HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Am Stat Assoc. 2023 ; 118(543): 1645–1658. doi:10.1080/01621459.2021.2003200.

A general framework for inference on algorithm-agnostic variable importance

Brian D. Williamson[1], **Peter B. Gilbert**[1,2], **Noah R. Simon**[2], **Marco Carone**[2,1]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center

[2]Department of Biostatistics, University of Washington

Abstract

In many applications, it is of interest to assess the relative contribution of features (or subsets of features) toward the goal of predicting a response — in other words, to gauge the variable importance of features. Most recent work on variable importance assessment has focused on describing the importance of features within the confines of a given prediction algorithm. However, such assessment does not necessarily characterize the prediction potential of features, and may provide a misleading reflection of the intrinsic value of these features. To address this limitation, we propose a general framework for nonparametric inference on interpretable algorithm-agnostic variable importance. We define variable importance as a population-level contrast between the oracle predictiveness of all available features versus all features except those under consideration. We propose a nonparametric efficient estimation procedure that allows the construction of valid confidence intervals, even when machine learning techniques are used. We also outline a valid strategy for testing the null importance hypothesis. Through simulations, we show that our proposal has good operating characteristics, and we illustrate its use with data from a study of an antibody against HIV-1 infection.

Keywords

variable importance; statistical inference; machine learning; targeted learning

1 Introduction

In many scientific problems, it is of interest to assess the contribution of features toward the objective of predicting a response, a notion that has been referred to as variable importance. Various approaches for quantifying variable importance have been proposed in the literature. In recent applications, variable importance has often been taken to reflect the extent to which a given algorithm makes use of particular features in rendering predictions (Breiman, 2001; Lundberg and Lee, 2017; Fisher et al., 2018; Murdoch et al., 2019). In this case, the

Software and supplementary material

We implement the methods discussed above in the R package vimp and the Python package vimpy, both available on CRAN and PyPI, respectively. Additional technical details are available in the supplementary material. All results may be reproduced using code available on GitHub at https://github.com/bdwilliamson/vimp_supplementary. The data from Section 6 are available at https://github.com/benkeser/vrc01/tree/1.0.

goal is thus to characterize a fixed algorithm. While this notion of variable importance can help provide greater transparency to otherwise opaque black-box prediction tools (Guidotti et al., 2018; Murdoch et al., 2019), it does not quantify the algorithm-agnostic relevance of features for the sake of prediction. Thus, a feature that holds great value for prediction may be deemed unimportant simply because it plays a minimal role in the given algorithm. This motivates the consideration of approaches in which the focus is instead on measuring the population-level predictiveness potential of features, which we can refer to as *intrinsic* variable importance. By definition, any measure of intrinsic variable importance should not involve the external specification of a particular prediction algorithm.

Traditionally, intrinsic variable importance has been considered in the context of simple population models (e.g., linear models) (see, e.g., Grömping, 2006; Nathans et al., 2012). For such models, both the prediction algorithm and the associated variable importance measure (VIM) are easy to compute from model outputs and straightforward to interpret. Common VIMs based on simple models include, for example, the difference in R^2 and deviance values based on (generalized) linear models (Nelder and Wedderburn, 1972; Grömping, 2006). However, overly simplistic models can lead to misleading estimates of intrinsic variable importance with little population relevance. In an effort to improve prediction performance, complex prediction algorithms, including machine learning tools, have been used as a substitute for algorithms resulting from simple population models. Many variable importance measures have been proposed for specific algorithms (see, e.g., reviews of the literature in Wei et al., 2015, Fisher et al., 2018, and Murdoch et al., 2019), with a particularly rich literature on variable importance for random forests (see, e.g., Breiman, 2001; Strobl et al., 2007; Ishwaran, 2007; Grömping, 2009) and neural networks (see, e.g., Garson, 1991; Bach et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017). Several recent proposals aim to describe a broad class of fixed algorithms (LeDell et al., 2015; Ribeiro et al., 2016; Benkeser et al., 2018; Lundberg and Lee, 2017; Aas et al., 2019). However, while some measures have been recently described for algorithm-independent variable importance (see, e.g., van der Laan, 2006; Lei et al., 2017; Williamson et al., 2020), there has been limited work on developing broad frameworks for algorithm-independent variable importance with corresponding theory for inference using machine learning tools.

In this article, we seek to circumvent the limitations of model-based approaches to assessing intrinsic variable importance. We provide a unified nonparametric approach to formulate variable importance as a model-agnostic population parameter, that is, a summary of the true but unknown data-generating mechanism. The VIMs we consider are defined as a contrast between the predictiveness of the best possible prediction function based on all available features versus all features except those under consideration. We allow predictiveness to be defined arbitrarily as relevant and appropriate for the task at hand, as we illustrate in several examples. In this framework, once a measure of predictiveness has been selected, estimation of VIM values from data can be carried out similarly as for any other statistical parameter of interest. This task involves estimation of oracle prediction functions based on all the features or various subsets of features, and the use of machine learning algorithms is advantageous for maximizing prediction performance for this purpose. Because we consider variable importance as a summary of the data-generating mechanism rather than a property of any

particular prediction algorithm, its definition and implementation does not hinge on the use of any particular prediction algorithm. This perspective contrasts with the model-based approach, where the probabilistic population-level mechanism that generates data and the algorithm that makes predictions based on data are usually entangled.

In Williamson et al. (2020), we focused on an application of the proposed framework to infer about a model-agnostic R^2-based variable importance, for which we described a nonparametric efficient estimator. We also presented the construction of valid confidence intervals and hypothesis tests for features with some importance but found it challenging to assess features with zero-importance. Here, we propose a general framework to study general predictiveness measures and propose a valid strategy for hypothesis testing. Our framework allows us to tackle cases involving complex predictiveness measures (e.g., defined in terms of counterfactual outcomes or involving missing data). It can be used to describe the importance of groups of variables as easily as individual variables. Our framework formally incorporates the use of machine learning tools to construct efficient estimators and perform valid statistical inference. We emphasize that the latter is especially important if high-impact decisions will be made on the basis of the resulting VIM estimates.

This article is organized as follows. In Section 2, we define variable importance as a contrast in population-level oracle predictiveness and provide simple examples. In Section 3, we construct an asymptotically efficient VIM estimator for a large class of measures using flexibly estimated prediction algorithms (e.g., predictive models constructed via machine learning methods) and provide a valid test of the zero-importance null hypothesis. These results allow us to analyze nonparametric extensions of common measures, including the area under the receiver operating characteristic curve (AUC) and classification accuracy. In Section 4, we explore an extension to deal with more complex predictiveness measures. In Section 5, we illustrate the use of the proposed approach in numerical experiments and detail its operating characteristics. Finally, we study the importance of various HIV-1 viral protein sequence features in predicting resistance to neutralization by an antibody in Section 6, and provide concluding remarks in Section 7. All technical details as well as results from additional simulation studies and data analyses can be found in the Supplementary Material.

2 Variable importance

2.1 Data structure and notation

Suppose that observations $Z_1, \ldots, Z_n$ are drawn independently from a data-generating distribution P_0 known only to belong to a rich (nonparametric) class $\mathcal{M}$ of distributions. For concreteness, suppose that $Z_i = (X_i, Y_i)$, where $X_i = (X_{i1}, \ldots, X_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$ is a covariate vector and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is the outcome. Here, $\mathcal{X}$ and $\mathcal{Y}$ denote the sample spaces of X and Y, respectively. Below, we will use the shorthand notation E_0 to refer to expectation under P_0.

We denote by $s \subseteq \{1, \ldots, p\}$ the index set of the covariate subgroup of interest, and for any p-dimensional vector w, we refer to the elements of w with index in s and not in s as w_s and w_{-s} respectively. We also denote by $\mathcal{X}_s$ and $\mathcal{X}_{-s}$ the sample space of X_s and X_{-s}, respectively. Finally, we consider a rich class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ endowed with a

norm $\|\cdot\|_{\mathscr{F}}$, and define the subset $\mathscr{F}_s := \{f \in \mathscr{F} : f(u) = f(v)$ for all $u, v \in \mathscr{X}$ satisfying $u_{-s} = v_{-s}\}$ of functions in $\mathscr{F}$ whose evaluation ignores elements of the input x with index in s. In all examples we consider, we will take $\mathscr{F}$ to be essentially unrestricted up to regularity conditions. Common choices include the class of all P_0-square-integrable functions from $\mathscr{X}$ to $\mathscr{Y}$ endowed with $L_2(P_0)$-norm $f \mapsto \|f\|_{2,P_0} := [\int \{f(x)\}^2 dP_0(x)]^{1/2}$, and of all bounded functions from $\mathscr{X}$ to $\mathscr{Y}$ endowed with the supremum norm $f \mapsto \|f\|_{\infty,\mathscr{X}} := \sup_{x \in \mathscr{X}} |f(x)|$.

2.2 Oracle predictiveness and variable importance

We now detail how we define variable importance as a population parameter. Suppose that $V(f, P)$ is a measure of the predictiveness of a given candidate prediction function $f \in \mathscr{F}$ when P is the true data-generating distribution, with large values of $V(f, P)$ implying high predictiveness. Examples of predictiveness measures — including those based on R^2, deviance, the area under the ROC curve, and classification accuracy — are discussed in detail in Section 2.3. If the true data-generating mechanism P_0 were known, a natural candidate prediction function would be any P_0-population maximizer f_0 of predictiveness over the class $\mathscr{F}$:

$$f_0 \in \underset{f \in \mathscr{F}}{\operatorname{argmax}}\, V(f, P_0)\,. \tag{1}$$

This population maximizer can be viewed as the oracle prediction function within $\mathscr{F}$ under P_0 relative to V. In particular, the definition of f_0 depends on the chosen predictiveness measure and on the data-generating mechanism. It can also depend on the choice of function class, although in contexts we consider this is not the case as long as $\mathscr{F}$ is sufficiently rich. It is often true that f_0 is the underlying target of machine learning-based prediction algorithms or a transformation thereof, which facilitates the integration of machine learning tools in the estimation of f_0. The *oracle predictiveness* $V(f_0, P_0)$ provides a measure of total prediction potential under P_0. Similarly, defining the oracle prediction function $f_{0,s}$ that maximizes $V(f, P_0)$ over all $f \in \mathscr{F}_s$, the *residual oracle predictiveness* $V(f_{0,s}, P_0)$ quantifies the remaining prediction potential after exclusion of covariate features with index in s.

We define the *population-level importance* of the variable (or subgroup of variables) X_s relative to the full covariate vector X as the amount of oracle predictiveness lost by excluding X_s from X. In other words, we consider the VIM value defined as

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0)\,. \tag{2}$$

By construction, we note that $\psi_{0,s} \geq 0$. Whether or not the loss in oracle predictiveness is sufficiently large to confer meaningful importance to a given subgroup of covariates depends on context. Once more, we emphasize that the definition of $\psi_{0,s}$ involves the oracle prediction function within $\mathscr{F}$, and if $\mathscr{F}$ is large enough, this definition is agnostic to this choice.

2.3 Examples of predictiveness measures

We now illustrate our definition of variable importance by listing common VIMs that are in this framework. As we will see, the conditional mean $\mu_0 : x \mapsto E_0(Y|X = x)$ plays a prominent role in the examples below. This is convenient since μ_0 is the implicit target of estimation for many standard machine learning algorithms for predictive modeling.

