})$. They are defined as follows. First, $f_\varnothing = \int f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$ and then

$$f_u(\boldsymbol{x}) = \int \bigl(f(\boldsymbol{x}) - \sum_{v \subsetneq u} f_v(\boldsymbol{x})\bigr)\, \mathrm{d}\boldsymbol{x}_{-u} \tag{1}$$

for $|u| > 0$. The integral in (1) is over $[0,1]^{d-|u|}$ and it yields a function f_u that depends on $\boldsymbol{x}$ only through $\boldsymbol{x}_u$. The effects f_u are orthogonal: $\int f_u(\boldsymbol{x}) f_v(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0$ when $u \neq v$.

The variance component for the set u is $\sigma_u^2 = \int f_u(\boldsymbol{x})^2\, \mathrm{d}\boldsymbol{x}$ for $|u| > 0$ and $\sigma_\varnothing^2 = 0$. The variance of f for $\boldsymbol{x} \sim \mathbf{U}[0,1]^d$ is $\sigma^2 = \sum_{u \subseteq 1:d} \sigma_u^2$.

We can define the importance of a set of variables by how much of the variance of f is explained by those variables. The best prediction of $f(\boldsymbol{x})$ given $\boldsymbol{x}_u$ is

$$f_{[u]}(\boldsymbol{x}) \equiv \mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u) = \sum_{v \subseteq u} f_v(\boldsymbol{x}).$$

This prediction explains

$$\underline{\tau}_u^2 \equiv \sum_{v \subseteq u} \sigma_v^2, \tag{2}$$

of the variance in f. This is one of Sobol's global sensitivity indices. His other index is

$$\overline{\tau}_u^2 \equiv \sum_{v \cap u \neq \varnothing} \sigma_v^2 = \sigma^2 - \underline{\tau}_{-u}^2.$$

It is more conventional to use normalized versions $\underline{\tau}_u^2/\sigma^2$ and $\overline{\tau}_u^2/\sigma^2$ but unnormalized ones are simpler for our purposes. The importance of an individual variable x_j is sometimes defined through $\underline{\tau}_{\{j\}}^2$ or $\overline{\tau}_{\{j\}}^2$. If $\underline{\tau}_{\{j\}}^2$ is large then x_j is important and if $\overline{\tau}_{\{j\}}^2$ is small then x_j is unimportant.

2.2 ANOVA for dependent variables

Now suppose that f is defined on $\mathbb{R}^d$ but the argument $\boldsymbol{x}$ does not have independent components. Instead $\boldsymbol{x}$ has distribution P. We could generalize (1) to the Stone-Hooker ANOVA

$$f_u(\boldsymbol{x}) = \int \bigl(f(\boldsymbol{x}) - \sum_{v \subsetneq u} f_v(\boldsymbol{x})\bigr)\, \mathrm{d}P(\boldsymbol{x}_{-u}) \tag{3}$$

but the result would not generally have orthogonal effects. To take a basic example, suppose that P is the $\mathcal{N}\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)\right)$ distribution for $0 < \rho < 1$ and let $f(\boldsymbol{x}) = \beta_1 x_1 + \beta_2 x_2$. Then (3) yields

$$f_\varnothing(\boldsymbol{x}) = 0, \quad f_{\{1\}}(\boldsymbol{x}) = (\beta_1 + \beta_2\rho)x_1, \quad f_{\{2\}}(\boldsymbol{x}) = (\beta_2 + \beta_1\rho)x_2$$

and $f_{\{1,2\}}(\boldsymbol{x}) = -\beta_2\rho x_1 - \beta_1\rho x_2$. These effects are not orthogonal under P and their mean squares do not sum to the variance of $f(\boldsymbol{x})$ for $\boldsymbol{x} \sim P$.

It is however possible to get a decomposition $f(\boldsymbol{x}) = \sum_{u \subseteq 1:d} f_u(\boldsymbol{x})$ with a hierarchical orthogonality property

$$\int f_u(\boldsymbol{x}) f_v(\boldsymbol{x})\, \mathrm{d}P(\boldsymbol{x}) = 0, \quad \forall v \subsetneq u. \tag{4}$$

Chastaing et al. (2012) give conditions under which a decomposition of f satisfying (4) exists and they use it to define variable importance.

They assume that the joint distribution P is absolutely continuous with respect to a product probability measure ν. That is $P(\mathrm{d}\boldsymbol{x}) = p(\boldsymbol{x}) \prod_{j \in 1:d} \nu_j(\mathrm{d}x_j)$ for a density function p. They require also that this density satisfies

$$\exists\, 0 < M \leqslant 1, \quad \forall u \subseteq 1{:}d, \quad p(\mathrm{d}\boldsymbol{x}) \geqslant M p(\mathrm{d}\boldsymbol{x}_u) p(\mathrm{d}\boldsymbol{x}_{-u}), \quad \nu - \text{a.e.} \tag{5}$$

The joint density is bounded below by a product of two marginal densities. Among other things, this criterion forbids 'holes' in the support of P. There cannot be regions $R_u \in \mathbb{R}^u$ and $R_{-u} \in \mathbb{R}^{-u}$ with $P(R_u \times R_{-u}) = 0$ while $\min(P(R_u \times \mathbb{R}^{-u}), P(\mathbb{R}^u \times R_{-u})) > 0$.

2.3 Challenges with dependent variable ANOVA

The no holes condition (5) is problematic in many applications. For example, when $\boldsymbol{x}$ is uniformly distributed on the triangle

$$\{(x_1, x_2) \in [0,1]^2 \,|\, x_1 \leqslant x_2\}$$

then (5) is violated. More generally, Gilquin et al. (2015) and Kucherenko et al. (2016) consider functions on non-rectangular regions defined by linear inequality constraints. These and similar regions arise in many engineering problems where safety or costs impose constraints on design parameters.

The simplest distribution with a hole is one with positive probability on the points

$$\{(0,0), (0,1), (1,0)\}$$

and no others. Sobol's 'pick-freeze' methods (Sobol', 1990, 1993) estimate variable importance by freezing the level of some inputs and then picking new values for the others. For the example here, setting $x_1 = 1$ implies that x_2 cannot be changed at all, which is a severe problem for a pick-freeze approach with dependent inputs.

It is not just probability zero holes that cause a problem for dependent variable ANOVA. When $\boldsymbol{x}$ is normally distributed with some nonzero correlations, then (5) does not hold, and then as we mentioned in the introduction, the dependent-variable ANOVA is unavailable. The second problem we mentioned there is that the dependent variable ANOVA can yield negative estimates of importance.

3 Shapley value

Shapley value is a way to attribute the economic output of a team to the individual members of that team. In our case, the team will be the set of variables $x_1, x_2, \ldots, x_d$. Given any subset $u \subseteq 1{:}d$ of variables, the value that subset creates on its own is its explanatory power. A convenient way to measure explanatory power is via

$$\mathrm{val}(u) = \underline{\tau}_u^2 \equiv \mathrm{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u)). \tag{6}$$

Here, the empty set creates no value and the entire team contributes σ^2 which we must now partition among the x_j.

