A unified approach to model-agnostic variable importance

Brian David Williamson

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Marco Carone, Chair

Noah Simon, Chair

Peter B. Gilbert

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

A unified approach to model-agnostic variable importance

Brian David Williamson

Co-Chairs of the Supervisory Committee:
Marco Carone
Department of Biostatistics

Noah Simon
Department of Biostatistics

Assessing the relative contribution of subsets of features towards predicting the response is often of interest in predictive modeling applications; this contribution is typically referred to as variable importance. Often, simple population models are used because the associated variable importance measure is easy to interpret; however, estimates may be misleading if the model used is overly simplistic. In an effort to improve prediction performance, complex prediction algorithms are often used instead; however, in these cases variable importance is often defined as a function of the algorithm rather than a summary of the population, rendering formal statistical inference on population importance difficult. In this dissertation, we propose a unified model-agnostic framework for statistical inference on population-level variable importance. Specifically, we define variable importance 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 discuss general conditions under which a simple estimator of this importance is nonparametric efficient and allows the construction of valid confidence intervals. We also propose a valid strategy for hypothesis testing. 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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As with any journey, the end result — this dissertation — is merely one step along the path. All of the steps that led me to this point are what truly matters. With that in mind, I want to thank the great number of people who made the journey to completing this dissertation an incredible learning experience.

Early in my undergraduate studies at Pomona College, Shahriar Shahriari taught me that studying mathematics could be incredibly rewarding, and has an invaluable resource ever since. Barbra Richardson encouraged me to participate in the Summer Institutes hosted by the University of Washington Department of Biostatistics. Scott Emerson and Jo Hardin taught me that statistics provided an opportunity to combine math research with scientific research to address exciting and important questions. The final tack in my journey came the summer after my third year, where I worked in a cancer research bioinformatics lab with Benedict Anchang, Andrew Gentles, and Sylvia Plevritis. I am incredibly grateful for this first research experience — it lit a fire that led me directly to studying biostatistics at UW.

My graduate experience was made possible by many people, but I first want to thank my advisors, Marco Carone and Noah Simon. During my first year in graduate school, Marco and Noah taught me about machine learning and semiparametric theory on the whiteboard for one hour per week. This led to a summer project, which turned into a paper, which led to this dissertation. I am incredibly thankful to both Marco and Noah for investing so much time in me, and for trusting me with interesting research projects and incredible collaborators throughout my time at UW. Throughout my graduate studies, both Marco and Noah have been trusted mentors, inspiring colleagues, and great friends.

I thank Scott Emerson, Jim Hughes, Peter Gilbert, and Amy Willis for their feedback,

support, and mentorship over the years. I also thank the people who at one time or another served on my dissertation committee: Marco, Noah, Scott, Peter, Carey Farquhar, and Annette Fitzpatrick. You all made the examination process much more bearable! Finally, I am thankful to the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under award number F31 AI140836 for financial support over the past 1.5 years. This support enabled me to pursue an incredible array of research projects. Please note that the opinions expressed in this dissertation are my own and do not necessarily represent the official views of the NIAID or the NIH.

Throughout my time at UW, I have had the honor to learn, collaborate, and work with many incredible people. A huge thank you to Gitana Garofolo, who goes above and beyond the call of duty to make sure that students are supported and have a positive learning environment. Additionally, I thank my cohort for being a source of support through all of the stress, for helping to get through exams and courses, and for everything you have taught me about statistics, science, and life. Thank you as well to the other students and collaborators who have taught and mentored me through my time at UW. Finally, I want to thank my intramural soccer team, Bert's Desserts, for hours of fun (and a free T-shirt!).