Example 1: R^2—The R^2 predictiveness measure is defined as $V(f, P_0) := 1 - E_0\{Y - f(X)\}^2/\sigma_0^2$, where we set $\sigma_0^2 := E_0\{Y - E_0(Y)\}^2 = E_0[Y - E_0\{\mu_0(X)\}]^2$, the variance of Y under P_0. This measure quantifies the proportion of variability in Y explained by $f(X)$ under P_0. Since μ_0 is the unrestricted minimizer of the mean squared error mapping $f \mapsto E_0\{Y - f(X)\}^2$, the optimizer of $V(f, P_0)$ is given by $f_0 = \mu_0$ as long as $\mu_0 \in \mathcal{F}$.

Example 2: deviance—When Y is binary, the deviance predictiveness measure is defined as

$$V(f, P_0) = 1 - \frac{E_0[Y \log f(X) + (1 - Y)\log\{1 - f(X)\}]}{\pi_0 \log \pi_0 + (1 - \pi_0)\log(1 - \pi_0)},$$

where $\pi_0 := P_0(Y = 1)$ is the marginal success probability of Y under P_0. This measure quantifies in a Kullback-Leibler sense the information gain from using X to predict Y relative to the null model that does not use X at all. Again, because the conditional mean μ_0 is the unconstrained population maximizer of the average log-likelihood, we find the optimizer of $f \mapsto V(f, P_0)$ to be $f_0 = \mu_0$ for any rich enough $\mathcal{F}$. This result similarly holds for a multinomial extension of deviance.

Example 3: classification accuracy—An alternative predictiveness measure in the context of binary outcomes is classification accuracy, defined as $V(f, P_0) = P_0\{Y = f(X)\}$. This measure quantifies how often the prediction $f(X)$ coincides with Y, and is commonly used in classification problems. As shown in the Supplementary Material, the Bayes classifier $b_0 : x \mapsto I\{\mu_0(x) > 1/2\}$ is the unconstrained maximizer of $f \mapsto V(f, P_0)$, and so, $f_0 = b_0$ as long as $b_0 \in \mathcal{F}$.

Example 4: area under the ROC curve—The area under the receiver operating characteristic curve (AUC) is another popular predictiveness measure for use when Y is binary. The AUC corresponding to f is given by $V(f, P_0) = P_0\{f(X_1) < f(X_2)|Y_1 = 0, Y_2 = 1\}$, where (X_1, Y_1) and (X_2, Y_2) represent independent draws from P_0. As shown in the Supplementary Material, the unrestricted maximizer of $f \mapsto V(f, P_0)$ is the population mean μ_0, so that once more $f_0 = \mu_0$ provided $\mu_0 \in \mathcal{F}$.

In all examples above, the unrestricted oracle prediction function f_0 equals or is a simple transformation of the conditional mean function μ_0. The unrestricted oracle prediction function $f_{0,s}$ based on all covariates except those with index in s is obtained similarly but with μ_0 replaced by $\mu_{0,s} : x \mapsto E_0(Y|X_{-s} = x_{-s})$.

3 Estimation and inference

3.1 Plug-in estimation

In our framework, the variable importance of X_s relative to X under P_0, denoted $\psi_{0,s}$, is a population parameter. Thus, assessing variable importance reduces to the task of inferring about $\psi_{0,s}$ from the available data. More formally, our goal is to construct a nonparametric (asymptotically) efficient estimator of $\psi_{0,s}$ using independent observations $Z_1, \ldots, Z_n$ from P_0. Definition (2) suggests considering the plug-in estimator

$$\psi_{n,s} := V(f_n, P_n) - V(f_{n,s}, P_n), \tag{3}$$

where P_n is the empirical distribution based on $Z_1, \ldots, Z_n$, and f_n and $f_{n,s}$ are estimators of the population optimizers f_0 and $f_{0,s}$, respectively. Often, f_n and $f_{n,s}$ are obtained by building a predictive model for outcome Y using all features in X or only those features in X_{-s}, respectively — this might be done, for example, using tree-based methods, deep learning, or other machine learning algorithms, including tuning via cross-validation. Using flexible learning techniques to construct f_n and $f_{n,s}$ minimizes the risk of systematic bias due to model misspecification.

As an illustration of the form of the resulting plug-in estimates, we note that, in the case of classification accuracy (Example 3), the VIM estimate is given by $\psi_{n,s} = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = f_n(X_i)\} - \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = f_{n,s}(X_i)\}$, where f_n and $f_{n,s}$ are estimates of the oracle prediction functions f_0 and $f_{0,s}$, respectively. Sensible estimates of f_0 and $f_{0,s}$ are given by

$$f_n : x \mapsto I\{\mu_n(x) > 0.5\} \quad \text{and} \quad f_{n,s} : x \mapsto I\{\mu_{n,s}(x) > 0.5\},$$

where μ_n and $\mu_{n,s}$ are estimates of conditional mean functions μ_0 and $\mu_{0,s}$, respectively. We provide the explicit form of $\psi_{n,s}$ for all examples in the Supplementary Material.

The simplicity of the plug-in construction makes it particularly appealing. However, the literature on semiparametric inference and targeted learning suggests that such naively constructed plug-in estimators may fail to even be consistent at rate $n^{-1/2}$, let alone efficient, if they involve nuisance functions — in this case, f_0 and $f_{0,s}$ — that are flexibly estimated. This phenomenon is due to the fact that excess bias is often inherited by the plug-in estimator from the nuisance estimators. Generally, this fact would motivate the use of debiasing procedures, such as the one-step correction or targeted maximum likelihood estimation (see, e.g., Pfanzagl, 1982; van der Laan and Rose, 2011). However, in Williamson et al. (2020) we noted the intriguing fact that the plug-in estimator of the R^2 VIM did not require debiasing, being itself already efficient. Below, we show that the same holds true for a large class of VIMs. These plug-in estimators therefore benefit from a combination of simplicity and statistical optimality.

3.2 Large-sample properties

We now study conditions under which $\psi_{n,s}$ is an asymptotically linear and nonparametric efficient estimator of the VIM value $\psi_{0,s}$, and we describe how to conduct valid inference on $\psi_{0,s}$. Below, we explicitly focus on inference for the oracle predictiveness value $v_0 := V(f_0, P_0)$ based on the plug-in estimator $v_n := V(f_n, P_n)$, since results can readily be extended to the residual oracle predictiveness value $v_{0,s} := V(f_{0,s}, P_0)$ and thus to the VIM value $\psi_{0,s}$. The behavior of v_n can be studied by first decomposing

$$v_n - v_0 = \{V(f_0, P_n) - V(f_0, P_0)\} + \{V(f_n, P_0) - V(f_0, P_0)\} + r_n, \tag{4}$$

where $r_n := [\{V(f_n, P_n) - V(f_n, P_0)\} - \{V(f_0, P_n) - V(f_0, P_0)\}]$. Each term on the right-hand side of (4) can be studied separately to determine the large-sample properties of v_n. The first term is the contribution from having had to estimate the second argument value P_0. The third term is a difference-of-differences remainder term that can be expected to tend to zero in probability at a rate faster than $n^{-1/2}$ under some conditions. We must pay particular attention to the second term, which represents the contribution from having had to estimate the first argument value f_0. A priori, we may expect this term to dominate since the rate at which $f_n - f_0$ tends to zero (in suitable norms) is generally slower than $n^{-1/2}$ when flexible learning techniques are used. However, because f_0 is a maximizer of $f \mapsto V(f, P_0)$ over $\mathscr{F}$, we may reasonably expect that

$$\frac{d}{d\epsilon} V(f_{0,\epsilon}, P_0)\big|_{\epsilon = 0} = 0$$

for any smooth path $\{f_{0,\epsilon} : -\infty < \epsilon < +\infty\} \subset \mathscr{F}$ through f_0 at $\epsilon = 0$, and thus that there is no first-order contribution of $V(f_n, P_0) - V(f_0, P_0)$ to the behavior of $v_n - v_0$. Under regularity conditions, this indeed turns out to be the case, and thus, if $f_n - f_0$ does not tend to zero too slowly, the second term will be asymptotically negligible.

Our first result will make use of several conditions requiring additional notation. Below, we define the linear space $\mathscr{R} := \{c(P_1 - P_2) : c \in [0, \infty), P_1, P_2 \in \mathscr{M}\}$ of finite signed measures generated by $\mathscr{M}$. For any $R \in \mathscr{R}$, say $R = c(P_1 - P_2)$, we refer to the supremum norm $\|R\|_\infty := c \cdot \sup_z |F_1(z) - F_2(z)|$, where F_1 and F_2 are the distribution functions corresponding to P_1 and P_2, respectively. Furthermore, we denote by $\dot{V}(f, P_0; h)$ the Gâteaux derivative of $P \mapsto V(f, P)$ at P_0 in the direction $h \in \mathscr{R}$, and define the random function $g_n : z \mapsto \dot{V}(f_n, P_0; \delta_z - P_0) - \dot{V}(f_0, P_0; \delta_z - P_0)$, where δ_z is the degenerate distribution on $\{z\}$. For any $P \in \mathscr{M}$, we also denote by f_P any P-population maximizer of $f \mapsto V(f, P)$ over $\mathscr{F}$. Finally, we define the following sets of conditions, classified as being either deterministic (A) or stochastic (B) in nature:

(A1) (*optimality*) there exists some constant $C > 0$ such that, for each sequence $f_1, f_2, \cdots \in \mathscr{F}$ such that $\|f_j - f_0\|_{\mathscr{F}} \to 0$, $|V(f_j, P_0) - V(f_0, P_0)| \leq C\|f_j - f_0\|_{\mathscr{F}}^2$ for each j large enough;

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

$$\sup_{f \in \mathscr{F}: \|f - f_0\|_{\mathscr{F}} < \delta} \left| \frac{V(f, P_0 + \epsilon_j h_j) - V(f, P_0)}{\epsilon_j} - \dot{V}(f, P_0; h_j) \right| \to 0 ;$$

(A3) (*continuity of optimization*) $\|f_{P_0 + \epsilon h} - f_0\|_{\mathscr{F}} = O(\epsilon)$ for each $h \in \mathscr{R}$;

(A4) (*continuity of derivative*) $f \mapsto \dot{V}(f, P_0; h)$ is continuous at f_0 relative to $\| \cdot \|_{\mathscr{F}}$ for each $h \in \mathscr{R}$;

(B1) (*minimum rate of convergence*) $\|f_n - f_0\|_{\mathscr{F}} = o_P(n^{-1/4})$;

(B2) (*weak consistency*) $\int \{g_n(z)\}^2 dP_0(z) = o_P(1)$;

(B3) (*limited complexity*) there exists some P_0-Donsker class $\mathscr{G}_0$ such that $P_0(g_n \in \mathscr{G}_0) \to 1$.

Theorem 1. *If conditions (A1)–(A2) and (B1)–(B3) hold, then v_n is an asymptotically linear estimator of v_0 with influence function equal to $\phi_0 : z \mapsto \dot{V}(f_0, P_0; \delta_z - P_0)$, that is,*

$$v_n - v_0 = \frac{1}{n} \sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + o_P(n^{-1/2})$$

under sampling from P_0. If conditions (A3)–(A4) also hold, then ϕ_0 coincides with the nonparametric efficient influence function (EIF) of $P \mapsto V(f_P, P)$ at P_0, and so, v_n is nonparametric efficient.

