Efficient nonparametric statistical inference on population feature importance using Shapley values

Brian D. Williamson [*1] **Jean Feng** [*2]

Abstract

The true population-level importance of a variable in a prediction task provides useful knowledge about the underlying data-generating mechanism and can help in deciding which measurements to collect in subsequent experiments. Valid statistical inference on this importance is a key component in understanding the population of interest. We present a computationally efficient procedure for estimating and obtaining valid statistical inference on the **S**hapley **P**opulation **V**ariable **I**mportance **M**easure (SPVIM). Although the computational complexity of the true SPVIM scales exponentially with the number of variables, we propose an estimator based on randomly sampling only $\Theta(n)$ feature subsets given n observations. We prove that our estimator converges at an asymptotically optimal rate. Moreover, by deriving the asymptotic distribution of our estimator, we construct valid confidence intervals and hypothesis tests. Our procedure has good finite-sample performance in simulations, and for an in-hospital mortality prediction task produces similar variable importance estimates when different machine learning algorithms are applied.

1. Introduction

In many scientific applications, understanding the intrinsic predictive value of a variable can shed light on the internal mechanisms relating the variable to the outcome of interest, help build future models, and guide experimental design. For example, hospital administrators may want to know the important features to collect for predicting patient outcomes. Likewise, vaccine researchers may want to know the most important molecular phenotypes to measure that are most predictive of binding or vaccine efficacy (see, e.g., Dunning, 2006). Variable importance measures (VIMs) provide necessary information towards answering these questions.

Our interest here is in statistical inference on the population VIM. This VIM quantifies the predictive value of a variable within the oracle prediction model f_0 defined relative to an arbitrary predictiveness measure V. For many choices of V, f_0 is either the conditional mean outcome given covariates (e.g., if $V = R^2$) or a simple functional of this conditional mean (e.g., if $V =$ classification accuracy). We note that population VIMs are distinct from algorithmic VIMs, which describe the importance of a variable within a fitted model $\hat{f}$ (see, e.g., Breiman, 2001; Garson, 1991; Murdoch et al., 2019). Although algorithmic VIMs have been used as a proxy for population VIMs out of convenience, differences between $\hat{f}$ and f_0 can often lead to substantially different interpretations of the resulting VIMs. Whereas an algorithmic VIM necessarily varies across fitted models, a population VIM is independent of the specific procedure used to estimate f_0.

Existing population VIMs suffer from a number of issues. Traditionally, population VIMs have relied on restrictive parametric assumptions (e.g., R^2 in linear models; see, e.g., Grömping, 2007; Nathans et al., 2012), which can lead to misleading results if the parametric model does not hold. Recent work has focused on extending these definitions by removing the parametric assumptions (Feng et al., 2018; Williamson et al., 2020b); however, these definitions define importance of a variable with respect to the others and assign near-zero importance when features are highly correlated. Other VIMs require strong assumptions on the design to be valid (e.g., ANOVA), but again fail in simple cases with correlated variables. To address this, Owen and Prieur (2017) proposed using Shapley values to quantify the population VIM, where the value function is the variance explained; these VIMs inherit many desirable theoretical properties from the Shapley value. In fact, contemporary work has also defined the ideal estimand of algorithmic VIM estimation procedures to be the Shapley population VIM (SPVIM) (Covert et al., 2020).

[*]Equal contribution [1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA [2]Department of Biostatistics, University of Washington, Seattle, WA. Correspondence to: Brian D. Williamson <bwillia2@fredhutch.org>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Unfortunately, exact estimation of SPVIM is computationally intractable in general settings (Owen and Prieur, 2017): the SPVIM is defined as the sum of 2^p terms, where p is the number of features and each term depends on estimating the conditional mean function with respect to a unique feature subset. Previous approaches have either suggested sampling as many subsets as possible to estimate the Shapley value (see, e.g., Castro et al., 2009) or utilized special properties of tree estimators to reduce the number of subsets required (Lundberg et al., 2020). Notably, Štrumbelj and Kononenko (2014) analyzed the asymptotic distribution of a sampling-based estimator of Shapley algorithmic variable importance to derive confidence intervals.

In this paper, we combine the aforementioned developments and provide a nonparametric statistical inference procedure for SPVIM. We generalize previous definitions of SPVIM and use an arbitrary measure of predictiveness V. We tackle the computational complexity of the problem by randomly sampling feature subsets according to the Shapley value weights and then fitting corresponding models. We derive the asymptotic distribution of this sampling-based SPVIM estimator and show that the error from our proposed procedure can be decomposed into two components: the error from estimating the oracle prediction models and the error from omitting summands from the Shapley value estimand. Given n training observations, we find that our estimator only needs to sample $m = \Theta(n)$ subsets to converge at an asymptotically optimal rate. Moreover, since the subset sampling distribution is highly skewed, the number of *unique* feature subsets is much smaller than m in practice. We then use the asymptotic distribution to construct asymptotically unbiased point estimates, valid confidence intervals, and hypothesis tests with proper type I error control.

We demonstrate the validity of our approach in a simulation study and estimate the SPVIM of hospital measurements for predicting mortality in the intensive care unit (ICU). All numerical results can be replicated using code available on GitHub at `bdwilliamson/spvim_supplementary`; the proposed methods are also implemented in the Python package `vimpy` and the R package `vimp`.

2. Variable importance

2.1. Data structure and notation

Let $\mathcal{M}$ be a nonparametric class of joint distributions over covariates $X = (X_1, \ldots, X_p) \in \mathcal{X} \subseteq \mathbb{R}^p$ and response $Y \in \mathcal{Y} \subseteq \mathbb{R}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the sample spaces of X and Y, respectively. Suppose that each observation O consists of (X, Y). In this article, we consider observations $O_1, \ldots, O_n$ drawn independently according to a joint probability distribution $P_0 \in \mathcal{M}$.