Finally, I am thankful to all of the people who supported me throughout my life and throughout graduate school with their friendship. Though I met you all at different times in my life and in different ways — including my close childhood friends, college roommates and teammates, hiking buddies, and so many more — you have made my journey fun and have made my life truly wonderful. I thank my family for instilling in me a lifelong love of learning, and for modeling honesty, ethics, and hard work. My final thanks go to Alex Lincoln for being my best friend and life partner; for your patience and understanding during times of stress and large workloads; and for helping me to go outside and enjoy nature every once in a while. I am incredibly lucky to have you in my life.

DEDICATION

In the order of my meeting them, this dissertation is dedicated to:

Susan G. Williamson

David M. Williamson

Christina L. Williamson

Alexandra E. Lincoln

Chapter 1

INTRODUCTION

1.1 Background and motivation

A traditional statistical analysis typically seeks to address one (or both) of two goals: (i) *information*, where we wish to extract some information about the association between predictors and response in nature; and (ii) *prediction*, where we desire to predict the value of the response for future predictors [Breiman, 2001b]. The main classical approach to (i) is to assume a stochastic model (e.g., a linear regression model) for the data, typically with a finite-dimensional parameter that indexes the effect of interest. Then parameter estimates and predictions are easily obtained from model fit outputs. The main classical approach to (ii) is to find the best algorithm for predicting the response given the features.

Each of these approaches suffer drawbacks. Traditional statistical modeling approaches may be misleading in view of potential model misspecification. If the proposed stochastic model does not hold in the population, what do the parameters – and our estimates of them – truly mean? What information, if any, do we gain about nature in this case? In an attempt to improve prediction performance and gain relevancy, algorithmic modeling approaches are often used instead. Here, it is increasingly common to fit a black-box algorithm to the data, with the goal of achieving the best possible predictions. However, these algorithmic approaches typically do not tell us directly about the true relationship between the predictors and response without additional assumptions; therefore, if our goal is to understand nature, algorithmic approaches may not be appropriate. This is a nuanced issue: Breiman [2001b] and the discussants therein showcase much of the ideological spectrum.

In this dissertation, we argue that the strengths of each approach may be combined by treating the statistical model and the parameter of interest as separate concepts. This serves

to divorce the estimation technique from the data-generating mechanism, and allows us to learn about nature while using flexible learning techniques. In particular, we define the parameter of interest as a finite-dimensional vector (or scalar) summary of a distribution, rather than using a parameter of interest implied by a (potentially restrictive) statistical model. This approach is based in ideas from semiparametric estimation theory.

Our specific goal is statistical inference on *variable importance*. This is often defined as the relative contribution of subsets of variables towards predicting the response. We use tools from semiparametric estimation theory to rigorously define and understand variable importance as a summary of the true, underlying data-generating mechanism. Then, we use state-of-the-art machine learning-based methods to obtain best-performing predictive algorithms. These predictive algorithms are used to estimate variable importance. Our proposed procedures allow consistent and efficient estimation of the true variable importance of features in predicting the outcome, and allow valid statistical inference for the true importance even when flexible machine learning-based methods are used in estimation. The resulting importance estimates, confidence intervals, and p-values provide researchers with an understanding of the true importance, which may in turn inform how nature relates the predictors to the response. We make this more concrete in the following example, which we revisit in more detail in Chapter 2.

Example: HIV-1 vaccine development

We still do not have a broadly efficacious vaccine against HIV-1. Developing an active vaccine that directly stimulates the immune system has proven difficult [McMichael and Hanke, 2003]. One promising route forward is passive antibody administration, where antibodies are delivered directly by injection or infusion [Kwong et al., 2013]. Human antibodies that effectively neutralize a large fraction of genetic variants of HIV-1 are known as broadly neutralizing antibodies (bnAbs). A number of bnAbs that neutralize large proportions of HIV-1 strains have been isolated from individuals with chronic HIV-1 infection [Klein et al., 2013, Walker et al., 2011, Wu et al., 2010, Zhou et al., 2010]. For instance, the bnAb VRC01 neutralizes more than 90% of all viral strains tested in vitro, including all major

subtypes of HIV-1 [Wu et al., 2010]. VRC01 has moved through three phase 1 clinical trials [Ledgerwood et al., 2015, Lynch et al., 2015a, Mayer et al., 2016]. The Antibody Mediated Prevention (AMP) trials [NCT02716675, NCT02568215, Gilbert et al., 2017] are the first proof-of-concept efficacy trials in humans to determine whether passive administration of a bnAb can prevent HIV-1 acquisition.