This result implies, in particular, that the plug-in estimator v_n of v_0 is often consistent as well as asymptotically normal and efficient. A similar theorem applies to the study of the estimator $v_{n,s} := V(f_{n,s}, P_n)$ of residual oracle predictiveness $v_{0,s}$ upon replacing instances of f_n, f_0 and $\mathscr{F}$ by $f_{n,s}$, $f_{0,s}$ and $\mathscr{F}_s$ in the conditions above, and denoting the resulting influence function by $\phi_{0,s}$. Thus, under the collection of all such conditions, the estimator $\psi_{n,s}$ of the VIM value $\psi_{0,s}$ is asymptotically linear with influence function $\varphi_{0,s} : z \mapsto \dot{V}(f_0, P_0; \delta_z - P_0) - \dot{V}(f_{0,s}, P_0; \delta_z - P_0)$ and nonparametric efficient. If $\psi_{0,s} > 0$ and $0 < \tau_{0,s}^2 := E_0\{\varphi_{0,s}^2(Z)\} < \infty$, this suggests that the asymptotic variance of $n^{1/2}(\psi_{n,s} - \psi_{0,s})$ can be estimated by

$$\tau_{n,s}^2 := \frac{1}{n} \sum_{i=1}^{n} \left[\dot{V}(f_n, P_n; \delta_{Z_i} - P_n) - \dot{V}(f_{n,s}, P_n; \delta_{Z_i} - P_n) \right]^2 ,$$

and that $(\psi_{n,s} - z_{1-\alpha/2}\tau_{n,s}n^{-1/2}, \psi_{n,s} + z_{1-\alpha/2}\tau_{n,s}n^{-1/2})$ is an interval for $\psi_{0,s}$ with asymptotic coverage $1-\alpha$, where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)^{th}$ quantile of the standard normal distribution. This procedure is summarized in Algorithm 1. We discuss settings in which $\psi_{0,s} = 0$ (and therefore $\tau_{0,s}^2 = 0$) in Section 3.4.

Algorithm 1

Inference on VIM value $\psi_{0,s}$ (valid in non-null settings)

1: construct estimators f_n of f_0 and $f_{n,s}$ of $f_{0,s}$;

2: construct empirical distribution estimator P_n of P_0;

3: compute estimator $\psi_{n,s} := V(f_n, P_n) - V(f_{n,s}, P_n)$ of $\psi_{0,s}$;

4: compute estimator

$$\tau^2_{n,s} := \frac{1}{n}\sum_{i=1}^{n}\left\{\dot{V}(f_n, P_n; \delta_{Z_i} - P_n) - \dot{V}(f_{n,s}, P_n; \delta_{Z_i} - P_n)\right\}^2$$

of the asymptotic variance $\tau^2_{0,s}$ of $n^{1/2}(\psi_{n,s} - \psi_{0,s})$.

Condition (A1) ensures that there is no first-order contribution that results from estimation of f_0. As indicated above, this condition can generally be established as a consequence of the optimality of f_0. However, in each particular problem, appropriate regularity conditions on P_0 and $\mathscr{F}$ must be determined for this condition to hold. We have provided details for Examples 1–4 in the Supplementary Material, though we summarize our findings here. In Example 1, we have that $|V(f, P_0) - V(f_0, P_0)| = E_0\{f(X) - f_0(X)\}^2/\sigma_0^2$ as long as $\mu_0 \in \mathscr{F}$, and so, condition (A1) holds with $C = 1/\sigma_0^2$ and $\|\cdot\|_{\mathscr{F}}$ taken to be either the $L_2(P_0)$ or supremum norm. In Example 2, provided that all elements of $\mathscr{F}$ are bounded between γ and $1 - \gamma$ for some $\gamma \in (0, 1)$ and that $\mu_0 \in \mathscr{F}$, then condition (A1) holds with $C = \{\gamma \log(1 - \gamma)\}^{-1}$ and $\|\cdot\|_{\mathscr{F}}$ taken to be either the $L_2(P_0)$ or supremum norm. In Example 3, condition (A1) holds for $C = 4\kappa$ and $\|\cdot\|_{\mathscr{F}}$ the supremum norm provided the classification margin condition $P_0\{|\mu_0(X) - 0.5| \leq t\} \leq \kappa t$ holds for some $0 < \kappa < \infty$ and all t small. Similarly, in Example 4, condition (A1) holds for $C = 2\kappa/\{\pi_0(1 - \pi_0)\}$ with $\pi_0 := P_0(Y = 1)$ and $\|\cdot\|_{\mathscr{F}}$ the supremum norm provided the margin condition $P_0\{|\mu_0(X_1) - \mu_0(X_2)| \leq t\} \leq \kappa t$ holds for some $0 < \kappa < \infty$ and all t small, where X_1 and X_2 are independent draws from P_0.

Condition (A2) is a form of locally uniform Hadamard differentiability of $P \mapsto V(f_0, P)$ at P_0 in a neighborhood of f_0. It can be readily verified in Examples 1–4; in fact, in Examples 1–3, this condition holds for any $\delta > 0$. Condition (A3) requires that the optimizer f_P vary smoothly in P around P_0, and is often straightforward to verify when f_P has a closed analytic form. Condition (A4) instead requires that the Hadamard derivative of $P \mapsto V(f, P)$ at P_0 vary smoothly in f around f_0. Condition (B1) requires that f_0 be estimated at a sufficiently fast rate in order for second-order terms to be asymptotically negligible, while condition (B2) states that a particular parameter-specific functional of f_n must tend to the corresponding evaluation of f_0, and is thus implied by consistency of f_n with respect to some norm under which this functional is continuous. Condition (B3) restricts the complexity of the algorithm used to generate f_n. We note that conditions (B1)–(B3) depend not only on the predictiveness measure chosen and on the true data-generating mechanism but also on properties of the estimator of the oracle prediction function.

3.3 Implementation based on cross-fitting

Condition (B3) puts constraints on the complexity of the algorithm used to generate f_n. This condition is prone to violations when flexible machine learning tools are employed, as discussed in Zheng and van der Laan (2011) and Chernozhukov et al. (2018), for example. However, it can be eliminated by dividing the entire dataset into two parts (say, training and testing sets), estimating f_0 using the training data, and then evaluating the predictiveness measure on the test data. This readily extends to K-fold cross-fitting. To construct a cross-fitted estimator in the current context, we begin by randomly partitioning the dataset into K subsets of roughly equal size. Setting aside one such subset, we construct an estimator $f_{k,n}$ of f_0 based on the bulk of the data, and then store $v_{k,n} := V(f_{k,n}, P_{k,n})$, where $P_{k,n}$ is the empirical distribution estimator based on the data set aside. We note that $f_{k,n}$ and $P_{k,n}$ are therefore estimated using non-overlapping subsets of the data. After repeating this operation for each of the K subsets, we finally construct the cross-fitted estimator $v_n^* := \frac{1}{K}\sum_{k=1}^{n} v_{k,n}$ of v_0.

To describe the large-sample behavior of v_n^*, we require an adaptation of the previously defined conditions (B1) and (B2) to the context of cross-fitted estimators. Below, the random function $g_{k,n}$ is defined identically as g_n but with f_n replaced by $f_{k,n}$.

(B1') (*minimum rate of convergence*) $\|f_{k,n} - f_0\|_{\mathcal{F}} = o_P(n^{-1/4})$ for each $k \in \{1, \ldots, K\}$;

(B2') (*weak consistency*) $\int \{g_{k,n}(z)\}^2 dP_0(z) = o_P(1)$ for each $k \in \{1, \ldots, K\}$.

The resulting cross-fit estimator v_n^* enjoys desirable large-sample properties under weaker conditions than those imposed on v_n, as the theorem below states. In particular, condition (B3), which in practice limits the complexity of machine learning tools used to estimate f_0, is no longer required.

Theorem 2. *If conditions (A1)–(A2) and (B1')–(B2') hold, then v_n^* is an asymptotically linear estimator of v_0 with influence function equal to $\phi_0 : x \mapsto \dot{V}(f_0, P_0; \delta_z - P_0)$, that is,*

$$v_n^* - v_0 = \frac{1}{n}\sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + o_P(n^{-1/2})$$

under sampling from P_0. If conditions (A3)–(A4) also hold, then v_n^* is nonparametric efficient.

The cross-fitted construction can be used to obtain an improved estimator $v_{n,s}^*$ of $v_{0,s}$ as well, thereby resulting in a cross-fitted estimator $\psi_{n,s}^* := v_n^* - v_{n,s}^*$ of the VIM value $\psi_{0,s}$. Cross-fitting can also be used to obtain an improved estimator $\tau_{n,s,*}^2$ of the asymptotic variance $\tau_{0,s}^2$. We summarize this construction in Algorithm 2, and provide the explicit form of $\psi_{n,s}^*$ for Examples 1–4 in the Supplementary Material. As before, Theorem 2 readily provides conditions under which $v_{n,s}^*$ is an asymptotically linear and nonparametric efficient estimator of $v_{0,s}$, and so, under which $\psi_{n,s}^*$ is an asymptotically linear and nonparametric efficient estimator of the VIM value $\psi_{0,s}$. Based on these theoretical results as well as numerical

experiments, we recommend this implementation whenever machine learning tools are used to estimate f_0 and $f_{0,s}$.

Algorithm 2

Cross-fitted inference on VIM value $\psi_{0,s}$ (valid in non-null settings)

1: generate $B_n \in \{1, \ldots, K\}^n$ by sampling uniformly from $\{1, \ldots, K\}$ with replacement, and for $j = 1, \ldots, K$, denote by D_j the subset of observations with index in $S_j := \{i : B_{n,i} = j\}$;

2: **for** $k = 1, \ldots, K$ **do**

3: using only data in $\cup_{j \neq k} D_j$, construct estimators $f_{k,n}$ of f_0 and $f_{k,n,s}$ of $f_{0,s}$;

4: using only data in D_k construct empirical distribution estimator $P_{k,n}$ of P_0;

5: with $n_k := \sum_{i=1}^{n} I\{i \in S_k\}$, compute $\psi_{k,n,s} := V(f_{k,n}, P_{k,n}) - V(f_{k,n,s}, P_{k,n})$ and

$$\tau^2_{k,n,s} := \frac{1}{n_k} \sum i \in S_k \{\dot{V}(f_{k,n}, P_{k,n}; \delta_{Z_i} - P_{k,n}) - \dot{V}(f_{k,n,s}, P_{k,n}; \delta_{Z_i} - P_{k,n})\}^2;$$

6: **end for**

7: compute estimator $\psi^*_{n,s} := \frac{1}{K} \sum_{k=1}^{K} \psi_{k,n,s}$ of $\psi_{0,s}$;

8: compute estimator $\tau^2_{n,s,*} := \frac{1}{K} \sum_{k=1}^{K} \tau^2_{k,n,s}$ of the asymptotic variance $\tau^2_{0,s}$ of $n^{1/2}(\psi^*_{n,s} - \psi_{0,s})$.

3.4 Inference under the zero-importance null hypothesis

When $\psi_{0,s} = 0$, in which case the variable group considered has null importance, the influence function of $\psi_{n,s}$ is identically zero. In these cases, even after standardization, $\psi_{n,s}$ generally does not tend to a non-degenerate law. As such, deriving an implementable test of the null hypothesis $\psi_{0,s} = 0$ or a confidence interval valid even when $\psi_{0,s} = 0$ is difficult. In such cases, standard Wald-type confidence intervals and tests based on $\tau^2_{n,s}$ will typically have incorrect coverage or type I error, as illustrated in numerical simulations reported in Williamson et al. (2020) While in parametric settings n-rate inference is possible under this type of degeneracy, this is not expected to be the case in nonparametric models, because the second-order contribution from estimation of f_0 and $f_{0,s}$ will generally have a rate slower than n^{-1}.