There are four very compelling properties that an attribution method should have. The following list is based on the account in Winter (2002). Let $\mathrm{val}(u) \in \mathbb{R}$ be the value attained by the subset $u \subseteq \{1, \ldots, d\} \equiv 1{:}d$. It is always assumed that $\mathrm{val}(\varnothing) = 0$, which holds in our variance explained setting. The values $\phi_j = \phi_j(\mathrm{val})$ should satisfy these properties:

1) (Efficiency) $\sum_{j=1}^d \phi_j = \mathrm{val}(1{:}d)$.
2) (Symmetry) If $\mathrm{val}(u \cup \{i\}) = \mathrm{val}(u \cup \{j\})$ for all $u \subseteq 1{:}d - \{i, j\}$, then $\phi_i = \phi_j$.

3) (Dummy) If $\mathrm{val}(u \cup \{i\}) = \mathrm{val}(u)$ for all $u \subseteq 1{:}d$, then $\phi_i = 0$.
4) (Additivity) If val and val$'$ have Shapley values ϕ and ϕ' respectively then the game with value val + val$'$ has Shapley value $\phi_j + \phi'_j$ for $j \in 1{:}d$.

Shapley (1953) showed that the unique valuation ϕ that satisfies these axioms attributes value

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} \big(\mathrm{val}(u \cup \{j\}) - \mathrm{val}(u)\big)$$

to variable j. Defining the value via (6) we get

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (\underline{\tau}^2_{u+\{j\}} - \underline{\tau}^2_u). \tag{7}$$

From (7) we see that the Shapley value is defined for any function for which $\mathrm{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u))$ is always defined. The components $\boldsymbol{x}_j$ do not have to be real valued, though $f(\boldsymbol{x})$ must be. Holes in the domain $\mathcal{X}$ do not make it impossible to define a Shapley value. Next, because $\boldsymbol{x}_{u+\{j\}}$ always has at least as much explanatory power as $\boldsymbol{x}_u$ has, we see that $\phi_j \geqslant 0$. That is, no variable has a negative Shapley value. As a result, the Shapley value addresses the two conceptual problems mentioned in the introduction.

Song et al. (2016) show that the same Shapley value arises if we use $\mathrm{val}(u) = \mathbb{E}(\mathrm{var}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_{-u}))$. That provides an alternative way to compute Shapley value. The Shapley value simplifies for independent inputs.

Theorem 1. *Let the ANOVA decomposition of a function with d independent inputs have variance components σ^2_u for $u \subseteq 1{:}d$. If the value of a subset u of variables is $\mathrm{val}(u) = \underline{\tau}^2_u$, then the Shapley value of variable j is*

$$\phi_j = \sum_{u \subseteq 1:d,\ j \in u} \sigma^2_u / |u|.$$

Proof. Owen (2014). □

It follows from Theorem 1 that $\underline{\tau}^2_{\{j\}} \leqslant \phi_j \leqslant \overline{\tau}^2_{\{j\}}$. This is how the Sobol' indices bracket the Shapley value.

4 Special cases

Here we consider some special case distributions and toy functions where we can work out the Shapley value in a closed or nearly closed form. The point of these examples is to show that Shapley gives sensible answers in both regular cases and corner cases. Because $\sigma^2 = \mathrm{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u)) + \mathbb{E}(\mathrm{var}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u))$ we may use

$$\underline{\tau}^2_u = \sigma^2 - \mathbb{E}(\mathrm{var}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u)). \tag{8}$$

4.1 Linear functions

Let $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^d \beta_j x_j$ where x_j are independent with variances σ_j^2. It is then easy to find that $\phi_j = \beta_j^2\sigma_j^2$. If we reparameterize x_j to cx_j for $c \neq 0$ then β_j becomes β_j/c and the importance of this variable remains unchanged as it should. Dependence among the x_j complicates the expression for Shapley effects in linear settings.

Shapley value for linear functions has historically been used to partition the R^2 quantity (proportion of sample variance explained) from a regression on d variables among those d variables. Taking the value of a subset u of variables to be R_u^2, the R^2 value when regressing a response on predictors x_j for $j \in u$, yields Shapley value

$$\phi_j = \frac{1}{d}\sum_{u\subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (R^2_{u+\{j\}} - R^2_u). \tag{9}$$

This is the LMG measure of variable importance, named after the authors of Lindeman et al. (1980). If we rearrange the d variables into all $d!$ orders, find the improvement in R^2 that comes at the moment the j'th variable is added to the regression, then (9) is the average of all those improvements. The LMG reference is difficult to obtain. Genizi (1993) is another reference, having (9) as equation (1). Grömping (2007) cites several more references on partitioning R^2 in regression and discusses alternative measures and criteria for choosing. It is clear that (9) is expensive for large d.

Here we consider a population/distribution version of partitioning variance explained among a set of variables acting linearly. We suppose that $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$ where $\Sigma \in \mathbb{R}^{d\times d}$ is a positive semi-definite symmetric matrix. The function of interest is $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^{\mathsf{T}}\beta$ where $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$. If there is an error term as in a linear regression on noisy data, then we can let x_d be that error variable with a corresponding $\beta_d = 1$.

If Σ is not diagonal then the Stone-Hooker ANOVA is not available because (5) does not hold. Shapley value gives an interpretable expression for general d.

Theorem 2. *If $f(\boldsymbol{x}) = \beta_0 + \beta^{\mathsf{T}}\boldsymbol{x}$ for $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$ where $\Sigma \in \mathbb{R}^{d\times d}$ has full rank, then the Shapley effect for variable j is*

$$\phi_j = \frac{1}{d}\sum_{u\subseteq -j} \binom{d-1}{|u|}^{-1} \frac{\operatorname{cov}\bigl(x_j, \boldsymbol{x}_{-u}^{\mathsf{T}}\beta_{-u} \mid \boldsymbol{x}_u\bigr)^2}{\operatorname{var}(x_j \mid \boldsymbol{x}_u)}.$$

Proof. See Section 6.1. □

A variable with $\beta_j = 0$ can still have $\phi_j > 0$. For instance if $\Sigma = \left(\begin{smallmatrix} 0 & \rho \\ \rho & 0 \end{smallmatrix}\right)$ and $f(\boldsymbol{x}) = x_1$, then we can find directly from (7) that $\phi_2 = \rho^2/2$ and $\phi_1 = 1 - \rho^2/2$. For $\rho = \pm 1$ we already know this by bijection.

The Shapley value works with conditional variances and the Gaussian distribution makes these very convenient. For non-Gaussian distributions the conditional covariance of $\boldsymbol{x}_v$ and $\boldsymbol{x}_w$ given $\boldsymbol{x}_u$ may depend on the specific value of $\boldsymbol{x}_u$, while in the Gaussian case it is simply $\Sigma_{vw} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uw}$ for all $\boldsymbol{x}_u$.

In a related problem, if we define $\mathrm{val}(u)$ to be $\mathrm{var}(\sum_{j\in u} x_j)$, instead of $\mathrm{var}(\mathbb{E}(\sum_j x_j \,|\, \boldsymbol{x}_u))$, then the Shapley value of variable j is $\phi_j = \mathrm{cov}(x_j, S)$, where $S = \sum_{j\in 1:d} x_j$. See Colini-Baldeschi et al. (2016). This quantity can be negative. For instance, if $d = 2$, then $\phi_1 = \mathrm{var}(x_1) + \mathrm{cov}(x_1, x_2)$ which is negative when x_1 and x_2 are negatively correlated and x_2 has much greater variance than x_1.