A key secondary objective of the AMP trials is a sieve analysis [Gilbert et al., 2001], which investigates how intervention-mediated protection from HIV-1 infection varies with genotypic characteristics of the exposing viruses. However, there are a large number of ways to define an HIV-1 genotype based on the HIV-1 amino acid sequence. Performing an exhaustive search over all genotypic characteristics results in low statistical power to detect effects after adjusting for multiple comparisons. Past genotypic sieve analyses for HIV-1 vaccine candidates [Gilbert et al., 2001, Rolland et al., 2011, Rolland and Gilbert, 2012] have thus focused on a small, pre-specified set of amino acid features, often defined based on prior knowledge of the vaccine or of HIV-1 [Rolland and Gilbert, 2012]. While excellent research has already been done to understand features of the HIV-1 virus that affect VRC01 resistance [Guo et al., 2012, Li et al., 2011, Lynch et al., 2015b, Utachee et al., 2014, Wibmer et al., 2013, Zhou et al., 2010], HIV researchers have shown that using machine learning-based techniques can often both corroborate prior knowledge and suggest new areas of focus [Korber et al., 2006, Li et al., 2011]. There is increased interest in combining the available data with flexible machine learning-based techniques to determine the amino acid features that are important in predicting IC_{50}, the 50% inhibitory concentration of an antibody against HIV-1 virus [Montefiori, 2009]. The knowledge generated by machine learning-based analyses, when combined with other prior knowledge, will be used to specify the statistical analysis plan for the sieve analysis in the AMP trials.

In Magaret et al. [2019], we used machine learning-based methods and data from the Compile, Neutralize, and Tally Neutralizing Antibody Panels (CATNAP) database [Yoon et al., 2015] to create a score that predicts resistance to VRC01 based on HIV-1 genotypic features. With this predictive score in hand, our second objective was to estimate the

importance of these genotypic features. The importance rankings, obtained on this external dataset, may help in pre-specifying regions of the HIV-1 genotype to focus on in the AMP trial sieve analysis.

Additionally, quantitative IC_{50} may not be the best measure of neutralization sensitivity. Recent work has shown — and we found in Magaret et al. [2019] — that dichotomous outcomes indicating a virus's resistance versus sensitivity status based on IC_{50} may be better predicted by HIV-1 amino acid sequence features. Therefore, methods that define importance relative to a scientifically meaningful measure are necessary to truly understand the importance of amino acid sequence features in predicting neutralization sensitivity of the HIV-1 virus to bnAbs. In the next section, we discuss variable importance measures and approaches to estimation in further detail.

1.2 Defining variable importance

1.2.1 Two notions of variable importance

As we discussed in the previous section, defining variable importance relative to a scientifically meaningful measure is a crucial step towards obtaining results that can be interpreted in a manner consistent with the goals of the analysis. This fact has been noted previously in the literature [Nathans et al., 2012]. However, the goal of a variable importance analysis is not immediately clear from the term *variable importance* alone. In a previous review article, Wei et al. [2015] identified three definitions of variable importance: (i) quantifying the change in model output when input variable(s) are changed or permuted; (ii) quantifying the contribution of uncertainty in the input variable(s) to the response variable; and (iii) quantifying the strength of dependence between the input variable(s) and the response variable. We argue that definitions (ii) and (iii) can be collapsed, leaving us with the following definitions of variable importance: *extrinsic* importance, which describes an algorithm used in data analysis or prediction and how it makes use of the features; and *intrinsic* importance, which describes the population-level relationships between the features and the outcome.