We note that, although $\psi_{n,s}$ has degenerate behavior under the null, each of v_n and $v_{n,s}$ are asymptotically linear with non-degenerate (but possibly identical) influence functions. Except for extreme cases in which the entire set of covariates has null predictiveness, we may leverage this fact to circumvent null degeneracy via sample-splitting. Indeed, if v_n and $v_{n,s}$ are constructed using different subsets of the data, then the resulting estimator $\psi_{n,s}$ is asymptotically linear with a non-degenerate influence function even if $\psi_{0,s} = 0$ so that a valid Wald test of the strict null $H_0 : \psi_{0,s} = 0$ versus $H_1 : \psi_{0,s} > 0$ can be constructed using $\psi_{n,s}$ and an estimator of the standard error of $\psi_{n,s}$. Of course, the same holds for the corresponding cross-fitted procedures, as we consider below. We emphasize here that sample-splitting and cross-fitting are distinct operations with distinct goals. Sample-splitting is used to ensure valid inference under the zero-importance null hypothesis, whereas cross-fitting is used to eliminate the need for Donsker class conditions, which otherwise limit how flexible the learning strategies for estimating the oracle prediction functions can be.

Sample-splitting and cross-fitting can be used simultaneously — Figure 1 provides an illustration of the subdivision of a dataset when equal subsets are used for sample-splitting and six splits are used for cross-fitting.

In practice, a group of variables may be considered scientifically unimportant even when $\psi_{0,s}$ is nonzero but small, yet such grouping would be deemed statistically significant in large enough samples. For this reason, given a threshold $\beta > 0$, it may be more scientifically appropriate to consider testing the β-null $H_0 : \psi_{0,s} \in [0, \beta]$ versus its complement alternative $H_1 : \psi_{0,s} > \beta$. The β-null approaches the strict null as β decreases to 0. The idea of sample-splitting also allows us to tackle β-null testing. Suppose that mutually exclusive portions of the dataset, say of respective sizes n—n_s and n_s, are used to construct v_n^* and $v_{n,s}^*$. Suppose further that η_n^2 and $\eta_{n,s}^2$ are consistent estimators of $\eta_0^2 := E_0\{\phi_0(Z)\}^2$ and $\eta_{0,s}^2 := E_0\{\phi_{0,s}(Z)\}^2$, respectively. Then, provided $v_0 > 0$, we may consider rejecting the β-null hypothesis H_0 in favor of its complement H_1 if and only if

$$t_n := \omega_{n,s}^{-1/2} (v_n^* - v_{n,s}^* - \beta) > z_{1-\alpha}, \tag{5}$$

where $\omega_{n,s} := \eta_n^2/(n - n_s) + \eta_{n,s}^2/n_s$ and $z_{1-\alpha}$ is the $(1 - \alpha)^{\text{th}}$ quantile of the standard normal distribution. The implementation of the resulting test, including computation of the corresponding p-value, is summarized in Algorithm 3. Its validity is guaranteed under conditions of Theorem 2 directly applied on the split used to estimate v_0 and modified appropriately (by replacing instances of f_n, f_0 and $\mathcal{F}$ by $f_{n,s}$, $f_{0,s}$ and $\mathcal{F}_s$ in all conditions) to the split used to estimate $v_{0,s}$. We note that, although there is no degeneracy under the β-null whenever $\psi_{0,s} \in (0, \beta)$, sample-splitting is still required for proper type I error control since the strict null $\psi_{0,s} = 0$ is contained in the β-null and must therefore be guarded against. We emphasize here that the use of distinct subsets of the data is critical for constructing v_n^* and $v_{n,s}^*$. If instead $\psi_{0,s} = 0$ and v_n^* and $v_{n,s}^*$ were constructed using the same data, the behavior of any testing procedure based on an estimator $\kappa_{n,s}$ of the standard error of $\psi_{n,s}$ would depend on the relative rates of convergence to zero of both $\psi_{n,s}$ and $\kappa_{n,s}$. In particular, this would lead to either uncontrolled type I error or type I error tending to zero depending on the procedures used to obtain f_n and $f_{n,s}$. Inference based on a cross-fitted version of this sample-split procedure is described in Algorithm 3.

The above testing procedure can be readily inverted to yield a one-sided confidence interval for $\psi_{0,s}$. Specifically, under regularity conditions and provided $v_0 > 0$ the random interval $(v_n^* - v_{n,s}^* - z_{1-\alpha}\omega_{n,s}^{1/2}, +\infty)$ contains $\psi_{0,s}$ with probability no less than $1 - \alpha$ asymptotically, even when $\psi_{0,s} = 0$. Then, rejecting the null hypothesis H_0 is equivalent to verifying that zero is contained in this one-sided interval. A two-sided confidence interval is instead given by $(v_n^* - v_{n,s}^* - z_{1-\alpha/2}\omega_{n,s}^{1/2}, v_n^* - v_{n,s}^* + z_{1-\alpha/2}\omega_{n,s}^{1/2})$. While the latter interval has the advantage of giving both a lower and upper bound on possible values for $\psi_{0,s}$ supported by the data, using it for testing purposes necessarily results in a reduction in power since the null value of $\psi_{0,s}$ is at the edge of the parameter space.

4 Extensions to more complex settings

In all examples studied thus far, the primary role P plays in $V(f, P)$ is to indicate the population with respect to which a particular measure of prediction performance should be averaged. In these

Algorithm 3

Sample-split, cross-fitted inference on VIM value $\psi_{0,s}$

1: generate $B_n \in \{1, \ldots, 2K\}^n$ by sampling uniformly from $\{1, \ldots, 2K\}$ with replacement, and for $j = 1, \ldots, 2K$, denote by D_j the set of observations with index in $S_j := \{i : B_{n,i} = j\}$ and $n_j := |D_j|$;

2: **for** $k = 1, \ldots, 2K$ **do**

3: using only data in $\cup_{j \neq k} D_j$, construct estimators $f_{k,n}$ of f_0 and $f_{k,n,s}$ of $f_{0,s}$;

4: using only data in D_k, construct estimator $P_{n,k}$ of P_0;

5: if k is odd, compute $\eta^2_{k,n} := \frac{1}{n_k}\sum_{i \in S_k} \dot{V}(f_{k,n}, P_{k,n}; \delta_{Z_i} - P_{k,n})^2$ and $v_{k,n} := V(f_{k,n}, P_{k,n})$;

6: if k is even, compute $\eta^2_{k,n,s} := \frac{1}{n_k}\sum_{i \in S_k} \dot{V}(f_{k,n,s}, P_{k,n}; \delta_{Z_i} - P_{k,n})^2$ and $v_{k,n,s} := V(f_{k,n,s}, P_{k,n})$;

7: **end for**

8: compute $v^*_n := \frac{1}{K}\sum_{k=1}^K v_{2k-1,n}$, $v^*_{n,s} := \frac{1}{K}\sum_{k=1}^K v_{2k,n,s}$ and estimator $\psi^*_{n,s} := v^*_n - v^*_{n,s}$ of $\psi_{0,s}$;

9: compute $\eta^2_n := \frac{1}{K}\sum_{k=1}^K \eta^2_{2k-1,n}$, $\eta^2_{n,s} := \frac{1}{K}\sum_{k=1}^K \eta^2_{2k,n,s}$ and estimator $\omega_{n,s} := \eta^2_n/(n - n_s) + \eta^2_{n,s}/n_s$ of the variance of $\psi^*_{n,s}$;

10: to test $H_0 : \psi_{0,s} \in [0, \beta]$ vs $H_1 : \psi_{0,s} > \beta$ at level $1-\alpha$, reject H_0 in favor of H_1 iff $p_n := 1 - \Phi(t_n) < \alpha$ with $t_n := \omega^{-1/2}_{n,s}(\psi^*_{n,s} - \beta)$ and Φ the standard normal distribution function.

cases, $P \mapsto V(f, P)$ is well-defined on discrete probability measures and sufficiently smooth so that $V(f_0, P_n) - V(f_0, P_0)$ is in first order a linear estimator in view of the functional delta method. However, there are other examples in which this requirement may not be true. In these examples, $V(f, P)$ involves P in a complex manner beyond some form of averaging, rendering $V(f, P)$ undefined for discrete P, let alone Hadamard differentiable. Complex predictiveness measures often arise when the sampling mechanism precludes from observation the ideal data unit on which a (possibly simpler) predictiveness measure is defined, and identification formulas must therefore be established to express predictiveness in terms of the observed data-generating distribution.

As a concrete illustration, we begin with an example from the causal inference literature. As before, we denote by Y and X the outcome of interest and a covariate vector, respectively. We suppose that larger values of Y correspond to better clinical outcomes, and consider a binary intervention $A \in \{0, 1\}$. A given treatment rule $f : \mathcal{X} \to \{0, 1\}$ for assigning the value of A based on X can be adjudicated, for example, on the basis of the population mean outcome that would arise if everyone in the population were treated according to f. We can consider the ideal data structure to be $\mathbb{Z} := (X, A, Y(0), Y(1)) \sim \mathbb{P}_0$, where for each $a \in \{0, 1\}$, $Y(a)$ denotes the counterfactual outcome corresponding to the intervention that deterministically sets $A = a$. The ideal-data predictiveness of f is then $\mathbb{V}(f, \mathbb{P}_0) := E_{\mathbb{P}_0}\{Y(f(X))\}$. In contrast, the observed data structure is $Z := (X, A, Y) \sim P_0$, and

we must find some observed-data predictiveness measure V such that $V(f, P_0) = \mathbb{V}(f, \mathbb{P}_0)$ to establish identification and proceed with estimation and inference. Defining the outcome regression $Q_P(a, x) := E_P(Y|A = a, X = x)$, it is not difficult to verify that

$$V(f, P) := E_P\left[Q_P(f(X), X)\right]$$

provides a valid identification of $\mathbb{V}(f, \mathbb{P})$ under standard causal identification conditions. We note that this predictiveness measure involves P through more than simple averaging, as the outcome regression Q_P also appears in the definition of $V(f, P)$. Unless the distribution of X is discrete under P, Q_P is ill-defined on the empirical distribution P_n, thus violating conditions (A1) and (A4) defined in Section 3. We also remark that, in this example, the moniker 'prediction function' is not entirely fitting for f, which represents a treatment rule and maps into the treatment (rather than outcome) space. Nevertheless, the proposed framework for variable importance remains applicable, underscoring the fact that it is sufficiently flexible to unify a large swath of variable importance problems. Restrictions imposed on the data structure and on the properties of the prediction function in Section 2 were largely for the sake of concreteness.

The simple plug-in approach described in Section 3 may fail in applications with more complex predictiveness measures. In such cases, we can instead employ a more general strategy based on nonparametric debiasing techniques to make valid inference about $V(f_0, P_0)$. For each $P \in \mathcal{M}$, we denote by f_P any optimizer of $f \mapsto V(f, P)$ over $\mathcal{F}$, and define the parameter mapping $V^* : P \mapsto V(f_P, P)$ so that v_0 can be expressed as $V^*(P_0)$. If $\hat{P}_n \in \mathcal{M}$ is an estimator of P_0, the plug-in estimator $V^*(\hat{P}_n)$ generally fails to be asymptotically linear unless $\hat{P}_n$ was purposefully constructed to ensure that it is indeed so. This happens because the plug-in estimator $V^*(\hat{P}_n)$ generally suffers from excessive bias whenever flexible learning techniques have been used, for example, because $V^*(P_0)$ involves local features of P_0 (e.g., the conditional mean or density function) — see Pfanzagl (1982) and van der Laan and Rose (2011). This fact renders the use of debiasing approaches necessary. In contrast, the one-step estimator

$$v_{n,OS} := V^*(\hat{P}_n) + \frac{1}{n}\sum_{i=1}^{n} \phi_n(Z_i),$$

where ϕ_n is the nonparametric EIF of V^* at $\hat{P}_n$, is nonparametric efficient (Pfanzagl, 1982) under regularity conditions. Alternatively, the framework of targeted minimum loss-based estimation describes how to convert $\hat{P}_n$ into a revised estimator $\hat{P}_n^*$ such that $V^*(\hat{P}_n^*)$ is itself nonparametric efficient without the need for further debiasing (van der Laan and Rose, 2011). Similarly as in Section 3, cross-fit versions of these debiasing procedures (see, e.g., Zheng and van der Laan, 2011; Chernozhukov et al, 2018) can be used to improve performance when flexible estimation algorithms are used.