4.2 Transformations, bijections and invariance

We can generalize the linear example to independent random variables that contribute additively: $f(\boldsymbol{x}) = \sum_{j=1}^d g_j(x_j)$. Then $\phi_j = \mathrm{var}(g_j(x_j))$. Replacing x_j by a bijection $\tau_j(x_j)$ and adjusting g_j to $g_j \circ \tau_j^{-1}$ leaves ϕ_j unchanged.

More generally, suppose that $y = f(\boldsymbol{x})$ and we transform the variables x_j into z_j by bijections: $z_j = \tau_j(x_j)$, $x_j = \tau_j^{-1}(z_j)$, for $j = 1, \ldots, d$. Now define $f'(\boldsymbol{z}) = f(\tau_1^{-1}(z_1), \ldots, \tau_d^{-1}(z_d))$ and let ϕ'_j be the Shapley importance of z_j as a predictor of $y' = f'(\boldsymbol{z})$. Because $\mathrm{var}(\mathbb{E}(f'(\boldsymbol{z}) \,|\, \boldsymbol{z}_u)) = \mathrm{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u))$, we find that $\phi'_j = \phi_j$ for $j = 1, \ldots, d$, where ϕ_j is the Shapley importance of x_j as a predictor of y. As a result we can apply invertible transformations to any or all of the x_j without changing the Shapley values.

Now lets revisit the linear setting with an extreme example: $f(x_1, x_2) = 10^6 x_1 + x_2$ with $x_1 = 10^6 x_2$ where x_2 (and hence x_1) has a finite positive variance. Because $\partial f/\partial x_1 \gg \partial f/\partial x_2 > 0$ and $\mathrm{var}(x_1) \gg \mathrm{var}(x_2)$ one might expect x_1 to be the more important variable. However, the Shapley formula easily yields $\phi_1 = \phi_2$; these variables are equally important. This is quite reasonable because f is a function of x_1 alone and equally a function of x_2 alone.

More generally, for $d \geqslant 2$, if there is a bijection between any two of the x_j then those two variables have the same Shapley value. To see this, let $x_1 = g_1(x_2)$ and $x_2 = g_2(x_1)$, both with probability one then for any $u \subset 1{:}d$ with $u \cap \{1, 2\} = \varnothing$ we have

$$\mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_{u+\{1\}}) = \mathbb{E}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_{u+\{2\}}).$$

It follows that $\underline{\tau}^2_{u+\{1\}} - \underline{\tau}^2_u = \underline{\tau}^2_{u+\{2\}} - \underline{\tau}^2_u$ and therefore $\phi_1 = \phi_2$ by the symmetry property of Shapley value.

To summarize:

1) Shapley value is preserved under invertible transformations, and
2) a bijection between variables implies that they have the same Shapley value.

4.3 Bivariate settings

When $d = 2$ we can get some simpler formulas for the importance of the two variables.

Proposition 1. *Let $f(\boldsymbol{x})$ have finite variance $\sigma^2 > 0$ for random $\boldsymbol{x} = (x_1, x_2)$. Then from (7),*

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\Big(1 + \frac{\mathrm{var}(\mathbb{E}(Y \,|\, x_1)) - \mathrm{var}(\mathbb{E}(Y \,|\, x_2)}{\sigma^2}\Big) \tag{10}$$

$$= \frac{1}{2}\Big(1 + \frac{\mathbb{E}(\mathrm{var}(Y \,|\, x_2)) - \mathbb{E}(\mathrm{var}(Y \,|\, x_1))}{\sigma^2}\Big), \quad \textit{and} \tag{11}$$

$$\frac{\phi_1}{\phi_2} = \frac{\mathrm{var}(\mathbb{E}(Y \,|\, x_1)) + \mathbb{E}(\mathrm{var}(Y \,|\, x_2))}{\mathrm{var}(\mathbb{E}(Y \,|\, x_2)) + \mathbb{E}(\mathrm{var}(Y \,|\, x_1))}. \tag{12}$$

Proof. Using $\underline{\tau}^2_{\{1,2\}} = \sigma^2$ and $\underline{\tau}^2_{\varnothing} = 0$, we find that

$$\phi_1 = \frac{1}{2}\big(\underline{\tau}^2_{\{1\}} + \sigma^2 - \underline{\tau}^2_{\{2\}}\big) = \frac{1}{2}\big(\sigma^2 + \mathrm{var}(\mathbb{E}(Y \,|\, x_1)) - \mathrm{var}(\mathbb{E}(Y \,|\, x_2)),$$

which gives us (10). The others are algebraic rearrangements. □

We can use Proposition 1 to get analogous expressions for ϕ_2/σ^2 and ϕ_2/ϕ_1 by exchanging indices.

4.3.1 Farlie-Gumbel-Morgenstern copula for $d = 2$

Here we focus on the case where the dependence between both components x_1 and x_2 is explicitly described by some copula. There exist simple conditional expectation formulas when considering some classical classes of copulas (see e.g., Crane and Hoek (2008) and references therein). Starting from such formulas, it is possible to derive explicit computations for Shapley values in a linear model. In this section, we state explicit results for the Farlie-Gumbel-Morgenstern family of copulas.

The Farlie-Gumbel-Morgenstern copula describes a random vector $\boldsymbol{x} \in [0,1]^2$ with each component $x_j \sim \mathbf{U}[0,1]$ and joint probability density function

$$c_\theta(x_1, x_2) = 1 + \theta(1 - 2x_1)(1 - 2x_2), \quad -1 \leqslant \theta \leqslant 1. \tag{13}$$

One can show that $\mathrm{cor}(x_1, x_2) = \theta/3$. Lai (1978) proved that, for $0 \leq \theta \leq 1$, x_1 and x_2 are positively quadrant dependent and positively regression dependent. Moreover,

$$\mathbb{E}(x_2 \,|\, x_1) = \frac{\theta}{3}x_1 + \Big(\frac{1}{2} - \frac{\theta}{6}\Big). \tag{14}$$

The linearity above is very useful for our purpose, as it will allow an explicit computation for Shapley values in that model.

Proposition 2. *Let $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\beta$ for $\boldsymbol{x}, \beta \in \mathbb{R}^2$ and $\boldsymbol{x} \sim c_\theta(x_1, x_2)$, with $-1 \leq \theta \leq 1$. Then*

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\left(1 + \Big(1 - \frac{\theta^2}{9}\Big)\frac{\beta_1^2 - \beta_2^2}{12\sigma^2}\right),$$

with $\sigma^2 = (\beta_1^2 + \beta_2^2)/12 + \beta_1\beta_2\theta/18$.