Next, we define the feature subsets and oracle prediction models of interest. We take $\mathcal{S}$ to be the power set of $N := \{1, \ldots, p\}$. Let $s_{(j)}$ for $j = 1, \ldots, 2^p$ be an ordered sequence of the subsets in $\mathcal{S}$, where $s_{(1)} = \emptyset$ and $s_{(2^p)} = N$. For any index set $s \in \mathcal{S}$, we denote by $\mathcal{X}_s$ and $\mathcal{X}_{-s}$ the sample spaces of X_s and X_{-s}, respectively. We denote by a_s and a_{-s} the elements of a vector a with indices in s and not in s, respectively. We also consider the binary vector $z(s) \in \mathbb{R}^{p+1}$ for each $s \in \mathcal{S}$, where $z(s)_1 = 1$ for all $s \in \mathcal{S}$ and $z(s)_{k+1} = I(k \in s)$ for $k = 1, \ldots, p$. Finally, we consider a rich class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ endowed with a norm $\|\cdot\|_{\mathcal{F}}$. For any $s \in \mathcal{S}$, we define the subset $\mathcal{F}_s := \{f \in \mathcal{F} : f(u) = f(v) \text{ for all } u, v \in \mathcal{X} \text{ satisfying } u_s = v_s\}$ of functions in $\mathcal{F}$ whose evaluation ignores elements of the input x with index not in s. In all examples we consider, we take $\mathcal{F}$ to be a rich class of functions that is essentially unrestricted up to regularity conditions.

2.2. Oracle predictiveness

We define the importance of a variable at the population level in terms of its oracle predictiveness. This predictiveness is measured by a real-valued functional $V : \mathcal{F} \times \mathcal{M} \mapsto \mathbb{R}$. We assume that larger values of $V(f, P)$ imply higher predictiveness. Examples of predictiveness measures — including R^2, deviance, area under the ROC curve, and classification accuracy — are provided in Williamson et al. (2020b).

The oracle predictiveness is the maximum achievable predictiveness over a class of prediction functions. More formally, we define the *total oracle predictiveness* $v_{0,N} := \max_{f \in \mathcal{F}} V(f, P_0)$ and its associated oracle prediction function $f_{0,N} \in \operatorname{argmax}_{f \in \mathcal{F}} V(f, P_0)$. For many machine learning algorithms, $f_{0,N}$ is the target of interest. We further define the oracle prediction function $f_{0,s}$ that maximizes $V(f, P_0)$ over all $f \in \mathcal{F}_s$; the *marginal oracle predictiveness* $v_{0,s} := V(f_{0,s}, P_0)$ quantifies the prediction potential of features with index in s. The *null oracle predictiveness* $v_{0,\emptyset} := V(f_{0,\emptyset}, P_0)$ quantifies the prediction potential of a model that uses no covariate information. Finally, let $v_0 := [v_{0,\emptyset}, v_{0,s_{(2)}}, \ldots, v_{0,N}]^\top$ denote the 2^p-dimensional vector of predictiveness measures for all subsets in $\mathcal{S}$. The predictiveness measure $v_{0,s_{(j)}}$ is defined relative to the population P_0, a joint distribution in the nonparametric statistical model $\mathcal{M}$; thus, its interpretation is tied to neither any particular estimation procedure nor any parametric assumptions.

2.3. The Shapley population variable importance measure

We now define a population VIM using the classical form of the Shapley value (see, e.g., Shapley, 1953; Charnes et al., 1988) with an arbitrary measure of predictiveness V. Specifically, the *Shapley population variable importance measure* (SPVIM) of the variable X_j is the average gain

in oracle predictiveness from including feature X_j over all possible subsets:

$$\psi_{0,0,j} := \sum_{s \in N \setminus \{j\}} \frac{1}{p} \binom{p-1}{|s|}^{-1} \{V(f_{0,s\cup j}, P_0) - V(f_{0,s}, P_0)\}, \tag{1}$$

where the indices of ψ describe the number of subsets, the distribution P_0, and the feature of interest j, respectively. We use the index 0 to indicate that the SPVIM is computed using all subsets and the true distribution P_0. SPVIMs inherit the following properties from Shapley values (Shapley, 1953):

- Non-negativity: by construction, $\psi_{0,0,j} \geq 0$.
- Additivity[1]: the sum of the SPVIMs across all variables is equal to the difference between the total and null oracle predictiveness,

$$\sum_{j=1}^{p} \psi_{0,0,j} = v_{0,N} - v_{0,\emptyset} \tag{2}$$

- Symmetry: if $X_i = X_j$, then $\psi_{0,0,i} = \psi_{0,0,j}$.
- Null feature: if X_j provides no added predictive value, i.e., $v_{0,s\cup j} = v_{0,s}$ for all $s \subseteq (N \setminus \{j\})$, then its SPVIM value is $\psi_{0,0,j} = 0$.
- Linearity: if $\tilde{V} \equiv \alpha V$, then its associated SPVIM values are $\tilde{\psi}_{0,0,j} = \alpha \psi_{0,0,j}$ for all $j \in \{1, \ldots, p\}$.

Because SPVIMs satisfy these properties, they clearly address the issue of correlated features: given collinear variables X_j and X_k that are each marginally predictive, previous nonparametric population VIMs (see, e.g., Williamson et al., 2020b) would assign zero importance to both variables whereas SPVIM would assign the same positive value to both variables.

In this paper, we take advantage of an alternate formulation of the Shapley value noted in previous work (see, e.g., Charnes et al., 1988; Lundberg and Lee, 2017). In particular, we can rewrite the weighted average in (1) as the solution of a weighted linear regression problem, where we treat the predictiveness of a feature subset $v_{0,s}$ as the response and the subset membership $z(s)$ as the covariates. Define a diagonal matrix of weights $W \in \mathbb{R}^{2^p \times 2^p}$ where $W_{1,1} = W_{2^p,2^p} = 1$, and for any $j \in 2, \ldots, 2^p - 1$, $W_{j,j} = \binom{p-2}{|s_{(j)}|-1}^{-1}$. The matrix $Z \in \mathbb{R}^{2^p \times (p+1)}$ consists of the stacked $z(s)$ vectors for each $s \in \mathcal{S}$. Setting $\psi_{0,0,\emptyset} := v_{0,\emptyset}$, we denote by $\psi_{0,0}$ the $(p+1)$-dimensional vector of population Shapley values. Then (1) is equivalent to

$$\psi_{0,0} := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|\sqrt{W}(Z\psi - v_0)\|_2^2, \tag{3}$$

a result that we prove in the Supplement. If we define the distribution Q_0 over subsets $\mathcal{S}$ with probability mass function assigning weight $\binom{p-2}{|S|-1}^{-1}$ for $S \in \mathcal{S} \setminus \{\emptyset, N\}$ and weight 1 for $S \in \{\emptyset, N\}$ (scaled so that the weights sum to one), then (3) is equivalent to a population average:

$$\psi_{0,0} \equiv \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} E_{Q_0}\left[(z(S)\psi - v_{0,S})^2\right].$$

We will use this fact in our estimation procedure below.