The generic approach above relies on deriving the nonparametric EIF of V^*. The definition of V^* involves P in various ways, including through the P-optimal prediction function f_P.

While in our examples f_P has a simple closed-form expression, this may not always be so. This fact can greatly complicate the derivation of the required EIF. However, as we shall see, the optimality of f_P often implies that f_P does not contribute to the nonparametric EIF of $P \mapsto V(f_P, P)$.

Before stating a formal result to this effect, we introduce a regularity condition. Below, $L_2^0(P_0)$ refers to the subset of all functions in $L_2(P_0)$ that have mean zero under P_0.

(A5) There exists a dense subset $\mathscr{H}$ of $L_2^0(P_0)$ such that, for each $h \in \mathscr{H}$ and regular univariate parametric submodel $\{P_{0,\epsilon}\} \subset \mathscr{M}$ through P_0 at $\epsilon = 0$ and with score for ϵ equal to h at $\epsilon = 0$ (see, e.g., Bickel et al, 1998), the following conditions hold, with $f_{0,\epsilon}$ denoting $f_{P_{0,\epsilon}}$:

- **(A5a)** (*second-order property of predictiveness perturbations*) $V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_{0,\epsilon}, P_0) = V(f_0, P_{0,\epsilon}) - V(f_0, P_0) + o(\epsilon)$ holds;
- **(A5b)** (*differentiability*) the mapping $\epsilon \mapsto V(f_{0,\epsilon}, P_0)$ is differentiable in a neighborhood of $\epsilon = 0$;
- **(A5c)** (*richness of function class*) the optimizer $f_{0,\epsilon}$ is in $\mathscr{F}$ for small enough ϵ.

Condition (A5a) essentially requires the pathwise derivative of $P \mapsto V(f, P)$ at P_0 to be insensitive to infinitesimal perturbations of f around f_0. In such case, the difference-in-differences term appearing in the condition can indeed be expected to be second-order in ϵ. Condition (A5b) will generally hold provided the functionals $f \mapsto V(f, P_0)$ and $P \mapsto f_P$ are sufficiently smooth around f_0 and P_0, respectively. Finally, condition (A5c) requires that $\mathscr{F}$ be sufficiently rich around f_0 so that, for a dense collection of paths through P_0, $\mathscr{F}$ contains $f_{0,\epsilon}$ for small enough ϵ.

Theorem 3. *Provided condition (A5) holds, if $P \mapsto V(f_0, P)$ is pathwise differentiable at P_0 relative to the nonparametric model $\mathscr{M}$, then so is $P \mapsto V(f_P, P)$, and the two parameters have the same EIF.*

This theorem indicates that, under a regularity condition, the computation of the nonparametric EIF ϕ_0 can be done treating f_P as fixed at f_0, thereby simplifying considerably this calculation. This fact is also useful because for a fixed prediction function f the parameter $P \mapsto V(f, P)$ will often have already been studied in the literature, thereby circumventing the need for any novel derivation. Armed with this observation, we revisit the motivating example we presented in this section, and also consider an additional example involving missing data.

Example 5: mean outcome under a binary intervention rule

As described above, in this example, the ideal-data parameter of interest, $\mathbb{V}(f, \mathbb{P}) := E_{\mathbb{P}}\{Y(f(X))\}$, can be identified by the observed-data parameter $V(f, P) = E_P\{Q_P(f(X), X)\}$ when the observed data unit consists of $Z = (X, A, Y) \sim P$. The map $f \mapsto V(f, P_0)$ is maximized over the unrestricted class $\mathscr{F}$ by the intervention rule

$f_0 : x \mapsto I\{Q_0(1, x) > Q_0(0, x)\}$, and over its subset $\mathcal{F}_s$ by $f_{0,s} : x \mapsto I\{Q_{0,s}(1, x) > Q_{0,s}(0, x)\}$, where we define $Q_{0,S}$ pointwise as $Q_{0,s}(a, x) := E_0\{Q_0(a, X)|X_{-s} = x_{-s}\}$. Furthermore, the parameter $P \mapsto V(f_0, P)$ is pathwise differentiable at a distribution P_0 if, for example, $Q_0(1, W) - Q_0(0, W) \neq 0$ occurs P_0-almost surely. The nonparametric EIF of $P \mapsto V(f_0, P)$ at P_0 is given by

$$\phi_0 : z \mapsto \frac{I\{a = f_0(x)\}}{g_0(f_0(x), x)}\{y - Q_0(f_0(x), x)\} + Q_0(f_0(x), x) - V(f_0, P_0),$$

where we define the propensity score $g_0(a, x) := P_0(A = a|X = x)$ for each $a \in \{0, 1\}$. Thus, under regularity conditions, the one-step debiased estimator

$$v_{n,OS} := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I\{A_i = f_n(X_i)\}}{g_n(f_n(X_i), X_i)}\{Y_i - Q_n(f_n(X_i), X_i)\} + Q_n(f_n(X_i), X_i)\right]$$

of v_0 is nonparametric efficient, where Q_n and g_n are estimators of Q_0 and g_0, respectively, and f_n is defined pointwise as $f_n(x) := I\{Q_n(1, x) > Q_n(0, x)\}$. The one-step debiased estimator of $v_{0,s}$ is defined similarly, with f_n replaced by any appropriate estimator of $f_{0,s}$, such as $f_{n,s}(x) := I\{Q_{n,s}(1, x) > Q_{n,s}(0, x)\}$ with $Q_{n,s}(a, x)$ obtained by flexibly regressing outcome $Q_n(a, x)$ onto X_{-s} for each $a \in \{0, 1\}$.

Example 6: Classification accuracy under outcome missingness

Suppose the ideal-data structure consists of $\mathbb{Z} := (X, Y) \sim \mathbb{P}$ and the predictiveness measure of interest based on this ideal data structure is the classification accuracy measure, $\mathbb{V}(f, \mathbb{P}) := \mathbb{P}\{Y = f(X)\}$, described in Example 3. Suppose that the outcome Y is subject to missingness, so that the observed data structure is $Z := (X, \Delta, U)$, where Δ is the indicator of having observed the outcome Y, and we have defined $U := \Delta Y$. The observed-data predictiveness measure

$$V(f, P) := E_P[P\{U = f(X)|\Delta = 1, X\}]$$

equals the ideal-data accuracy measure provided that (a) Δ and Y are independent given X, and (b) $P(\Delta = 1|X = x) > 0$ for P-almost every value x. In other words, the provided identification holds provided the outcome is missing at random (relative to X), and there is no subpopulation of patients (as defined by the value of X) for which the outcome can never be observed. Defining $\pi_0(x) := P_0(U = 1|\Delta = 1, X = x)$, the unrestricted optimizers f_0 and $f_{0,S}$ are given pointwise by $f_0(x) = I\{\pi_0(x) > 0.5\}$ and $f_{0,s}(x) = I\{E_0\{\pi_0(X)|X_{-s} = x_{-s}\} > 0.5\}$. Finally, the nonparametric EIF of $P \mapsto V(f_0, P)$ at P_0 is given by

$$\phi_0 : z \mapsto \frac{\delta}{g_0(x)}[I\{u = f_0(x)\} - Q_0(x)] + Q_0(x) - V(f_0, P_0),$$

where we now have defined the nuisance parameters $g_0(x) := P_0(\Delta = 1|X = x)$ and $Q_0(x) := P_0\{Y = f_0(x)|\Delta = 1, X = x\} = f_0(x)\pi_0(x) + \{1 - f_0(x)\}\{1 - \pi_0(x)\}$, so that $Q_0(x)$ is no

more than a simple transformation of $\pi_0(x)$. Under regularity conditions, the one-step debiased estimator

$$v_{n,OS} := \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{g_n(X_i)}[I\{U_i = f_n(X_i)\} - Q_n(X_i)] + Q_n(X_i)$$

of v_0 is nonparametric efficient, where g_n and π_n are consistent estimators of g_0 and π_0, and we define f_n and Q_n pointwise as $f_n(x) := I\{\pi_n(x) > 0.5\}$ and $Q_n(x) := f_n(x)\pi_n(x) + \{1 - f_n(x)\}\{1 - \pi_n(x)\} = \max\{\pi_n(x), 1 - \pi_n(x)\}$. The one-step debiased estimator of $v_{0,s}$ is defined identically except that all instances of f_n are replaced by $f_{n,s}$, which we define pointwise as $f_n(x) := I\{\pi_{n,s}(x) > 0.5\}$, with $\pi_{n,s}$ representing an appropriate estimator of $\pi_{0,s} := E_0\{\pi_0(X)|X_{-s} = x_{-s}\}$, obtained, for example, by flexibly regressing outcome $\pi_n(X)$ onto X_{-s}.

5 Numerical experiments

5.1 Simulation setup

We now present empirical results describing the performance of our proposed plug-in VIM estimator. In all cases, our simulated dataset included independent replicates of (X, Y), where X is a covariate vector with independent components $X_1, \ldots, X_p$ each following a standard normal distribution and a binary outcome Y following a Bernoulli distribution with success probability $\Phi(\beta_{01}x_1 + \ldots + \beta_{0p}x_p)$ conditional on $X = x$, where Φ is the standard normal distribution function. In Scenario 1, we set $p = 2$ and $\beta_0 = (2.5, 3.5)$, whereas in Scenario 2, we took $p = 4$ and $\beta_0 = (2.5, 3.5, 0, 0)$; thus, in all cases, the first two features had nonzero importance and the remaining features (if any) had zero importance. In this specification, Y follows a probit model. For each scenario considered, we generated 1000 random datasets of size $n \in \{100, 500, 1000, \ldots, 4000\}$, and considered the importance of both X_1 and X_2 in Scenario 1 and the importance of X_2 and X_3 in Scenario 2. In each scenario, we considered VIMs based on classification accuracy (Example 3) and the area under the ROC curve (Example 4). The true values of these VIMs implied by the data-generating mechanisms considered under Scenarios 1 and 2 are provided in Table 1. All analyses were performed using our R package vimp and may be reproduced using code available online (see details in the Supplementary Material). Since results were similar for accuracy and AUC, we only display results for accuracy here but provide results for AUC in the Supplementary Material.