Proof. From the linearity of the regression function (14),

$$\mathbb{E}(f(\mathbf{x}) \,|\, x_1) = x_1\Big(\beta_1 + \frac{\theta}{3}\beta_2\Big) + \beta_2\Big(\frac{1}{2} - \frac{\theta}{6}\Big),$$

thus

$$\operatorname{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, x_1)) = \frac{1}{12}\Big(\beta_1 + \frac{\theta}{3}\beta_2\Big)^2.$$

Symmetry gets us the corresponding expression for $\operatorname{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, x_2))$. Then Proposition 1 establishes the expression for ϕ_1/σ^2. Finally, because $\operatorname{var}(x_j) = 1/12$ and $\operatorname{cor}(x_1, x_2) = \theta/3$, we get $\sigma^2 = (\beta_1^2 + \beta_2^2)/12 + \beta_1\beta_2\theta/18$. □

Now we consider the Farlie-Gumbel-Morgenstern copula, but we assume x_j has as cumulative distribution function F_j, and probability density function F_j', not necessarily from the uniform distribution.

Lemma 1. *Let $\boldsymbol{x} \in \mathbb{R}^2$ have probability density $F_1'(x_1)F_2'(x_2)c_\theta(F_1(x_1), F_2(x_2))$, with $-1 \leq \theta \leq 1$. Then*

$$\mathbb{E}(x_2 \,|\, x_1) = \mathbb{E}(x_2) + \theta(1 - 2F_1(x_1)) \int_{\mathbb{R}} y(1 - 2F_2(y))F_2'(y)\,\mathrm{d}y.$$

For exponential x_j with $F_j(x_j) = 1 - \exp(-\lambda_j x_j)$ for $\lambda_j > 0$, we get

$$\mathbb{E}(x_2 \,|\, x_1) = \frac{1}{\lambda_2} + \frac{\theta}{2\lambda_2}(1 - 2e^{-\lambda_1 x_1}). \tag{15}$$

Proof. Crane and Hoek (2008). □

Next we assume that $\boldsymbol{x}$ has exponential margins and we transform these margins to be unit exponential by making a corresponding scale adjustment to β. From Section 4.2, we know that such transformations do not change the Shapley value.

Proposition 3. *Let $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\beta$ for $\boldsymbol{x}, \beta \in \mathbb{R}^2$ where $\boldsymbol{x}$ has probability density function $e^{-x_1-x_2}c_\theta(1 - e^{-x_1}, 1 - e^{-x_2})$, where $-1 \leq \theta \leq 1$. Then*

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\Big(1 + \Big(1 - \frac{\theta^2}{12}\Big)\frac{\beta_1^2 - \beta_2^2}{\sigma^2}\Big) \tag{16}$$

with $\sigma^2 = \beta_1^2 + \beta_2^2 + \theta\beta_1\beta_2/2$.

Proof. From Lemma 1, $\mathbb{E}(x_2 \,|\, x_1) = 1 + \theta/2 - \theta e^{-x_1}$ so

$$\mathbb{E}(f(\boldsymbol{x}) \,|\, x_1) = \beta_1 x_1 + \beta_2(1 + \theta/2 - \theta e^{-x_1}).$$

Therefore

$$\mathrm{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, x_1)) = \beta_1^2 + \beta_2^2\theta^2\mathrm{var}(e^{-x_1}) - 2\beta_1\beta_2\theta\mathrm{cov}(x_1, e^{-x_1}).$$

Now $\mathrm{var}(e^{-x_1}) = \mathbb{E}(e^{-2x_1}) - \mathbb{E}(e^{-x_1})^2 = 1/12$ and

$$\mathrm{cov}(x_1, e^{-x_1}) = \int_0^\infty x e^{-2x}\,\mathrm{d}x - \frac{1}{2} = -\frac{1}{4},$$

so $\mathrm{var}(\mathbb{E}(f(\boldsymbol{x}) \,|\, x_1)) = \beta_1^2 + \beta_2^2\theta^2/12 + \beta_1\beta_2\theta/2$. This establishes (16) by Proposition 1. □

Suppose that $\beta_1 > \beta_2 > 0$. Then of course $\phi_1/\sigma^2 > 1/2$. Equation (16) shows that ϕ_1/σ^2 decreases as θ increases from 0 to 1. It does not approach 1/2 because even at $\theta = 1$, x_2 is not a deterministic function of x_1.

4.3.2 Gaussian variables, exponential f, $d = 2$

Let $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$ and take $Y = e^{\beta_0 + \sum_{j=1}^d x_j\beta_j}$. The effect of β_0 and μ_j is simply to scale Y and so we can take $\beta_0 = 0$ and $\mu = 0$ without affecting ϕ_j/σ^2. Next we suppose that the diagonal elements of Σ are nonzero. By the transformation result in Section 4.2 we can replace each x_j by x_j/Σ_{jj} if need be without changing ϕ_j and so we suppose that each $x_j \sim \mathcal{N}(0, 1)$. Here we find variable importances for $d = 2$.

Proposition 4. *Let $f(\boldsymbol{x}) = \exp\bigl(\boldsymbol{x}^\mathsf{T}\beta\bigr)$ for $\boldsymbol{x}, \beta \in \mathbb{R}^2$ and $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, for $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Then*

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\left(1 + \frac{e^{(\beta_1+\beta_2\rho)^2} - e^{(\beta_2+\beta_1\rho)^2}}{e^{\beta_1^2+\beta_2^2+2\rho\beta_1\beta_2} - 1}\right), \tag{17}$$

where the variance of $f(\boldsymbol{x})$ is

$$\sigma^2 = e^{\beta_1^2+\beta_2^2+2\rho\beta_1\beta_2}(e^{\beta_1^2+\beta_2^2+2\rho\beta_1\beta_2} - 1). \tag{18}$$

Proof. Recall the lognormal moments: if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{E}(e^Z) = e^{\mu+\sigma^2/2}$ and $\mathrm{var}(e^Z) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$. Taking $Z = \boldsymbol{x}^\mathsf{T}\beta$ we find that $Y = e^Z$ has variance σ^2 given by (18).

The distribution of $x_2\beta_2$ given x_1 is $\mathcal{N}(\rho x_1\beta_2, (1-\rho^2)\beta_2^2)$. Therefore

$$\mathbb{E}(Y \,|\, x_1) = e^{(\beta_1+\rho\beta_2)x_1 + \beta_2^2(1-\rho^2)/2}, \quad \text{and so}$$
$$\mathrm{var}(\mathbb{E}(Y \,|\, x_1)) = e^{\beta_2^2(1-\rho^2)}e^{(\beta_1+\rho\beta_2)^2}(e^{(\beta_1+\rho\beta_2)^2} - 1)$$