3. Estimation and inference

3.1. Plug-in estimation

We now discuss how to estimate the SPVIM values for all p features using independent observations $O_1, \ldots, O_n$ drawn from P_0. Definition (3) suggests considering an estimator based on plugging in estimators of each individual component. We discuss each component in turn.

First, we estimate the predictiveness measure $v_{0,s} = V(f_{0,s}, P_0)$ for a subset $s \in \mathcal{S}$ by plugging in estimates of the oracle function $f_{0,s}$ and the distribution P_0. A simple approach is to partition the data into a training set and a validation set, construct an estimator $f_{n,s}$ for $f_{0,s}$ on the training data (using only the observed covariates in s), and estimate P_0 using the empirical distribution of the validation set P_V. Using this training-validation split, our estimate of predictiveness is then

$$v_{n,s} = V(f_{n,s}, P_V). \tag{4}$$

An alternative approach is to perform K-fold cross-fitting, where we partition the data into K subsets of roughly equal size and, for each $k \in \{1, \ldots, K\}$, construct an estimator $f_{k,n,s}$ based on all the data except for the kth subset. Let $P_{k,n}$ be the empirical distribution of the kth subset. Then we could estimate $v_{0,s}$ using

$$v_{n,s} = \frac{1}{K} \sum_{k=1}^{K} V(f_{k,n,s}, P_{k,n}). \tag{5}$$

If we had the entire estimated vector of predictiveness measures v_n, we could estimate $\psi_{0,0}$ using the plug-in estimator

$$\psi_{0,n} := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} E_{Q_0}\left[(Z(S)\psi - v_{n,S})^2\right]. \tag{6}$$

Unfortunately, obtaining v_n requires training 2^p models, rendering this a computationally intractable task in general.

[1] In the Shapley value literature, this additivity property is referred to as "efficiency". However, this notion of efficiency is very different from statistical efficiency, which is related to the asymptotic variance of a statistical estimator.

Instead, we can replace Q_0 in (6) with an empirical distribution estimator Q_m obtained by sampling m subsets from $\mathcal{S}$ according to Q_0. This leads us to the SPVIM estimator $\psi_{m,n}$ which solves the constrained least squares problem

$$\min_{\psi \in \mathbb{R}^{p+1}} E_{Q_m}\left[(Z(S)\psi - v_{n,S})^2\right] \text{ subject to } G\psi = c_n, \tag{7}$$

where $G := [z(\emptyset)^\top, z(N)^\top]^\top \in \mathbb{R}^{2\times(p+1)}$ and $c_n := [v_{n,\emptyset}, v_{n,N}]^\top \in \mathbb{R}^2$. The constraint ensures that the estimated SPVIMs satisfy the additivity property (2) and that the estimated SPVIM for the null set is the estimated null predictiveness value.

This constrained least squares problem can be solved by forming a Lagrangian and inverting its Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004). More specifically, let $s_1, \ldots, s_\ell$ be the unique subsets in Q_m. Let W_m be the $\ell \times \ell$ diagonal matrix where the kth diagonal element is the probability mass of s_k in Q_m. Let $v_{m,n} = (v_{n,s_1}, \ldots, v_{n,s_\ell})$ be the estimated predictiveness measures for the ℓ subsets. Let Z_m be the stack of vectors $z(s_1), \ldots, z(s_\ell)$. Then (7) can also be written as

$$\min_{\psi \in \mathbb{R}^{p+1}} \left\| \sqrt{W_m}\left(Z_m\psi - v_{m,n}\right)\right\|_2^2 \text{ subject to } G\psi = c_n.$$

Solving the KKT conditions with Lagrange multipliers denoted by λ, we obtain a closed-form SPVIM estimator:

$$\begin{bmatrix}\psi_{m,n}\\ \lambda\end{bmatrix} = \begin{bmatrix}2Z_m^\top W_m Z_m & G^\top\\ G & 0\end{bmatrix}^{-1}\begin{bmatrix}2\sqrt{W_m}v_{m,n}\\ c_n\end{bmatrix}. \tag{8}$$

To ensure that (7) has a unique solution, we select a sufficiently large value of m so that Q_m inclues at least $p+1$ unique subsets. The full estimation procedure is given in Algorithm 1.

We now describe the properties listed in Section 2.3 that are satisfied by this sampling-based SPVIM estimator. It is easy to see that the additivity, symmetry, and linearity properties always hold. One possible concern is that the nonnegativity property can be violated. Nevertheless, in practice we find that negative SPVIM estimates are close to zero and the 95% confidence intervals cover zero. If nonnegativity is truly a concern, one can also add a nonnegative constraint to (7). Finally, the null feature property holds with respect to *estimated* predictiveness values and the sampled subsets. Note that this property is only relevant for discrete predictiveness measures like 0-1 classification accuracy, since the estimated predictiveness values are rarely exactly the same for continuous predictiveness measures like R^2.

The plug-in estimator $\psi_{m,n}$ is appealing due to its simplicity. In general, however, such an estimator may fail to be consistent at rate $n^{-1/2}$ if the population optimizers $f_{0,s}$ are flexibly estimated. This phenomenon is due in large part to the optimal bias-variance tradeoff for estimating $f_{0,s}$ differing in general from the optimal bias-variance tradeoff for estimating $v_{n,s}$. Plug-in estimators typically inherit much of the bias from estimating $f_{0,s}$, and this bias does not in general tend to zero sufficiently fast to allow $n^{-1/2}$-rate estimation of $\psi_{0,0}$ (Williamson et al., 2020a). In the next section, we extend the results of Williamson et al. (2020b) to describe conditions under which the estimator $\psi_{m,n}$ is asymptotically normal.