In Scenario 1, we investigate the finite-sample properties of our proposal in a setting in which all features are truly important. We also use this setting to explore the effect of cross-fitting when using flexible estimators of f_0 and $f_{0,s}$. Specifically, we compare the performance of our estimation procedure with and without five-fold cross-fitting when using the following estimators of f_0 and $f_{0,s}$: a correctly specified (parametric) probit regression model; a generalized additive model (GAM; Hastie and Tibshirani, 1990, implemented in the R package mgcv); random forests (RF; Breiman, 2001, implemented in the R package ranger); and the Super Learner (SL; van der Laan et al, 2007, implemented in the R package SuperLearner). The latter estimator is a particular implementation of stacking (Wolpert,

1992) with favorable finite-sample and asymptotic performance guarantees (van der Laan et al., 2007). For the Super Learner, we used a library consisting of gradient boosted trees (Friedman, 2001, implemented in the R package xgboost), GAMs (implemented in the R package gam), and random forests, each with the default tuning parameter choices, in addition to parametric probit regression, with five-fold cross-validation to determine the optimal convex combination of these learners that minimizes the cross-validated negative log-likelihood risk. The resulting optimal convex combination of these individual algorithms is the Super Learner-based conditional mean estimator we adopt in any case where the Super Learner was fit. The tuning parameters considered for each algorithm are provided in the Supplementary Material. We do not use sample-splitting, since the results of Section 3 are valid under the alternative. We use Algorithm 2 to compute the cross-fitted point and standard error estimators for the importance of X_1 and X_2, from which we computed nominal 95% Wald-type confidence intervals. We then computed the empirical bias scaled by $n^{1/2}$, the empirical variance scaled by n, the empirical coverage of confidence intervals, and the width of these intervals.

In Scenario 2, we study the properties of our proposal under the null hypothesis. In this case, we used sample-splitting since the importance of X_3 and X_4 is zero. We again ran both cross-fitted and non-cross-fitted implementations, and considered the same learning strategies as in Scenario 1, with one exception: in this case, we added the lasso (Tibshirani, 1996, implemented in the R package glmnet) to the library of candidate learners in the Super Learner. As before, we computed point estimates and nominal 95% Wald-type confidence intervals but also obtained p-values for the null hypothesis using the sample-splitting procedure of Algorithm 3. We then computed the empirical bias scaled by $n^{1/2}$, the empirical variance scaled by n, the empirical coverage of confidence intervals, and the rejection probability for the proposed hypothesis test.

5.2 Primary empirical results

In Figure 2, we display the results of the experiment conducted under Scenario 1, in which both features have nonzero importance. For ease of visualization, we only display the results for X_2; the results for X_1 are similar and available in the Supplementary Material. In the top-left panel, we observe that the bias of the proposed estimators decreases to zero at rate faster than $n^{1/2}$ for all non-cross-fitted estimators except those based on random forests and Super Learner, whereas it does so for all cross-fitted estimators. This reflects the need for cross-fitting in cases where the Donsker class conditions of Theorem 1 may fail to hold. The top-right panel shows that the variance of all estimators is approximately proportional to n. In the bottom-left panel, we observe that coverage of nominal 95% confidence intervals increases to the nominal level with increasing sample size for all cases except the non-cross-fitted estimators based on random forests and Super Learner. In the bottom-right panel, we see that the width of these intervals decreases with increasing sample size, as expected.

In Figure 3, we display the results pertaining to null feature X_3 in the experiments conducted under Scenario 2. Here, it appears that the bias vanishes at a rate faster than $n^{-1/2}$ for both the cross-fitted and non-cross-fitted estimators (top-left panel), but that the variance of the non-cross-fitted estimators tends to increase with increasing sample size, especially for the

more flexible learning algorithms (top-right panel). We observe that empirical coverage is near the nominal level at all sample sizes (bottom-left panel). Finally, we see that the type I error of the proposed hypothesis test is controlled at the nominal level for all cross-fitted procedures, but not so for their non-cross-fitted counterparts, yielding an inflated type I error in that case (bottom-right panel). In the Supplementary Material, we present results for the non-null feature X_2, which show that power of the proposed test is large for all sample sizes considered here.

This simulation study suggests that the estimation and inferential procedures proposed, including our null testing approach, have good practical performance and are properly calibrated, as suggested by theory. Our findings suggest that cross-fitting is critical when flexible algorithms are used, in which case the estimation procedure without cross-fitting performs poorly while its cross-fitted counterpart instead shows good performance. This is the case both for point and interval estimation, as we explicitly show in the Supplementary Material. When correctly-specified parametric regression models are implemented, both procedures (with and without cross-fitting) perform similarly well. This reflects the fact that when parametric estimators are used, condition (B3) is typically satisfied and cross-fitting is then not needed.

5.3 Additional empirical results

In the Supplementary Material, we present results for additional features under Scenarios 1 and 2, observing similar patterns to those presented in Figures 2 and 3. We also consider pairing a non-cross-fitted standard error estimator with the cross-fitted estimation procedure, observing reduced coverage compared to the cross-fitted standard error estimator of Algorithm 2. Finally, we present results from additional investigations scrutinizing the performance of our proposal in higher dimensions, both with and without correlated features. We found, in small samples, that the presence of many independent null features results in an increased bias in the estimation of the importance of non-null features, with a corresponding decrease in empirical interval coverage. However, this inflated bias and undercoverage dissipate as the sample size increases. A similar pattern was seen in the presence of correlated null features. This suggests that greater dimensionality indeed increases the difficulty of the statistical problem at hand, but that correlation between features does not exacerbate this challenge beyond rendering more difficult the interpretation of the population VIM values.

6 Studying an antibody against HIV-1 infection

Broadly neutralizing antibodies (bnAbs) against HIV-1 neutralize a large fraction of genetic variants of HIV-1. Two harmonized, placebo-controlled randomized trials were conducted to evaluate VRC01, a promising bnAb, for its ability to prevent HIV-1 infection (Corey et al, 2021). A secondary objective was to assess how VRC01 prevention efficacy depends on amino acid (AA) sequence features of HIV-1. Because there are thousands of AA features, the statistical analysis plan for addressing this objective requires first restricting attention to a subset of AA features that putatively affect prevention efficacy. Given the underlying assumption that VRC01 prevents infection via *in vivo* neutralization, a useful approach may

be to rank AA features based on their estimated VIM for predicting *in vitro* neutralization — whether or not an HIV-1 virus is sensitive to neutralization by VRC01 — and select only the top-ranked features for further analyses.

In an effort to determine these important AA features, we analyzed the HIV-1 envelope (Env) AA sequence features of 611 publicly-available HIV-1 Env pseudoviruses made from blood samples of HIV-1 infected individuals (Magaret et al, 2019). All analyses accounted for the geographic region of the infected individuals. Among AA sequence features, approximately 800 individual features and 13 groups of features were of interest, e.g., polymorphic AA positions in Env AA that comprise the VRC01 antibody footprint to which VRC01 binds. These groups of features are described more fully in the Methods section of Magaret et al. (2019). There, we focused on a definition of variable importance as the difference in nonparametric R^2, and used as outcome an indicator of whether or not the 50% inhibitory concentration (IC_{50}, defined as the concentration of VRC01 necessary to neutralize 50% of viruses *in vitro*, with large values of the IC_{50} indicating that the virus was resistant to neutralization; Montefiori, 2009) was right-censored. However, the AMP trials have identified the 80% inhibitory concentration (IC_{80}) as a possible biomarker of prevention efficacy, with 75.4% estimated efficacy against the most sensitive viruses ($IC_{80} < 1$). Since many observations in our dataset are missing IC_{80} values, we use as outcome the binary indicator that $IC_{50} < 1$. Here, analyzing the same data set, we compare results based on the outcome of Magaret et al. (2019) with a variable importance analysis based on classification accuracy and AUC and the AMP-based outcome $IC_{50} < 1$. We consider a *marginal* VIM value, evaluating the intrinsic importance of each feature group of interest relative to geographic confounding variables – this can be achieved by considering the full feature vector in (2) to be simply the geographic confounders plus the feature group of interest. We provide a replication of Magaret et al. (2019) using a harmonized outcome in the Supplementary Material.

We used the Super Learner with a large library of candidate learners to estimate the involved regression functions. These learners included the lasso, random forests, and boosted decision trees, each with varying tuning parameters. Details on our library of learners are described in the Supplementary Material. Our resulting estimator is the convex combination of the candidate estimators, where we used five-fold cross-validation to determine the convex combination that minimized the negative log-likelihood risk. Finally, to make inference on the VIM values considered, we used the sample-split cross-fitted method (Algorithm 3) studied in the simulations under Scenario 2.

In Figure 4, we display the results of this analysis and the feature groups of interest. The top-ranked feature groups do not differ much between different VIMs but the magnitude of both importance and p-values depends greatly on the measure chosen. Both VIMs result suggest that the CD4 binding sites, the VRC01 binding footprint, sites with sufficient exposed surface area (ESA sites), sites with residues that co-vary with the VRC01 binding footprint (co-varying sites), and sites for indicating N-linked glycosylation (glycosylation sites) are the five most important groups. The finding that CD4 binding sites are in the most important groups across VIMs matches our expectations from basic science experiments that have identified AA substitutions at CD4 binding sites that altered VRC01 neutralization

sensitivity. This result is in line with Magaret et al. (2019). Based on our proposed hypothesis test, we computed p-values for a test of the strict null hypothesis (that is, $\beta = 0$) for each group. We found that AA features in the CD4 binding sites (group 2), VRC01 binding footprint (group 1), ESA sites (group 3), co-varying sites (group 5), and glycosylation sites (group 8) had p-values of 6.98×10^{-9}, 8.14×10^{-9}, 1.69×10^{-7}, 4.66×10^{-6}, and 8.07×10^{-6}, respectively, based on AUC (denoted by stars in Figure 4). Based on these analyses, AA features in these groups may be prioritized for the forthcoming trial data analyses. Additionally, taking the set of top-ranked features above a minimum threshold may help to narrow the set of gp160 AA sequence features to pre-specify for the analysis of the AMP trial data sets. Our recommendation, nonetheless, is to analyze all feature sets in secondary or supporting analysis of the AMP trial data sets to ensure that the results generated are comprehensive.

7 Discussion

We have proposed a general model-agnostic framework for statistical inference on population-level VIMs. These measures are summaries of the true data-generating mechanism, defined as a contrast between the predictiveness of the best possible prediction function based on all available features versus all features but those under consideration. We found that plug-in estimators of these VIMs are asymptotically linear and nonparametric efficient under regularity conditions. Through examples, we showed that many simple and commonly used VIMs fall within this framework. We found in numerical experiments that our proposed cross-fitted VIM estimator enjoys good operating characteristics, and that these characteristics match our theoretical expectations. More complex predictiveness measures and sampling scenarios, including missing data, may also be analyzed within our proposed framework, though these cases typically require more effort, including the computation of an influence function. Interpretation of the estimated VIMs depends on the application, and may include considering the ranked VIM values, or considering features with VIM values above some scientifically meaningful threshold.

Defining the importance of individual features in cases with large amounts of correlation is challenging. In practice, we recommend making use of any available background scientific knowledge either to group variables that are expected to be highly correlated or to develop an appropriate causal model. In settings where this knowledge is lacking, it may be useful to consider, for example, unsupervised methods to cluster variables before assessing variable importance; however, further work is needed to determine how to preserve inferential validity with any such procedure. One alternative approach to handling correlated features is to consider *marginal* importance, wherein each feature in turn could be considered as the ‘full set of covariates’ and its importance could be assessed relative to the null feature vector; if there are concerns about confounding factors, these can constitute the ‘null feature vector’ and each feature could be added to the potential confounders. A second alternative approach is to use measures like the Shapley Population VIM (SPVIM; Williamson and Feng, 2020). Since SPVIM is defined as the average increase in predictive power from including a particular feature in *all possible subsets* of the remaining features, use of this approach comes at the cost of significantly increased complexity.