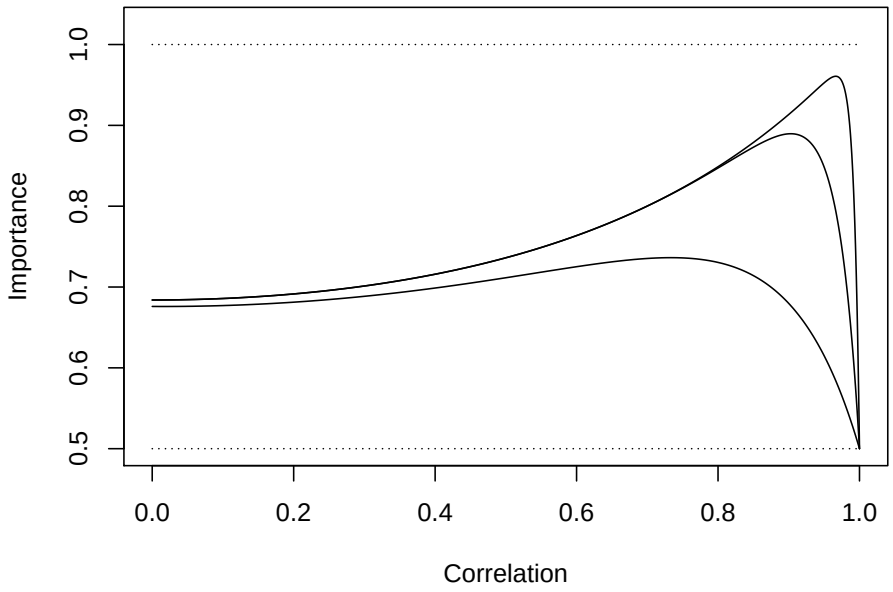

Figure 1: Relative importance ϕ_1/σ^2 versus correlation $|\rho|$ from Proposition 2. From top to bottom, β^{T} is $(8,1)$, $(4,1)$, and $(2,1)$.

$$= e^{\beta^{\mathsf{T}}\Sigma\beta}(e^{(\beta_1+\rho\beta_2)^2} - 1).$$

Similarly, $\mathrm{var}(\mathbb{E}(Y \mid x_2)) = e^{\beta^{\mathsf{T}}\Sigma\beta}(e^{(\beta_2+\rho\beta_1)^2} - 1)$. Then applying Proposition 1 and noticing that the lead factor $e^{\beta^{\mathsf{T}}\Sigma\beta}$ appears also in σ^2, yields the result. □

If $\rho = \pm 1$ then $\phi_1/\sigma^2 = 1/2$ as it must because there is then a bijection between the variables. The value of ϕ_1/σ^2 in (17) is unchanged if we replace ρ by $-\rho$. The formula is not obviously symmetric, but the fraction within parentheses there can be divided by the corresponding one for $-\rho$ and the ratio reduces to 1. More directly, we know from Section 4.2 that making the transformation $x_2 \to -x_2$ and $\beta_2 \to -\beta_2$ would leave the variable importances unchanged while switching $\rho \to -\rho$.

It is clear that for $\beta_1 > \beta_2$ we must have $\phi_1/\sigma^2 \geqslant 1/2$. Even with the closed form (17), it is not obvious how ϕ_1/σ^2 should depend on ρ or on β. Figure 1 shows that increasing $|\rho|$ from zero generally raises the importance of x_1 until at some high correlation level the relative importance quickly drops down to $1/2$. Also, for $\rho = 0$ the effect of β_1 over the range $2 \leqslant \beta_1 \leqslant 8$ is quite small when $\beta_2 = 1$.

The lognormal case is different from the bivariate normal case. There, the value of ϕ_1 converges monotonically towards $1/2$ as $|\rho|$ increases from 0 to 1.

p	x_1	x_2	y
p_0	0	0	y_0
p_1	1	0	y_1
p_2	0	1	y_2

Table 1: The random variable $y = f(\boldsymbol{x})$ is the given function of $\boldsymbol{x} = (x_1, x_2)$. That vector takes three values with the probabilities in this table. For example, $\Pr(\boldsymbol{x} = (1, 0)) = p_1$ and then $y = y_1$.

4.4 Holes

Here we consider the simplest setting where there is an unreachable part of the $\boldsymbol{x}$ space. We consider two binary variables x_1 and x_2 but $x_1 = x_2 = 1$ never occurs. For instance f could be the weight of a sea turtle, x_1 could be 1 iff the turtle is bearing eggs and x_2 could be 1 iff the turtle is male. It may seem unreasonable to even attempt to compare the importance of these variables (male/female versus eggs/none) but Shapley value does provide such a comparison based on compelling axioms in the event that we do seek a comparison.

This simplest setting is depicted in Table 1 where $p_0 + p_1 + p_2 = 1$. We assume that $p_1 > 0$ and $p_2 > 0$ for otherwise the function does not have two input variables.

Theorem 3. *Let y be a function of the random vector $\boldsymbol{x}$ as given in Table 1. Assume that $\sigma^2 = \mathrm{var}(y) > 0$, and $\min(p_1, p_2) > 0$. Then the Shapley relative importance of variable x_1 is*

$$\frac{1}{2}\Big(1 + \frac{p_0}{\sigma^2} \times \frac{p_1(1-p_1)\bar{y}_1^2 - p_2(1-p_2)\bar{y}_2^2}{(1-p_1)(1-p_2)}\Big) \tag{19}$$

where $\bar{y}_j = y_j - y_0$ for $j = 1, 2$.

Proof. See section 6.2. □

We see that when $p_0 = 0$, then the Shapley relative importance of x_1 is $1/2$. That is what it must be because there is then a bijection between x_1 and x_2 via $x_1 + x_2 = 1$.

Now suppose that $\bar{y}_1 = \bar{y}_2$. For instance $y_1 = y_2 = 1$ while $y_0 = 0$. Then the more important variable is the one with the larger variance. That is x_1 is more important if $p_1(1-p_1) > p_2(1-p_2)$. This can only happen if $p_1 > p_2$. So the more probable input is the more important one in this case.

4.5 Maximum of exponential random variables

Keinan et al. (2004) considered a network of neurons $e_1, \ldots, e_d$ where the e_j have independent lifetimes x_j that are exponentially distributed with mean $1/\lambda_j$. In their setting the value of a set of neurons is $\phi(u) = \mathbb{E}(\max_{j \in u} x_j)$, that is the

expected amount of time that at least part of that subset survives. For $d = 3$, they give a Shapley value of

$$\phi_j = \frac{1}{\lambda_1} - \frac{1}{2}\frac{1}{\lambda_1+\lambda_2} - \frac{1}{2}\frac{1}{\lambda_1+\lambda_3} + \frac{1}{3}\frac{1}{\lambda_1+\lambda_2+\lambda_3},$$

but they do not give a proof. While value in this example is not based on prediction error, we include it because it is another example of a closed form for Shapley value based on random variables. We prove their formula here and generalize it to any $d \geqslant 1$.

Theorem 4. *Let the value of a set $u \subseteq 1{:}d$ be $\mathrm{val}(u) = \mathbb{E}(\max_{j\in u} x_j)$ where $x_1, \ldots, x_d$ are independent exponential random variables with $\mathbb{E}(x_j) = 1/\lambda_j$. Then*

$$\phi_j = \sum_{r=1}^{d} \frac{(-1)^{r-1}}{r} \sum_{w\subseteq 1:d, j\in w, |w|=r} \frac{1}{\sum_{\ell\in w}\lambda_\ell}.$$

Proof. See section 6.3. □

5 Conclusions

The Shapley value from economics remedies the conceptual difficulties in measuring importance of dependent variables via ANOVA. Like ANOVA it uses variances, but unlike the dependent data ANOVA, Shapley value never goes negative and it can be defined without onerous assumptions on the input distribution.