Algorithm 1 Estimation of SPVIM

1: Input initial parameter $\gamma \geq 1$.
2: Sample $m = \gamma n$ subsets from Q_0, denoted $s_1, \ldots, s_m$.
3: Estimate prediction functions $f_{n,s}$ for each $s \in \{s_1, \ldots, s_m\} \cup \{\emptyset, N\}$.
4: Compute predictiveness estimates $v_{n,s}$ for $s \in \{s_1, \ldots, s_m\} \cup \{\emptyset, N\}$ using a training-validation split (see Equation (4)) or K-fold cross-fitting (see Equation (5)).
5: Solve for $\psi_{m,n}$ using Equation (8).

3.2. Large-sample inferential properties

We now study the conditions under which $\psi_{m,n}$ is an asymptotically normal estimator of the SPVIM $\psi_{0,0}$. Using these conditions, we can design a procedure to construct confidence intervals and hypothesis tests. To do this, we decompose the error of our estimator $\psi_{m,n}$ into the following components:

$$\psi_{m,n} - \psi_{0,0} = (\psi_{0,n} - \psi_{0,0}) + (\psi_{m,0} - \psi_{0,0}) + r_{m,n}, \tag{9}$$

where $\psi_{m,0}$ is obtained by replacing $v_{n,S}$ with $v_{0,S}$ in (7) and $r_{m,n} := (\psi_{m,n} - \psi_{m,0}) - (\psi_{0,n} - \psi_{0,0})$. Each term on the right-hand side of (9) can then be studied separately to determine the large-sample behavior of $\psi_{m,n}$. The first term is the error of the estimator $\psi_{0,n}$ (6) constructed using prediction functions $f_{n,s}$ estimated using n observations for all subsets s. The second term is the error of the estimator $\psi_{m,0}$ constructed using oracle prediction functions for sampled subsets in Q_m. In other words, the first term characterizes the error contribution from sampling training observations and the second term characterizes the error contribution from sampling subsets. The third term is a difference-in-differences remainder term that we prove to be negligible under some regularity conditions. Based on this decomposition, we will show that the asymptotic variance of $\sqrt{n}(\psi_{m,n} - \psi_{0,0})$ is simply the sum of the asymptotic variances of the first and second error terms.

Our result makes use of several conditions that require additional notation. These conditions were initially provided in Williamson et al. (2020b). We define the linear

space $\mathcal{R} := \{c(P_1 - P_2) : c \in \mathbb{R}, P_1, P_2 \in \mathcal{M}\}$ of finite signed measures generated by $\mathcal{M}$. For any $R \in \mathcal{R}$, e.g., $R = c(P_1 - P_2)$, we consider the supremum norm $\|R\|_\infty := |c| \sup_o |F_1(o) - F_2(o)|$, where F_1 and F_2 are the distribution functions corresponding to P_1 and P_2, respectively. Next, we define the following notation for each subset $s \in \mathcal{S}$. For distribution $P_{0,\epsilon} := P_0 + \epsilon h$ with $\epsilon \in \mathbb{R}$ and $h \in \mathcal{R}$, we define $f_{0,\epsilon,s} = f_{P_{0,\epsilon},s}$ to be its corresponding oracle prediction function with respect to subset s. Let $\dot{V}(f, P_0; h)$ denote the Gâteaux derivative of $P \mapsto V(f, P)$ at P_0 in the direction $h \in \mathcal{R}$, and define the random function $g_{n,s} : o \mapsto \dot{V}(f_{n,s}, P_0; \delta_o - P_0) - \dot{V}(f_{0,s}, P_0; \delta_o - P_0)$, where δ_o is the degenerate distribution on $\{o\}$. Consider the following set of deterministic [(A1)–(A4)] and stochastic [(B1)–(B2)] conditions for each subset $s \in \mathcal{S}$:

(A1) (*optimality*) there is some $C > 0$ such that for each sequence $f_1, f_2, \cdots \in \mathcal{F}_s$ with $\|f_j - f_{0,s}\|_{\mathcal{F}_s} \to 0$, there is a J such that for all $j > J$, $|V(f_j, P_0) - V(f_{0,s}, P_0)| \leq C\|f_j - f_{0,s}\|^2_{\mathcal{F}_s}$;

(A2) (*differentiability*) there is some $\delta > 0$ such that for each sequence $\epsilon_1, \epsilon_2, \ldots \in \mathbb{R}$ and $h, h_1, h_2, \ldots \in \mathcal{R}$ satisfying that $\epsilon_j \to 0$ and $\|h_j - h\|_\infty \to 0$, it holds that

$$\sup_{f \in \mathcal{F}_s : \|f - f_{0,s}\|_{\mathcal{F}_s} < \delta} \left| \frac{V(f, P_0 + \epsilon_j h_j) - V(f, P_0)}{\epsilon_j} - \dot{V}(f, P_0; h_j) \right| \longrightarrow 0 ;$$

(A3) (*optimizer continuity*) $\|f_{0,\epsilon,s} - f_{0,s}\|_{\mathcal{F}_s} = o(\epsilon)$ for each $h \in \mathcal{R}$;

(A4) (*derivative continuity*) $f \mapsto \dot{V}(f, P_0; h)$ is continuous at $f_{0,s}$ relative to $\mathcal{F}_s$ for each $h \in \mathcal{R}$;

(B1) (*minimum rate of convergence*) $\|f_{n,s} - f_{0,s}\|_{\mathcal{F}_s} = o_P(n^{-1/4})$;

(B2) (*weak consistency*) $E_0[\int \{g_{n,s}(o)\}^2 dP_0(o)] = o_P(1)$;

The Gâteaux derivative $\dot{V}$ is provided in Williamson et al. (2020b) for several common measures of predictiveness, including classification accuracy, AUC, and R^2. Assuming conditions (A1)–(A4) and (B1)–(B2) hold for every subset in $\mathcal{S}$, v_n is an asymptotically linear estimator of v_0 with influence function $\dot{V}_0 : o \mapsto [\dot{V}(f_{0,\emptyset}, P_0; \delta_o - P_0), \ldots, \dot{V}(f_{0,N}, P_0; \delta_o - P_0)]^\top$ by Theorem 2 in Williamson et al. (2020b). Finally, we introduce a condition that specifies the number of subsets to sample:

(C1) (*minimum number of subsets*) For $\gamma > 0$ and sequence $\gamma_1, \gamma_2, \ldots \in \mathbb{R}^+$ satisfying that $|\gamma_j - \gamma| \to 0$, $m = \gamma_n n$.