The inferential procedures following Theorems 1 and 2 can be used whenever it is known a priori that the features of interest have non-zero importance. We note that, as an alternative, a nonparametric bootstrap scheme could be used in which f_0 and $f_{0,s}$ are not re-estimated over bootstrap samples but rather fixed at their original estimates. The use of this bootstrap is illustrated in the Supplementary Material, where it is shown to yield similar results as the inferential procedures described in this paper. If the features of interest may have zero importance, inference should generally be conducted using sample-splitting, as described in Section 3.4. There, we propose confidence intervals valid even when a feature of interest has zero importance and a test of the zero-importance hypothesis. Our numerical results suggest that the resulting test controls type I error rate at the desired level. However, since our procedure involves sample-splitting without data reuse, it does not fully exploit the information available in the data, and may possibly be improved upon. Use of the bootstrap in this context is complicated by the need to re-estimate f_0 and $f_{0,s}$. Developing a more powerful test of the null importance hypothesis is an important unresolved need. This objective could be achieved, on one hand, by considering modifications of our current approach, including averaging results over multiple splits of the dataset or choosing split sizes more judiciously, or on the other hand, by utilizing more complex analytical tools, including approximate higher-order influence functions. These ideas are being pursued in ongoing research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants F31AI140836, R01AI029168, R01HL137808, UM1AI068635, and S10OD028685. The opinions expressed in this article are those of the authors and do not necessarily represent the official views of the NIH.

Appendix A: Special case: standardized V-measures

Beyond smoothness requirements, the results presented in Section 3 do not impose much structure on the predictiveness measure. However, as is often the case that the predictiveness measure has the form $V(f, P) = a + V_1(f, P)/V_2(P)$ with

$$V_1(f, P) \coloneqq E_P\{G((Y_1, f(X_1)), \ldots, (Y_m, f(X_m)))\}$$

for some symmetric function $G : (\mathcal{Y} \times \mathcal{Y})^m \to \mathbb{R}$, where $a \in \mathbb{R}$ is a fixed constant, $V_2 : \mathcal{M} \to \mathbb{R}$ is Hadamard differentiable, and the expectation defining V_1 is over the distribution of independent draws $(X_1, Y_1), \ldots, (X_m, Y_m)$ from P. In this case, the plug-in estimator $V_1(f_n, P_n)$ of $V_1(f_0, P_0)$ is a V-statistic of degree m (Hoeffding, 1948), whereas the denominator $V_2(P_0)$ does not depend on f_0 and typically serves as a normalization constant. As such, we refer to any predictiveness measure of this form as a *standardized V-measure*. We note that each example presented in Section 2.3 is a standardized V-measure, defined respectively by:

1. $a = 1$, $G((u, v)) = -(u - v)^2$, $V_2(P) = var_P(Y)$, $m = 1$;
2. $a = 1$, $G((u, v)) = -\{u \log v + (1 - u)\log(1 - v)\}$, $V_2(P) = P(Y = 1)\log P(Y = 1) + P(Y = 0)\log P(Y = 0)$, $m = 1$;
3. $a = 0$, $G((u, v)) = I(u = v)$, $V_2(P) = 1$, $m = 1$;
4. $a = 0$, $G((u_1, v_1), (u_2, v_2)) = \{I(u_1 = 0, u_2 = 1, v_1 < v_2) + I(u_2 = 0, u_1 = 1, v_2 < v_1)\}/2$, $V_2(P) = P(Y = 1)\ P(Y = 0)$, $m = 2$.

This is useful to note because whenever V is a standardized V-measure, the influence function ϕ_0 of $V(f_n, P_n)$ can be described more explicitly. Specifically, its pointwise evaluation $\phi_0(z)$ at a given observation value $z = (x, y)$ is given by

$$m\left[\frac{E_0\{G((y, f_0(x)), (Y_2, f_0(X_2)), \ldots, (Y_m, f_0(X_m)))\}}{V_2(P_0)} - V(f_0, P_0)\right] - \frac{\dot{V}_2(P_0; \delta_z - P_0)}{V_2(P_0)} V(f_0, P_0)$$

with $\dot{V}_2(P_0; \delta_z - P_0)$ denoting the Gâteaux derivative of V_2 at P_0 in the direction $h = \delta_z - P_0$. Except for the influence function of the normalization estimator $V_2(P_n)$, which is typically straightforward to compute, this is an explicit form. In Examples 1–4, the influence function of $V(f_n, P_n)$ can thus be derived respectively as:

1. $\phi_0(z) = -\{y - \mu_0(x)\}^2/\sigma_0^2 + v_0\ \{2 - (y - \mu_0)^2/\sigma_0^2\}$;
2. $\phi_0(z) = -2[y \log \mu_0(x) + (1 - y)\log\{1 - \mu_0(x)\}]/\bar{\pi}_0 + v_0[2\log\{\pi_0/(1 - \pi_0)\}(y - \pi_0)/\bar{\pi}_0 - 1]$;
3. $\phi_0(z) = yI\{\mu_0(x) > 0.5\} + (1 - y)I\{\mu_0(x) \leq 0.5\} - v_0$;
4. $\phi_0(z) = (1 - y)P_0\{\mu_0(X) > \mu_0(x) \mid Y = 1\}/(1 - \pi_0) + yP_0\ \{\mu_0(x) > \mu_0(X) \mid Y = 0\}/\pi_0 - v_0[2 + (1 - 2\pi_0)(y - \pi_0)/\{\pi_0(1 - \pi_0)\}]$,

where here we have used the shorthand notation $\mu_0(x) \coloneqq E_0(Y|X = x)$, $\mu_0 \coloneqq E_0(Y)$, $\sigma_0^2 \coloneqq var_0(Y)$, $\pi_0 \coloneqq P_0(Y = 1)$, and $\bar{\pi}_0 \coloneqq \pi_0 \log \pi_0 + (1 - \pi_0)\log(1 - \pi_0)$. Furthermore, for standardized V-measures, condition (A2) is often easier to verify. For example, if $m = 1$, then it holds trivially since $V_1(f, P)$, the only component of $V(f, P)$ involving f, is linear in P.

References

Aas K, Jullum M, and Løland A (2019). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. arXiv:1903.10464.

Bach S, Binder A, Montavon G, Klauschen F, Müller K, and Samek W (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10(7), e0130140.

Benkeser D, Mertens A, Arnold B, Colford J, Hubbard A, Jumbe N, and van der Laan M (2018). A machine learning-based approach for estimating and testing associations with multivariate outcomes. arXiv:1803.04877.

Bickel P, Klaasen C, Ritov Y, and Wellner J (1998). Efficient and Adaptive Estimation for Semiparametric Models. Springer.

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

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, and Robins J (2018). Double/debiased machine learning for treatment and structural parameters.

Corey L, Gilbert P, Juraska M, Montefiori D, Morris L, Karuna S, Edupuganti S, Mgodi N, deCamp A, Rudnicki E, et al. (2021). Two randomized trials of neutralizing antibodies to prevent HIV-1 acquisition. New England Journal of Medicine 384 (11), 1003–1014.

Fisher A, Rudin C, and Dominici F (2018). All models are wrong but *many* are useful: variable importance for black-box, proprietary, or misspecified prediction models, using *model class reliance*. arXiv:1801.01489.

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

Garson D (1991). Interpreting neural network connection weights. Artificial Intelligence Expert.

Grömping U (2006). Relative importance for linear regression in r: the package relaimpo. Journal of Statistical Software.

Grömping U (2009). Variable importance in regression: linear regression versus random forest. The American Statistician 63(4), 308–319.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, and Pedreschi D (2018). A survey of methods for explaining black box models. ACM Computer Surveys 51 (5), 93:1–93:42.

Hastie T and Tibshirani R (1990). Generalized Additive Models, Volume 43. CRC Press.

Hoeffding W (1948). A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics 19(3), 293–325.

Ishwaran H (2007). Variable importance in binary regression trees and forests. Electronic Journal of Statistics 1, 519–537.

LeDell E, Petersen M, and van der Laan M (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electronic Journal of Statistics.

Lei J, G'Sell M, Rinaldo A, Tibshirani R, and Wasserman L (2017). Distribution-free predictive inference for regression. Journal of the American Statistical Association.

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

Magaret C, Benkeser D, Williamson B, Borate B, Carpp L, et al. (2019). Prediction of VRC01 neutralization sensitivity by HIV-1 gp160 sequence features. PLoS Computational Biology 15(4), e1006952.

Montefiori D (2009). Measuring HIV neutralization in a luciferase reporter gene assay. In: Prasad VR, Kalpana GV (eds) HIV Protocols. Methods in Molecular Biology 485, 395–405.

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

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

Nelder J and Wedderburn R (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A 135 (3), 370–384.

Pfanzagl J (1982). Contributions to a general asymptotic statistical theory. Springer.

Ribeiro M, Singh S, and Guestrin C (2016). 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, pp. 1135–1144.

Shrikumar A, Greenside P, and Kundaje A (2017). Learning important features through propagating activation differences. arXiv:1704.02685.

Strobl C, Boulesteix A, Zeileis A, and Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(1), 1.

Sundararajan M, Taly A, and Yan Q (2017). Axiomatic attribution for deep networks. arXiv:1703.01365.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 267–288.

van der Laan M (2006). Statistical inference for variable importance. The International Journal of Biostatistics 2(1). doi: 10.2202/1557-4679.1008.

van der Laan M, Polley E, and Hubbard A (2007). Super learner. Statistical Applications in Genetics and Molecular Biology 6(1), Online Article 25.

van der Laan M and Rose S (2011). Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.

Wei P, Lu Z, and Song J (2015). Variable importance analysis: a comprehensive review. Reliability Engineering & System Safety 142, 399–432.

Williamson B and Feng J (2020). Efficient nonparametric statistical inference on population feature importance using Shapley values. In Proceedings of the 37th International Conference on Machine Learning, Volume 119 of Proceedings of Machine Learning Research, pp. 10282–10291.

Williamson B, Gilbert P, Carone M, and Simon N (2020). Nonparametric variable importance assessment using machine learning techniques. Biometrics 77, 9–22.

Wolpert D (1992). Stacked generalization. Neural Networks 5(2), 241–259.

Zheng W and van der Laan M (2011). Cross-validated targeted minimum-loss-based estimation. In Targeted Learning, pp. 459–474. Springer.