We find that Shapley value has useful properties. When two variables are functionally equivalent, then they get equal Shapley value. When an invertible transformation is made to a variable, it retains its Shapley value. We thus conclude that Song et al. (2016) had the right idea proposing Shapley value for dependent inputs. Computation of Shapley values remains a challenge outside of special cases like the ones we discuss here.

A potential application that we find interesting is measuring the importance of parameters in a Bayesian context. When the parameter vector β has an approximate Gaussian posterior distribution, as the central limit theorem often provides, then Theorem 2 yields a measure $\phi_j(\boldsymbol{x}_0)$ for the importance of parameter β_j for the posterior uncertainty of the prediction $\boldsymbol{x}_0^{\mathsf{T}}\beta$. We hasten to add that parameter independence is quite different from variable importance, which is a more common goal. By this measure an important parameter is one whose uncertainty dominates uncertainty in $\boldsymbol{x}_0^{\mathsf{T}}\beta$. The corresponding variable may or may not be important. Another potential application is in modeling the importance of order statistics. They naturally belong to a non-rectangular set Lebrun and Dutfoy (2014).

Acknowledgments

This work was supported by grant DMS-1521145 from the U.S. National Science Foundation. We thank Marco Scarsini, Jiangming Xiang, Bertrand Iooss, two anonymous referees and an associate editor for valuable comments.

References

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

Chastaing, G., Gamboa, F., and Prieur, C. (2012). Generalized Hoeffding-Sobol' decomposition for dependent variables – applications to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448.

Chastaing, G., Gamboa, F., and Prieur, C. (2015). Generalized Sobol' sensitivity indices for dependent variables: Numerical methods. *Journal of Statistical Computation and Simulation*, 85(7):1306–1333.

Colini-Baldeschi, R., Scarsini, M., and Vaccari, S. (2016). Variance allocation and Shapley value. *Methodology and Computing in Applied Probability*, pages 1–15.

Crane, G. J. and Hoek, J. v. d. (2008). Conditional expectation formulae for copulas. *Australian & New Zealand Journal of Statistics*, 50(1):53–67.

Genizi, A. (1993). Decomposition of r^2 in multiple regression with correlated regressors. *Statistica Sinica*, pages 407–420.

Gilquin, L., Prieur, C., and Arnaud, E. (2015). Replication procedure for grouped Sobol' indices estimation in dependent uncertainty spaces. *Information and Inference*, 4(4):354–379.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2).

Hooker, G. (2012). Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*.

Keinan, A., Hilgetag, C. C., Meilijson, I., and Ruppin, E. (2004). Causal localization of neural function: the Shapley value method. *Neurocomputing*, 58:215–222.

Kucherenko, S., Klymenko, O. V., and Shah, N. (2016). Sobol' indices for problems defined in non-rectangular domains. Technical report, arXiv:1605.05069.

Lai, C. D. (1978). Morgenstern's bivariate distribution and its application to point processes. *Journal of Mathematical Analysis and Applications*, 65(2):247–256.

Lebrun, R. and Dutfoy, A. (2014). Copulas for order statistics with prescribed margins. *Journal of Multivariate Analysis*, 128:120–133.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott Foresman and Company, Glenview, IL.

Owen, A. B. (2013). Variance components and generalized Sobol' indices. *Journal of Uncertainty Quantification*, 1(1):19–41.

Owen, A. B. (2014). Sobol' indices and Shapley value. *Journal on Uncertainty Quantification*, 2:245–251.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contribution to the Theory of Games II (Annals of Mathematics Studies 28)*, pages 307–317. Princeton University Press, Princeton, NJ.

Sobol', I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2(1):112–118. (In Russian).

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414.

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

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118–184.

Winter, E. (2002). The Shapley value. *Handbook of game theory with economic applications*, 3:2025–2054.

Wu, C. F. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*. John Wiley & Sons.

6 Proofs

6.1 Proof of Theorem 2

Recall that $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\beta$ where $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$. We also assumed that Σ is of full rank. Now $\mathrm{var}(\boldsymbol{x}_{-u} \,|\, \boldsymbol{x}_u) = \Sigma_{-u,-u} - \Sigma_{-u,u}\Sigma_{u,u}^{-1}\Sigma_{u,-u}$, and so

$$\mathrm{var}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u) = \mathrm{var}(\boldsymbol{x}_u^{\mathsf{T}}\beta_u + \boldsymbol{x}_{-u}^{\mathsf{T}}\beta_{-u} \,|\, \boldsymbol{x}_u)$$

$$
\begin{aligned}
&= \operatorname{var}(\boldsymbol{x}_{-u}^{\mathsf{T}}\beta_{-u} \,|\, \boldsymbol{x}_u) \\
&= \beta_{-u}^{\mathsf{T}}\big(\Sigma_{-u,-u} - \Sigma_{-u,u}\Sigma_{u,u}^{-1}\Sigma_{u,-u}\big)\beta_{-u}.
\end{aligned}
$$

We will use $v = v(j,u) \equiv -u - \{j\}$. It helps to visualize the partitioned covariance matrix

$$
\Sigma = \begin{pmatrix} \Sigma_{uu} & \Sigma_{uj} & \Sigma_{uv} \\ \Sigma_{ju} & \Sigma_{jj} & \Sigma_{jv} \\ \Sigma_{vu} & \Sigma_{vj} & \Sigma_{vv} \end{pmatrix}
$$

if the indices have been ordered for those in u to precede j which precedes those in v. For this section only, we make a further notational compression shortening $u + \{j\}$ to $u + j$. Next

$$
\begin{aligned}
\underline{\tau}_{u+j}^2 - \underline{\tau}_u^2 &= \operatorname{var}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_u) - \operatorname{var}(f(\boldsymbol{x}) \,|\, \boldsymbol{x}_{u+j}) \\
&= \beta_{-u}^{\mathsf{T}}\big(\Sigma_{-u,-u} - \Sigma_{-u,u}\Sigma_{u,u}^{-1}\Sigma_{u,-u}\big)\beta_{-u} \\
&\quad - \beta_v^{\mathsf{T}}\big(\Sigma_{vv} - \Sigma_{v,u+j}\Sigma_{u+j,u+j}^{-1}\Sigma_{u+j,v}\big)\beta_v.
\end{aligned}
$$