For convenience, we define several objects that simplify the notation in our main result below. Set $A := Z^\top W Z$, where Z is the stack of vectors $z(s)$ for all $s \in \mathcal{S}$, and define $C := A^{-1} G (G^\top A^{-1} G)^{-1}$. Let the QR decomposition of $G^\top$ be

$$G^\top = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix},$$

where R is an upper-triangular matrix. We define the functions

$$\phi_{0,1}(O) = A^{-1} Z^\top \sqrt{W} \dot{V}_0(O) \text{ and}$$
$$\phi_{0,2}(S; v_0) = -U_2 V^{-1} \left[z(S)^\top \psi_{0,0} - v_{0,S} \right] U_2^\top z(S),$$

where $V = U_2^\top Z^\top W Z U_2$. Assuming all of the aforementioned conditions hold, then $\psi_{m,n}$ is a consistent and an asymptotically normal estimator of $\psi_{0,0}$.

Theorem 1. *If the collection of conditions implied by (A1)–(A4) and (B1)–(B2) hold for every subset in $\mathcal{S}$ and condition (C1) holds, then $\psi_{m,n}$ has the asymptotic distribution*

$$\sqrt{n}(\psi_{m,n} - \psi_{0,0}) \to_d N(0, \Sigma_0),$$

where $\Sigma_0 := \mathrm{Cov}_{P_0}(\phi_{0,1}(O)) + \gamma^{-1} \mathrm{Cov}_{Q_0}(\phi_{0,2}(S; v_0))$.

To construct Wald-based confidence intervals (CIs) for $\psi_{0,0}$, we estimate the asymptotic covariance Σ_0 by plugging in consistent estimators of each component. That is, we use consistent estimators A_m, Z_m, and W_m of A, Z, and W, respectively. Note that the estimators and CIs may be constructed using only the sampled subsets. If $\psi_{0,0,j} = 0$ for any j, then the contribution from sampling observations to the asymptotic covariance term corresponding to index j will be zero, leading to some additional complications. We discuss this case further in the next section.

Conditions (A1)–(A4) are required to control the contribution from estimating $f_{0,s}$ for each $s \in \mathcal{S}$. Williamson et al. (2020b) show that these conditions are satisfied for R^2, deviance, accuracy, and AUC. Conditions (B1)–(B2) place restrictions on the class of estimators of $f_{0,s}$ that we may consider. While condition (B1) holds for many estimators (e.g., generalized additive models (Hastie and Tibshirani, 1990)), we show in Section 5 that this condition may only need to be approximately satisfied. Condition (B2) is implied by a form of consistency of $f_{n,s}$.

Finally, condition (C1) is necessary to control the contribution from having had to estimate Q_0. Because $\psi_{0,n}$ is an asymptotically efficient estimator of $\psi_{0,0}$, this condition implies that sampling $m = \Theta(n)$ subsets is asymptotically optimal, up to a constant factor proportional to γ^{-1}. Intuitively, this is because there is an irremovable error contribution from having sampled n training observations. As such, we simply need to sample enough subsets for the second error term in (9) to be on the same order as the first term.

Moreover, because the distribution Q_0 places the heaviest weight on subset sizes at the extremes (closest in size to the empty set and full set), we do not need to estimate a large number of unique prediction functions in practice. To our knowledge, this is the first result that delineates the number of feature subsets to sample for constructing an asymptotically normal estimator of Shapley values.

3.3. Testing the null SPVIM hypothesis

We now use Theorem 1 to construct a test for the null hypothesis that a variable is not important, i.e., $\psi_{0,0,j} = 0$ for some j. When a variable X_j has null importance, the true value $\psi_{0,0,j}$ is at the boundary of the parameter space, and the contribution to the asymptotic variance from sampling observations in Theorem 1 is zero. This may cause difficulties in hypothesis testing: as the number of sampled subsets grows, the contribution to the asymptotic variance from sampling subsets tends to zero. Thus, in the limit, a hypothesis test based on the estimator of this asymptotic variance proposed in the previous section will fail to appropriately control the type I error.

Instead, we rely on sample-splitting to construct a valid test of the δ-null hypothesis of the jth SPVIM value, i.e., $H_{0,j} : \psi_{0,0,j} \in [0, \delta]$. In our approach, we make use of the fact that $\psi_{0,0,\emptyset}$ may be nonzero for some predictiveness measures (e.g., AUC). Based on one portion of the data, construct estimator $\psi_{m,n,j,+} := \psi_{m,n,j} + \psi_{m,n,\emptyset}$ of $\psi_{0,0,j} + \psi_{0,0,\emptyset}$ and obtain an estimator $\sigma^2_{n,j}$ of the variance $\sigma^2_{0,j} := (\Sigma_0)_{jj}$. Based on the remaining data, obtain an estimator $\psi_{m,n,\emptyset,1}$ of $\psi_{0,0,\emptyset}$ with corresponding variance estimator $\sigma^2_{n,\emptyset}$. Then, we calculate a test statistic $T_n := \frac{(\psi_{m,n,j,+} - \psi_{m,n,\emptyset,1}) - \delta}{\sqrt{n_1^{-1}\sigma^2_{n,j} + 2 * n_2^{-1}\sigma^2_{n,\emptyset}}}$ and its corresponding p-value $p_n := 1 - \Phi(T_n)$, where n_1 and n_2 denote the respective sample sizes of the split dataset and Φ denotes the standard normal cumulative distribution function. We reject H_0 if and only if $p_n < \alpha$ for some prespecified level α. Under conditions (A1)–(A4), (B1)–(B2), and (C1), for any $\alpha \in (0, 1)$, the proposed test is consistent and has type I error equal to α.

4. Local and group variable importance

Until now, we have focused on a global measure of importance by integrating over the entire distribution P_0. For certain settings, we may be interested instead in a local version of variable importance. A simple extension of (1) or (3) allows us to define a local version of variable importance: for a subpopulation $A \subseteq \mathcal{X}$,

$$\psi_{0,0,j}(A) := \frac{1}{p} \sum_{s \in \mathcal{S}} \binom{p-1}{|s|}^{-1} \{V(f_{0,s\cup j}, P_{0|X\in A}) - V(f_{0,s}, P_{0|X\in A})\},$$

where we have simply plugged the conditional distribution $P_{0|X\in A}$ into (1). Taken to the extreme, where the subpopulation A consists only of a single observation, this definition of local feature importance is equivalent to the SHAP values considered by Lundberg and Lee (2017), though here we use an arbitrary measure of predictiveness in place of the conditional expectation. Unfortunately, valid statistical inference on this individual-observation-level importance appears difficult, if not impossible.