Each of these goals is an important part of scientific research. Hopefully, our classification is a useful step towards specifying the goal of a variable importance analysis. We provide further discussion and examples of these two types of variable importance in the following two sections.

1.2.2 Extrinsic variable importance measures

Black-box algorithms have been increasingly used in a wide variety of domains, leading to a corresponding increase in research on interpreting these black boxes [see, e.g., Guidotti et al., 2018]. Interpreting algorithms touches on principles of fairness, transparency, and justice, among other areas [see, e.g., Ribeiro et al., 2016, Fisher et al., 2018]. Random forests [Breiman, 2001a], boosted trees [Friedman, 2001], and neural networks [Barron, 1989] are among the most commonly used machine learning-based methods in both regression and classification prediction problems. While each has strengths and weaknesses in terms of prediction performance, we will not discuss this further; our main goal is variable importance. Methods for explaining black boxes — among these, methods for random forests, boosted trees, and neural networks — can be broadly classified as extrinsic variable importance measures: they describe the reliance of an algorithm on a set of features.

Extrinsic variable importance in tree-based methods, including random forests and boosted trees, is well studied [see, e.g., Breiman, 2001a, Strobl et al., 2007, Ishwaran, 2007, Grömping, 2009], and is typically based on the tree structure. Consider a tree T with J terminal nodes. Each non-terminal node is the result of splitting the sample space into two regions based on the value of a particular variable, chosen to minimize the squared error (or classification error) at each step. At each non-terminal node t, one of the predictors is used to partition the region associated with that node; this chosen variable maximizes this estimated improvement in squared error over fitting a constant to the entire region. Importance, then, is defined as the sum of these improvements for each time the variable was chosen. This definition is used in many studies [see, e.g., Breiman, 2001a, Friedman, 2001, Hastie et al., 2009], and may be generalized to boosted trees or random forests by averaging over all trees.

This idea may be generalized to a permutation-based approach. The key idea is that, since bootstrapping is typically used to fit either random forests or boosted trees, approximately one-third of the observations are not used to fit a given tree at each step [Breiman, 2001a]. These "out-of-bag" observations may be used to estimate the reduction in mean squared error (MSE) resulting from a permutation of the observed values of a given covariate; these values are computed in a similar fashion to the importance values given above.

Extrinsic variable importance in neural networks is typically defined in terms of the weights between the nodes that make up the neural network. Both overall [Garson, 1991, Olden and Jackson, 2002, Lipovetsky and Conklin, 2001] and local [Bach et al., 2015, Shrikumar et al., 2017, Sundararajan et al., 2017] measures have been proposed. More specifically, a neural network is typically defined by layers of regression: in a fully connected neural network, each predictor is connected to each node in the first hidden layer, with edge weights denoting connection strength; each node in a hidden layer is connected to each node in the next hidden layer; and the final hidden layer is connected to the outcome (see Figure 1.1 for an example neural network). Importance is typically a weighted average of the edge weights through the network that connects a given predictor to the response.

In addition to algorithm-specific extrinsic measures, recent proposals for extrinsic variable importance may be used with multiple types of algorithms. Examples include local interpretable model-agnostic explanations [Ribeiro et al., 2016], Shapley additive explanations [Lundberg and Lee, 2017], and model reliance [Fisher et al., 2018]. Additionally, recent work in developing data-adaptive target parameters [see, e.g., LeDell et al., 2015, Hubbard et al., 2016, Benkeser et al., 2018] provides population-level summaries of a chosen algorithm.