Using the formula for the inverse of a partitioned matrix, we find that

$$
\Sigma_{u+j,u+j}^{-1} = \begin{pmatrix} \Sigma_{uu}^{-1} + \Sigma_{uu}^{-1}\Sigma_{uj}D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1} & -\Sigma_{uu}^{-1}\Sigma_{uj}D_j(u) \\ -D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1} & D_j(u) \end{pmatrix},
$$

where $D_j(u) = (\Sigma_{jj} - \Sigma_{ju}\Sigma_{uu}^{-1}\Sigma_{uj})^{-1} = \operatorname{var}(x_j \,|\, \boldsymbol{x}_u)^{-1}$, which exists because Σ has full rank. Continuing,

$$
\begin{aligned}
&\Sigma_{v,u+j}\Sigma_{u+j,u+j}^{-1}\Sigma_{u+j,v} \\
&= \begin{pmatrix} \Sigma_{vu} & \Sigma_{vj} \end{pmatrix} \begin{pmatrix} \Sigma_{uu}^{-1} + \Sigma_{uu}^{-1}\Sigma_{uj}D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1} & -\Sigma_{uu}^{-1}\Sigma_{uj}D_j(u) \\ -D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1} & D_j(u) \end{pmatrix} \begin{pmatrix} \Sigma_{uv} \\ \Sigma_{jv} \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{vu} & \Sigma_{vj} \end{pmatrix} \begin{pmatrix} \Sigma_{uu}^{-1}\Sigma_{uv} + \Sigma_{uu}^{-1}\Sigma_{uj}D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1}\Sigma_{uv} - \Sigma_{uu}^{-1}\Sigma_{uj}D_j(u)\Sigma_{jv} \\ -D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1}\Sigma_{uv} + D_j(u)\Sigma_{jv} \end{pmatrix} \\
&= \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv} + \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uj}D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1}\Sigma_{uv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uj}D_j(u)\Sigma_{jv} \\
&\quad - \Sigma_{vj}D_j(u)\Sigma_{ju}\Sigma_{uu}^{-1}\Sigma_{uv} + \Sigma_{vj}D_j(u)\Sigma_{jv} \\
&= \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv} + D_j(u)\big(\Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uj} - \Sigma_{vj}\big)\big(\Sigma_{ju}\Sigma_{uu}^{-1}\Sigma_{uv} - \Sigma_{jv}\big) \\
&= \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv} + D_j(u)\operatorname{cov}(\boldsymbol{x}_v, x_j \,|\, \boldsymbol{x}_u)\operatorname{cov}(x_j, \boldsymbol{x}_v \,|\, \boldsymbol{x}_u)
\end{aligned}
$$

recalling that $D_j(u)$ is a scalar.

Now $\underline{\tau}_{u+j}^2 - \underline{\tau}_u^2$ is

$$
\begin{aligned}
&\beta_{-u}^{\mathsf{T}}\operatorname{cov}(\boldsymbol{x}_{-u} \,|\, \boldsymbol{x}_u)\beta_{-u} - \beta_v^{\mathsf{T}}\Sigma_{vv}\beta_v \\
&\quad + \beta_v^{\mathsf{T}}\big(\Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv} + D_j(u)\operatorname{cov}(\boldsymbol{x}_v, x_j \,|\, \boldsymbol{x}_u)\operatorname{cov}(x_j, \boldsymbol{x}_v \,|\, \boldsymbol{x}_u)\big)\beta_v \\
&= \beta_{-u}^{\mathsf{T}}\operatorname{cov}(\boldsymbol{x}_{-u} \,|\, \boldsymbol{x}_u)\beta_{-u} - \beta_v^{\mathsf{T}}\operatorname{cov}(\boldsymbol{x}_v \,|\, \boldsymbol{x}_u)\beta_v \\
&\quad + D_j(u)\beta_v^{\mathsf{T}}\operatorname{cov}(\boldsymbol{x}_v, x_j \,|\, \boldsymbol{x}_u)\operatorname{cov}(x_j, \boldsymbol{x}_v \,|\, \boldsymbol{x}_u)\beta_v \\
&= \Sigma_{jj}\beta_j^2 + \beta_j\Sigma_{jv}\beta_v + \beta_v^{\mathsf{T}}\Sigma_{vj}\beta_j
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \beta_j^2 \Sigma_{ju} \Sigma_{uu}^{-1} \Sigma_{uj} - \beta_j \Sigma_{ju} \Sigma_{uu}^{-1} \Sigma_{uv} \beta_v - \beta_v^\mathsf{T} \Sigma_{vu} \Sigma_{uu}^{-1} \Sigma_{uj} \beta_j \\
&\quad + D_j(u) \beta_v^\mathsf{T} \mathrm{cov}(\boldsymbol{x}_v, x_j \,|\, \boldsymbol{x}_u) \mathrm{cov}(x_j, \boldsymbol{x}_v \,|\, \boldsymbol{x}_u) \beta_v \\
&= \beta_j^2 \mathrm{var}(x_j \,|\, \boldsymbol{x}_u) + 2\beta_j \mathrm{cov}(x_j, \boldsymbol{x}_v \,|\, \boldsymbol{x}_u) \beta_v \\
&\quad + D_j(u) \beta_v^\mathsf{T} \mathrm{cov}(\boldsymbol{x}_v, x_j \,|\, \boldsymbol{x}_u) \mathrm{cov}(x_j, \boldsymbol{x}_v \,|\, \boldsymbol{x}_u) \beta_v.
\end{aligned}
$$

Putting this together, the Shapley value of variable j is

$$
\begin{aligned}
\phi_j = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \Big(& \beta_j^2 \mathrm{var}(x_j \,|\, \boldsymbol{x}_u) + 2\beta_j \mathrm{cov}(x_j, \boldsymbol{x}_{-u-j} \,|\, \boldsymbol{x}_u) \beta_{-u-j} \\
& + \mathrm{var}(x_j \,|\, \boldsymbol{x}_u)^{-1} \beta_{-u-j}^\mathsf{T} \mathrm{cov}(\boldsymbol{x}_{-u-j}, x_j \,|\, \boldsymbol{x}_u) \mathrm{cov}(x_j, \boldsymbol{x}_{-u-j} \,|\, \boldsymbol{x}_u) \beta_{-u-j} \Big).
\end{aligned} \tag{20}
$$

Writing

$$
\mathrm{cov}(x_j, \boldsymbol{x}_{-u}^\mathsf{T} \beta_{-u} \,|\, \boldsymbol{x}_u) = \mathrm{cov}(x_j, \boldsymbol{x}_{-u-j}^\mathsf{T} \beta_{-u-j} \,|\, \boldsymbol{x}_u) + \beta_j \mathrm{var}(x_j \,|\, \boldsymbol{x}_u)
$$

we then find that $\mathrm{cov}(x_j, \boldsymbol{x}_{-u}^\mathsf{T} \beta_{-u} \,|\, \boldsymbol{x}_u)^2 / \mathrm{var}(x_j \,|\, \boldsymbol{x}_u)$ equals the factor to the right of $\binom{d-1}{|u|}$ in (20).

6.2 Proof of Theorem 3

Without loss of generality take $y_0 = 0$. Then $\mu = p_1 y_1 + p_2 y_2$ and $\sigma^2 = p_1 y_1^2 + p_2 y_2^2 - \mu^2$.