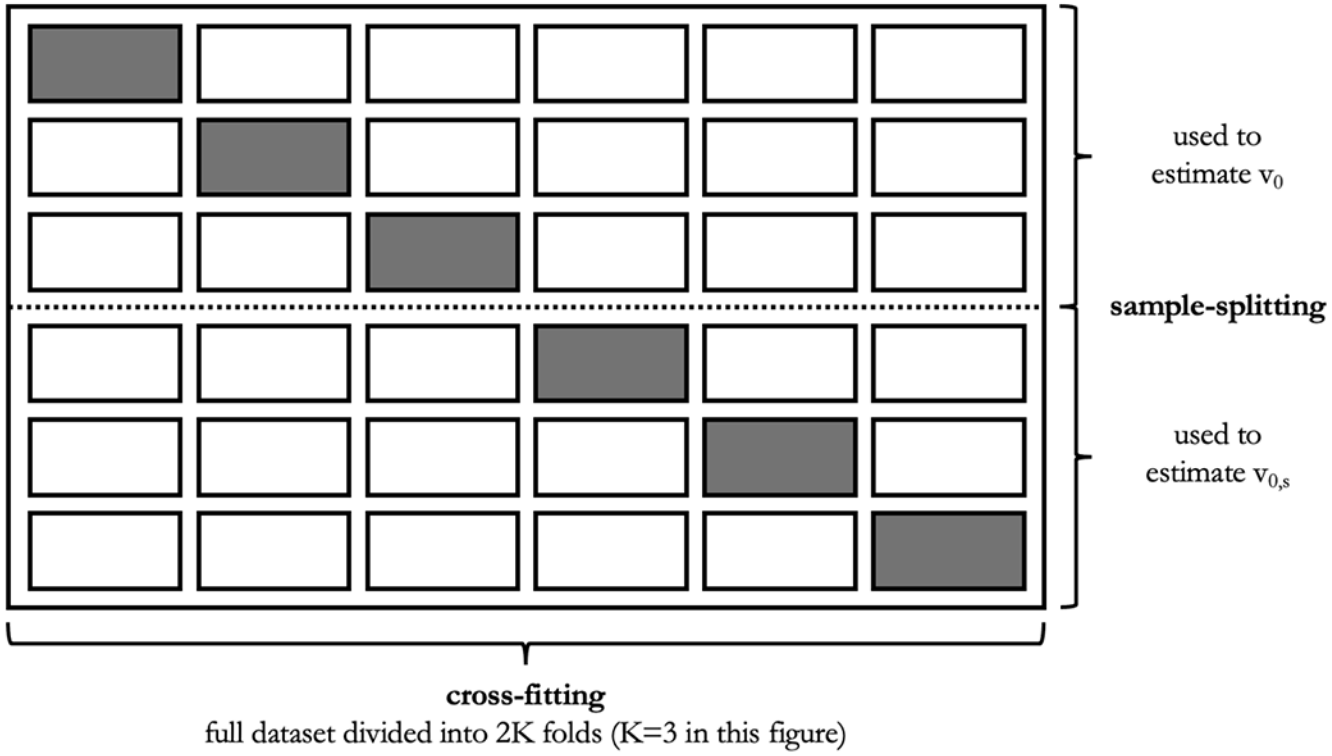

Figure 1:

Illustration of dataset subdivision when sample-splitting and cross-fitting are used simultaneously for valid inference under the zero-importance hypothesis (sample-splitting) without requiring Donsker class conditions (cross-fitting). Each row represents the entire dataset with a different subset singled out (in grey) as testing set. To estimate v_0, the top three rows are used. In each such row, f_0 is estimated using data in the white cells, and v_0 is estimated using the resulting estimate of f_0 and data in the grey cells. Row-specific estimates of v_0 are then averaged. The process is repeated for estimating $v_{0,s}$ but instead using the bottom three rows and estimating $f_{0,s}$ rather than f_0.

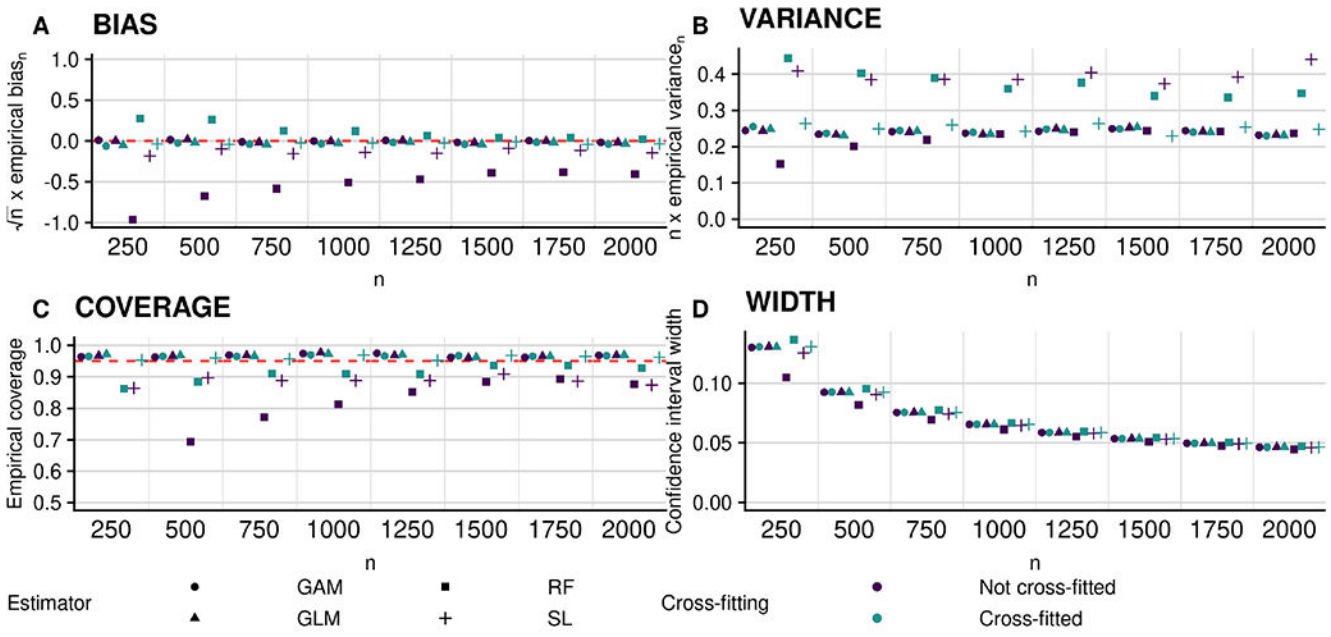

Figure 2:

Performance of plug-in estimators for estimating (non-zero) importance of X_2 in terms of accuracy under Scenario 1 (all features have non-zero importance). Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and average width of these intervals. Circles, triangles, squares and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF), and the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively.

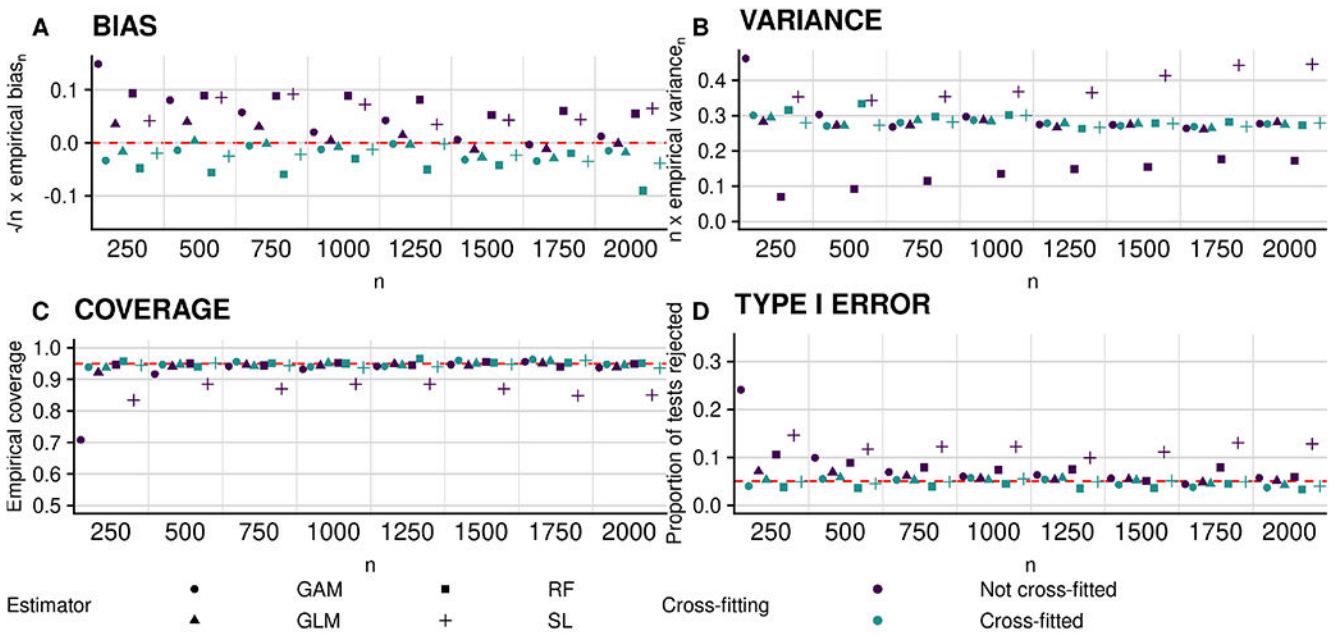

Figure 3:

Performance of plug-in estimators for estimating (zero) importance of X_3 in terms of accuracy under Scenario 2. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and empirical type I error of the proposed hypothesis test. Circles, triangles, squares and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF), and the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively.

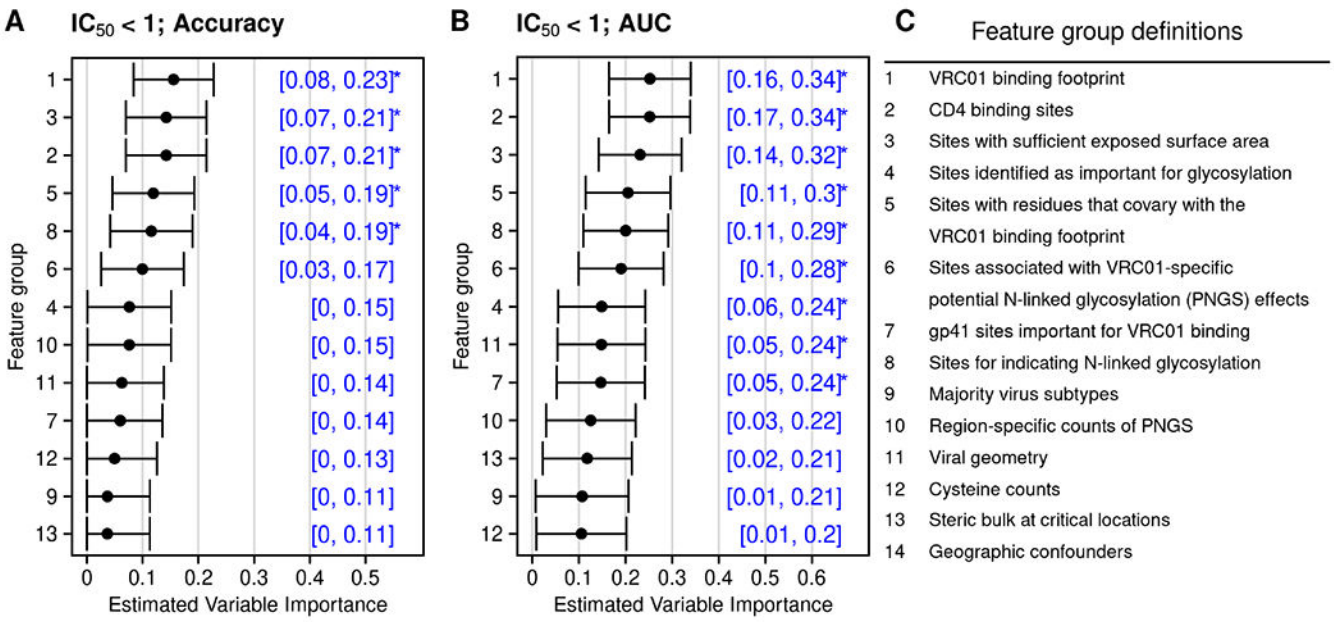

Figure 4:

Variable importance measured by accuracy (panel A) and AUC (panel B) for the groups defined in panel C. Stars denote importance deemed statistically significantly different from zero at the 0.0038 (0.05 / 13) level.

Table 1:

Approximate values of $\psi_{0,s}$ in the numerical experiments.

	Feature of interest			
Importance measure	X_1	X_2	X_3	X_4
Accuracy	0.136	0.236	0	0
Area under the ROC curve	0.105	0.221	0	0