In addition, if there is some scientifically meaningful partition of the features, we can extend SPVIM to these feature subgroups. For example, one may group together all measurements from the same medical device. Let the partition of features into groups be denoted $\mathcal{P} := \{s_1, \ldots, s_k\}$ where $s_i \in \mathcal{S}$ and $\bigcup_{i=1}^k s_i = N$, and $s_i \bigcap s_j = \emptyset$ for every (i, j) pair. Then the Shapley-based population variable *group* importance measure may be determined as in (1), where the sum is taken over all subsets in $\mathcal{P}$.

5. Simulation study

In this section, we present simulation results validating our statistical inference procedure for SPVIM in finite samples. We consider 200 covariates $X \sim N_{200}(0, \Sigma)$. The variance-covariance matrix Σ has diagonal equal to 1 and several correlated features: $Cov(X_1, X_{11}) = 0.7$; $Cov(X_3, X_{12}) = Cov(X_3, X_{13}) = 0.3$; and $Cov(X_5, X_{14}) = 0.05$. The covariance of the remaining feature pairs is zero. Based on these covariates, we observe a continuous outcome $Y \mid X = x \sim N(f(x), 1)$, where

$$\begin{aligned}
f(x) &= \sum_{j\in\{1,3,5\}} f_j(x_j), \\
f_1(x) &= \text{sign}(x), \\
f_3(x) &= (-6)I(x \le -4) + (-4)I(-4 < x \le -2) \\
&\quad + (-2)I(0 \le x < -2) + 2I(2 < x \le 4) \\
&\quad + 4I(x > 4), \text{ and} \\
f_5(x) &= (-1)I(x \le -4 \text{ or } -2 < x \le 0 \text{ or } 2 < x \le 4) \\
&\quad + I(-4 < x \le -2 \text{ or } 0 < x \le 2 \text{ or } x > 4).
\end{aligned}$$

In this data-generating mechanism, the vector (X_1, X_3, X_5) is directly relevant to predicting the outcome, while the vector $(X_{11}, \ldots, X_{14})$ is only related to the outcome through correlation with (X_1, X_3, X_5); the remaining 193 features are pure noise. We generated 1,000 random datasets of size $n \in \{500, 1000, 2000, 3000, 4000\}$. The true SPVIM values for predictiveness defined in terms of R^2 are approximately $(0.19, 0.29, 0.23, 0.04, 0.01, 0.01, 0)$ for the non-noise features, respectively, and zero for the remaining features.

To obtain each $f_{n,s}$ we fit boosted trees (Friedman, 2001) using the Python package `xgboost` (Chen

and Guestrin, 2016) with maximum tree depth equal to one, learning rate equal to 10^{-2}, and ℓ_2-regularization parameter equal to zero. The number of trees varied among $\{50, 100, 250, 500, 1000, \ldots, 3000\}$ and the ℓ_1-regularization parameter varied among $\{10^{-3}, 10^{-2}, 0.1, 1, 5, 10\}$; the combination of these parameters was tuned using five-fold cross-validation to minimize the mean squared error (MSE).

We computed the relevant SPVIM estimator using Algorithm 1, where we sampled $m = 2n$ subsets and estimated predictiveness using five-fold cross-fitting. For comparison, we computed the mean absolute SHAP value (Lundberg and Lee, 2017), where the average was taken over all observations. This allows us to directly evaluate the accuracy of algorithmic VIMs for estimating the population VIMs. We then computed the empirical MSE scaled by n, the empirical coverage of nominal 95% CIs, and the empirical power of our proposed hypothesis test. Finally, we compare the accuracy of our SPVIM estimates and the mean SHAP values in terms of their correlation with the true SPVIM values. All analyses were performed on a computer cluster with 32-core CPU nodes with 64 GB RAM.

We display the results of this experiment in Figure 1. We see that as n increases, the scaled empirical MSE of our estimator decreases to a fixed level — namely, the scaled empirical variance — for each feature. This matches our expectations from Section 3.2: the scaled empirical bias of our proposed estimator should tend to zero with increasing sample size, while the scaled empirical variance tends to the asymptotic variance. We note here that while boosted trees are a popular estimation procedure, they do not necessarily satisfy condition (B1) (see, e.g., Zhang and Yu, 2005). Thus, the convergence observed here provides some empirical evidence that condition (B1) may only need to hold approximately in practice. We also find that the coverage of nominal 95% confidence intervals increases to the nominal level as the sample size increases. Our proposed hypothesis test controls the type I error rate and is consistent: the empirical type I error rate is at the nominal level for null feature X_6, while the empirical power is near one for each of the directly important features. Power tends to be small for the indirectly important features $(X_{11}, \ldots, X_{14})$, especially at small sample sizes; this reflects the fact that the importance of these features is closer to the null hypothesis than the importance of the directly relevant features. Finally, we see that SPVIM estimates are more correlated with the true population importance than SHAP values. We provide the estimated SPVIM and mean absolute SHAP values in the Supplement.

6. Predicting mortality of patients in the intensive care unit

We now analyze data on patients' stays in the ICU from the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database (Silva et al., 2012). We consider 4000 records on several features: five general descriptors collected upon admission to the ICU, and 15 features — including the Glasgow Coma Scale (GCS), blood urea nitrogen (BUN), and heart rate — measured over the course of the first 48 hours after admission to the ICU. The outcome of interest is in-hospital mortality. Rather than use the entire time series, we simplify the analysis by first computing the minimum, average, and maximum value of each of the time-series features used in the simplified acute physiology (SAPS) I or II scores. The SAPS scores are established measures for estimating the mortality risk of ICU patients. We then remove any features that are measured in fewer than 70% of the patients. When combined with the general descriptor variables, a total of 37 extracted features remain. We provide a full list of these extracted features in the Supplement.