Now with $y_0 = 0$,

$$
\begin{aligned}
\mathrm{var}(\mathbb{E}(y \,|\, x_1)) &= (p_0 + p_2) \Big(\frac{y_2 p_2}{p_0 + p_2} - \mu \Big)^2 + p_1 (y_1 - \mu)^2 \\
&= (1 - p_1) \Big(\frac{y_2 p_2}{1 - p_1} - \mu \Big)^2 + p_1 (y_1 - \mu)^2 \\
&= \frac{p_2^2 y_2^2}{1 - p_1} - 2\mu y_2 p_2 + \mu^2 (1 - p_1) + p_1 (y_1 - \mu)^2 \\
&= \frac{p_2^2 y_2^2}{1 - p_1} - 2(p_1 y_1 + p_2 y_2) y_2 p_2 + (p_1 y_1 + p_2 y_2)^2 (1 - p_1) + p_1 (y_1 (1 - p_1) - p_2 y_2)^2 \\
&= y_2^2 \Big(\frac{p_2^2}{1 - p_1} - 2p_2^2 + p_2^2 (1 - p_1) + p_1 p_2^2 \Big) \\
&\quad + y_1^2 \Big(p_1^2 (1 - p_1) + p_1 (1 - p_1)^2 \Big) \\
&\quad + y_1 y_2 \Big(-2p_1 p_2 + 2p_1 p_2 (1 - p_1) - 2p_1 p_2 (1 - p_1) \Big) \\
&= y_2^2 \Big(\frac{p_2^2}{1 - p_1} - p_2^2 \Big) + y_1^2 p_1 (1 - p_1) - 2 y_1 y_2 p_1 p_2 \\
&= y_2^2 \frac{p_1 p_2^2}{1 - p_1} + y_1^2 p_1 (1 - p_1) - 2 y_1 y_2 p_1 p_2.
\end{aligned}
$$

Then $\mathrm{var}(\mathbb{E}(y \mid x_1)) - \mathrm{var}(\mathbb{E}(y \mid x_2))$ equals

$$\begin{aligned}
&y_2^2 \frac{p_1 p_2^2}{1-p_1} + y_1^2 p_1(1-p_1) - y_1^2 \frac{p_2 p_1^2}{1-p_2} - y_2^2 p_2(1-p_2) \\
&= y_2^2 \Big(\frac{p_1 p_2^2}{1-p_1} - p_2(1-p_2) \Big) + y_1^2 \Big(p_1(1-p_1) - \frac{p_2 p_1^2}{1-p_2} \Big) \\
&= y_1^2 \Big(\frac{p_0 p_1}{1-p_2} \Big) - y_2^2 \Big(\frac{p_0 p_2}{1-p_1} \Big).
\end{aligned}$$

Finally, the relative importance of variable x_1 is

$$\begin{aligned}
\frac{1}{2}\Big(1 + \frac{y_1^2\big(\frac{p_0 p_1}{1-p_2}\big) - y_2^2\big(\frac{p_0 p_2}{1-p_1}\big)}{\sigma^2}\Big) &= \frac{1}{2}\Big(1 + \frac{p_0}{\sigma^2} \frac{y_1^2 p_1(1-p_1) - y_2^2 p_2(1-p_2)}{(1-p_1)(1-p_2)}\Big) \\
&= \frac{1}{2}\Big(1 + \frac{p_0}{\sigma^2}\Big(\frac{p_1 y_1^2}{1-p_2} - \frac{p_2 y_2^2}{1-p_1}\Big)\Big).
\end{aligned}$$

6.3 Proof of Theorem 4

Recall that the random vector $\boldsymbol{x} \in [0,\infty)^d$ has independent components x_j. They are exponentially distributed and $\mathbb{E}(x_j) = 1/\lambda_j$ for $0 < \lambda_j < \infty$. Let $M_u = \max_{j \in u} x_j$ and define value $\mathrm{val}(u) = \mathbb{E}(M_u)$. Our first step is to evaluate the expected value of a maximum of independent not identically distributed exponential random variables.

Proposition 5.

$$\mathbb{E}(M_u) = \sum_{\varnothing \neq v \subseteq u} (-1)^{|v|-1} \frac{1}{\sum_{j \in v} \lambda_j}.$$

Proof. First $\Pr(M_u < x) = \prod_{j \in u} \Pr(x_j < x) = \prod_{j \in u}(1 - e^{-\lambda_j x})$. Then,

$$\begin{aligned}
\mathbb{E}(M_u) &= \int_0^\infty \left(1 - \prod_{j \in u}(1 - e^{-\lambda_j x})\right) \mathrm{d}x \\
&= \int_0^\infty \left(1 - \sum_{v \subseteq u}(-e^{-\lambda_j x})\right) \mathrm{d}x \\
&= \sum_{\varnothing \neq v \subseteq u} (-1)^{|v|-1} \int_0^\infty e^{-x \sum_{j \in v} \lambda_j} \, \mathrm{d}x \\
&= \sum_{\varnothing \neq v \subseteq u} (-1)^{|v|-1} \frac{1}{\sum_{j \in v} \lambda_j}.
\end{aligned}$$

□

Using Proposition 5 we get Shapley value

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (\mathrm{val}(u+j) - \mathrm{val}(u))$$

$$= \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} \left(\sum_{\varnothing \neq v \subseteq u+j} (-1)^{|v|-1} \frac{1}{\sum_{\ell \in v} \lambda_\ell} - \sum_{\varnothing \neq v \subseteq u} (-1)^{|v|-1} \frac{1}{\sum_{\ell \in v} \lambda_\ell} \right)$$
$$= \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} \sum_{w \subseteq u} (-1)^{|w|} \frac{1}{\sum_{\ell \in w+j} \lambda_\ell}.$$

Introducing the 'slack variable' v with $u = v + w$,

$$\begin{aligned}
\phi_j &= \frac{1}{d} \sum_{w \subseteq -\{j\}} (-1)^{|w|} \frac{1}{\sum_{\ell \in w+j} \lambda_\ell} \sum_{v \subseteq -\{j\}-w} \binom{d-1}{|v+w|}^{-1} \\
&= \frac{1}{d} \sum_{w \subseteq -\{j\}} (-1)^{|w|} \frac{1}{\sum_{\ell \in w+j} \lambda_\ell} \sum_{r=0}^{d-1-|w|} \binom{d-1-|w|}{r} \Big/ \binom{d-1}{r+|w|} \\
&= \frac{1}{d} \sum_{w : j \in w} (-1)^{|w|-1} \frac{1}{\sum_{\ell \in w} \lambda_\ell} \sum_{r=0}^{d-|w|} \binom{d-|w|}{r} \Big/ \binom{d-1}{r-1+|w|}.
\end{aligned}$$

The following diagonal sum identity for binomial coefficients will be useful:

$$\sum_{r=0}^{A} \binom{L+r}{r} = \binom{A+L+1}{A}.$$

Using that identity at the third step below,

$$\begin{aligned}
\sum_{r=0}^{d-|w|} \binom{d-|w|}{r} \Big/ \binom{d-1}{r-1+|w|} &= \sum_{r=0}^{d-|w|} \frac{(d-|w|)!}{r!} \Big/ \frac{(d-1)!}{(r-1+|w|)!} \\
&= \frac{(d-|w|)!}{(d-1)!} (|w|-1)! \sum_{r=0}^{d-|w|} \binom{r-1+|w|}{r} \\
&= \frac{(d-|w|)!}{(d-1)!} (|w|-1)! \binom{d}{d-|w|} \\
&= \frac{d}{|w|}.
\end{aligned}$$

As a result,

$$\phi_j = \sum_{w : j \in w} \frac{1}{|w|} (-1)^{|w|-1} \frac{1}{\sum_{\ell \in w} \lambda_\ell}$$

which after slight rearrangement gives the conclusion of Theorem 4.