We estimate the SPVIM for each variable using AUC to measure predictiveness. For comparison, we also provide the mean absolute SHAP value obtained from Tree SHAP (Lundberg et al., 2020) and Kernel SHAP (Lundberg and Lee, 2017); and the proportion of times a feature was selected across test instances using LIME (Ribeiro et al., 2016). We discuss conditions under which the mean absolute SHAP value is a suitable proxy for the SPVIM in Section 2.2 in the Supplement.

We obtained estimates of each $f_{0,s}$ using two separate procedures. In the first analysis, we maximized the empirical log likelihood using boosted trees with maximum depth equal to four, learning rate equal to 10^{-3}, and a number of estimators in $\{2000, 4000, \ldots, 12000\}$ selected using five-fold cross-validation. In the second analysis, we maximized the empirical log likelihood by fitting ensembles of five dense ReLU neural networks (NNs) with architectures chosen from $\{(37, 25, 25, 20, 10, 1), (37, 25, 20, 1), (37, 25, 20, 20, 1)\}$ using 5-fold cross-validation. The NNs were trained using Adam (Kingma and Ba, 2014) with a maximum of 2000 iterations and with ℓ_2 regularization parameter equal to 0.1. We again used 5-fold cross-fitting to estimate the predictiveness measures for the sampled subsets. Using our procedure, we fit models for only 119 unique subsets and computed SPVIM estimates in two hours for each analysis. LIME had similar computation time (1.7 hours) in the case of NNs, but longer computation time (4 hours) in the case of trees. The computation time of both our procedure and LIME falls between the highly specialized Tree SHAP algorithm, which completed in a few minutes, and the general-purpose Kernel SHAP, which took approximately

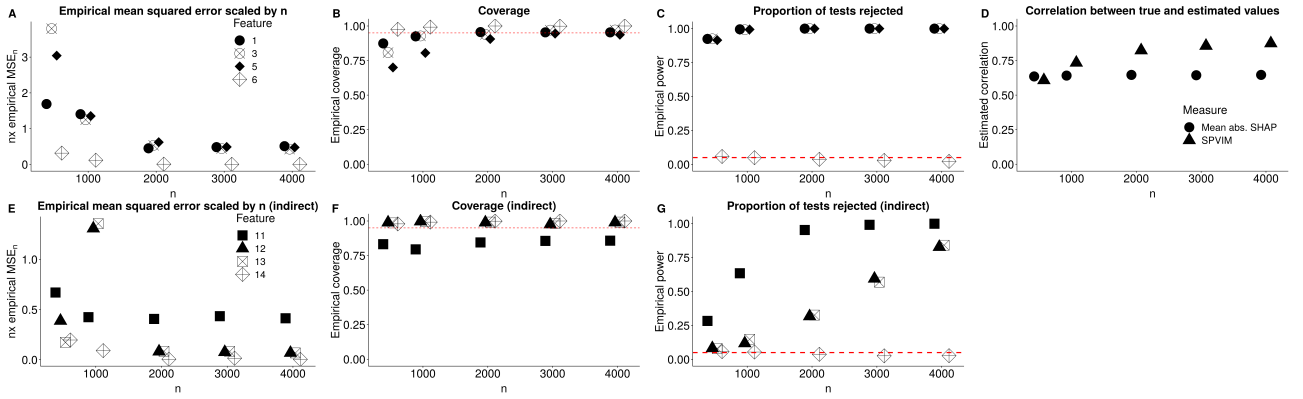

Figure 1. Performance of our statistical inference procedure for estimating the Shapley-based population variable importance (SPVIM) with respect to R^2 using n training observations and $2n$ sampled subsets. (A, E) Empirical MSE for the proposed plug-in estimator scaled by n for $j \in \{1, 3, 5, 6\}$ and $j \in \{11, 12, 13, 14\}$, respectively; (B, F) Empirical coverage of nominal 95% confidence intervals; (C, G) Empirical power of the hypothesis testing procedure for null hypothesis that the jth variable has null importance; (D) Kendall's tau between the true and estimated SPVIM values using our approach versus the mean absolute SHAP value.

20 hours.

In Figure 2, we display the estimates from each VIM and both estimation procedures. We first focus on the SPVIM estimates provided in Panel A. The GCS is estimated to be the most important feature using both trees and NNs, though different summaries of the GCS are most important across the two procedures (mean for trees and max for NNs). This result matches prior knowledge: GCS is used to assess the level of consciousness of patients and is the highest scoring item in the SAPS scores. We find that the confidence intervals for SPVIM are quite wide, which is important information for placing the results in context.

Next, we compare the agreement between rankings calculated based on the fitted boosted trees and NNs for the SPVIM estimates, mean absolute SHAP values (Figure 2 panel B), and LIME (panel C). There is considerably more agreement between the two procedures for the SPVIM estimates than for the SHAP value estimates and LIME proportions. The estimated Kendall's tau between procedures is 0.71 for our SPVIM estimator vs 0.37 for SHAP and 0.39 for LIME. Given the large discrepancies between the algorithmic VIMs, we conclude that they are poor proxies for our population VIM. Instead, one should use a procedure specifically designed to estimate SPVIM.

Finally, we find that the feature rankings within trees or NNs from our SPVIM estimator, SHAP, and LIME are substantively different. One noticeable difference is that SHAP and LIME values for several summary statistics derived from the same measurement (e.g., min, mean, and max GCS) differ widely; this should not occur, since these summary statistics are highly correlated. On the other hand, SPVIM estimates for summary statistics derived from the same measurement tend to be more similar.

7. Discussion

We have proposed a computationally tractable statistical inference procedure for the Shapley population variable importance measure (SPVIM). Methods for estimating SPVIM are complementary to those for estimating algorithmic variable importance. The former helps us understand the underlying data-generating mechanism and can help guide future experiments; the latter helps us interpret a particular fitted model. Here, we define SPVIM with respect to an arbitrary measure of predictiveness, allowing the data analyst to select the most appropriate measure for the task at hand. Since the SPVIM is also defined relative to the population, the target of inference is not affected by the choice of prediction algorithm. We have derived the asymptotic distribution of an SPVIM estimator based on randomly sampled feature subsets, and have used this distribution to construct asymptotically normal point estimates, valid confidence intervals, and hypothesis tests with the correct type I error control. Notably, we determined a minimum number of feature subsets to sample: we show that our estimator only needs to fit prediction models for $m = \Theta(n)$ sampled subsets for its error to be on the same order as an estimator that fits prediction models for all possible subsets.

In this manuscript, we have focused on quantifying the importance of a variable averaging across the entire population. Local importance measures can be obtained by restricting to smaller subpopulations. However, as the subpopulations decrease in size, the uncertainty of our estimates increases. Our asymptotic results do not apply to the most extreme case, the variable importance at the level of a single obser-

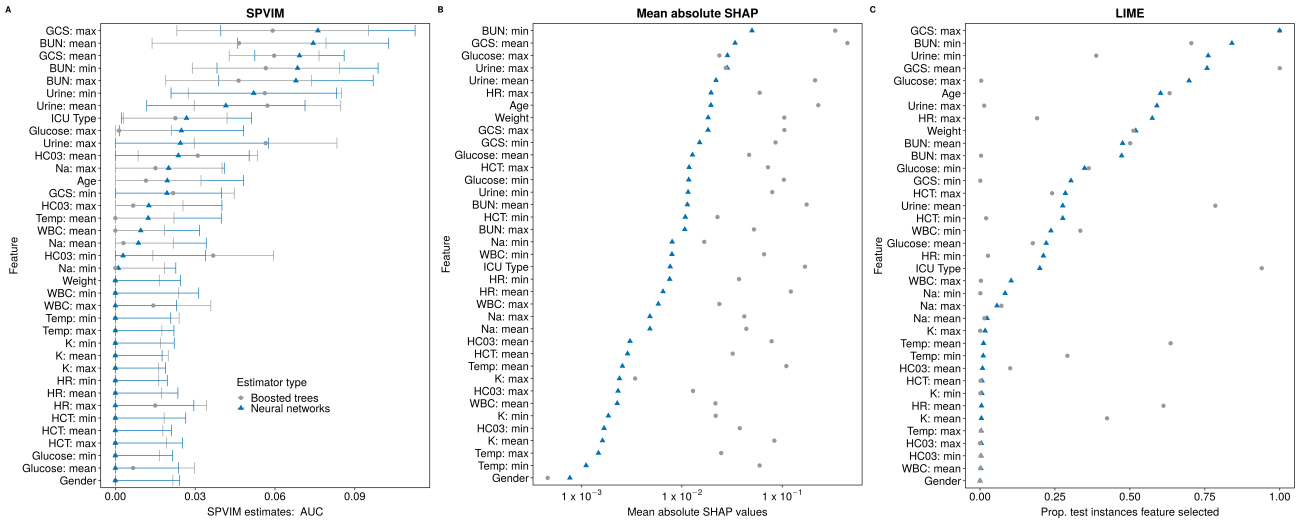

Figure 2. We estimated importance of features for predicting in-hospital death in the ICU using our statistical inference procedure for SPVIM with respect to AUC (A), the mean absolute SHAP value (B), and LIME (C). Gray circles and blue triangles denote estimates from fitting boosted trees and neural networks, respectively. The features are ordered from top to bottom by their point estimate from the neural networks procedure. 95% confidence intervals only appear in (A) since there is no statistical inference procedure for mean absolute SHAP values or LIME.

vation. Nevertheless, this value may be of interest in some tasks. Further work is necessary to define relevant importance measures at the single-observation-level and derive procedures with the desired performance.

Finally, we caution against interpreting SPVIM estimates in a causal manner. SPVIM reflects importance in the oracle prediction model rather than importance in the oracle causal model. In many scientific applications, the importance in the causal model is of ultimate interest. To get causal importance, one may need to employ techniques from causal inference. Recent developments relating prediction models and causal models may also be of use in these cases (Arjovsky et al., 2019).

Acknowledgments

The authors wish to thank Jessica Perry, Noah Simon, the anonymous reviewers, and the meta-reviewer for insightful comments that improved this manuscript. BDW was supported by NIH award F31 AI140836. The opinions expressed in this manuscript are those of the authors and do not necessarily represent the official views of the NIAID or the NIH.

Supplementary Material

Technical details are available in the Supplement. Code is available on Github at `bdwilliamson/spvim_supplementary`.

References

M Arjovsky, L Bottou, I Gulrajani, and D Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

S Boyd and L Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

L Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

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

A Charnes, B Golany, M Keane, and J Rousseau. Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations. In JK Sengupta and GK Kadekodi, editors, *Econometrics of Planning and Efficiency*, pages 123–133. Springer, 1988.

T Chen and C Guestrin. XGBoost: A Scalable Tree Boosting System. *arXiv:1603.02754*, 2016.

I Covert, S Lundberg, and SI Lee. Understanding global feature contributions through additive importance measures. *arXiv*, 2020. https://arxiv.org/abs/2004.00668.

AJ Dunning. A model for immunological correlates of protection. *Statistics in Medicine*, 25(9):1485–1497, 2006.

J Feng, BD Williamson, M Carone, and N Simon. Nonparametric variable importance using an augmented neural network with multi-task learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1495–1504, 2018.

JH Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

DG Garson. Interpreting neural network connection weights. *Artificial Intelligence Expert*, 1991.

U Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

TJ Hastie and RJ Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

D Kingma and J Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

SM Lundberg and S-I Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

SM Lundberg, G Erion, H Chen, A DeGrave, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 2522–5839, 2020.

WJ Murdoch, C Singh, K Kumbier, R Abbasi-Asl, and B Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv:1901.04592*, 2019.

LL Nathans, FL Oswald, and K Nimon. Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, 17 (9), 2012.

AB Owen and C Prieur. On Shapley value for measuring importance of dependent units. *SIAM/ASA Journal on Uncertainty Quantification*, 5, 2017. doi: 10.1137/16M1097717.

MT Ribeiro, S Singh, and C Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

LS Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

I Silva, G Moody, DJ Scott, LA Celi, and RG Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC), 2012*. IEEE, 2012.

E Štrumbelj and I Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

BD Williamson, PB Gilbert, M Carone, and N Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, (to appear), 2020a.

BD Williamson, PB Gilbert, N Simon, and M Carone. A unified approach for inference on algorithm-agnostic variable importance. *arXiv*, 2020b. https://arxiv.org/abs/2004.03683.

T Zhang and B Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4): 1538–1579, 2005.