These algorithm-specific approaches have a few relative strengths. First, they are often computationally efficient: variable importance estimates and conditional mean estimates are often obtained simultaneously. Second, each importance measure provides information about the performance of the fitted algorithm. Third, these measures may provide information on how the fitted algorithm makes predictions. This extrinsic importance may lead to improved interpretation of these black-box algorithms. However, of the approaches to assessing extrin-

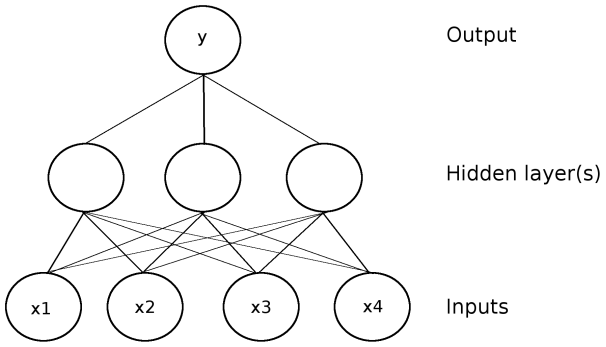

Figure 1.1: An example neural network structure with four input variables and a single hidden layer with three hidden nodes.

sic variable importance laid out in this section only model reliance [Fisher et al., 2018] and the data-adaptive target parameters appear to allow formal statistical inference.

1.2.3 Intrinsic variable importance measures

A second approach to measuring variable importance is to separate the definition of importance from both the statistical model and the algorithm used to estimate the conditional mean. Here, importance is defined as a property of the data-generating mechanism, and is the parameter of interest. We refer to this as *intrinsic* variable importance assessment, since the parameter provides a population-level summary of the importance of measured features. Valid statistical inference requires that the procedure for estimating importance correctly account for using flexible estimation techniques to estimate the conditional means. Several variable importance parameters have been proposed in recent work. These parameters may be best compared through their use in either regression or classification problems.

Extrinsic variable importance measures in regression problems include nonparametric extensions of classical measures. This includes the nonparametric R^2 [Doksum and Samarov, 1995, Huang and Chen, 2008, Yao et al., 2005]; the marginalized mean difference [van der Laan, 2006, Chambaz et al., 2012, Sapp et al., 2014]; and the mean absolute difference [Lei

et al., 2017]. We discuss each of these measures in more detail in Chapter 3.

Each of these importance measures may also be used in classification problems; however, using importance measures tailored to binary data may yield additional insights. Both the classically defined population deviance [Nelder and Wedderburn, 1972] and the population area under the receiver operating characteristic curve (AUC) are intrinsic importance measures; current implementations, however, typically use a slightly different definition of each of these measures depending on the algorithm used. Nonparametric extensions of the deviance appear to be under-studied. Nonparametric estimation of AUC is classically done using a method-of-moments estimator, but this approach does not allow the inclusion of covariates. Extensions of AUC exist; many of these, however, are parametric or semiparametric and rely on a generalized linear model for the inclusion of covariates [see, e.g., Thompson and Zucchini, 1989, Pepe, 2000, Pepe et al., 2006, Dodd and Pepe, 2003]. LeDell et al. [2015] describe a fully nonparametric procedure, but their parameter of interest is tied to the chosen classification algorithm, as we mentioned in the previous section; it truly gives more insight into the performance of the classification algorithm than the underlying population.

The major strength of these intrinsic measures with an algorithm-agnostic approach to estimation is that valid statistical inference on intrinsic measures is possible. We believe that taking this approach leads to strong science by not relying heavily on potentially restrictive modeling assumptions, while still gaining insight into the interplay between features and outcome. However, work remains to propose and evaluate model-agnostic parameters that have been studied in full generality, are entirely agnostic to the estimation technique, allow valid inference, and provide an interpretation that is well suited to the scientific task at hand. We provide some tools towards this goal in this dissertation.

1.3 Concluding remarks and outline of the dissertation

There is a wide spectrum of beliefs on how best to analyze data, summarized well in Breiman [2001b] and the discussion therein. However, we argue that by (1) defining the statistical model as only a collection of plausible distributions that contains the true data-generating

mechanism; (2) only encoding knowledge backed by science in any restrictions to this model, or leaving it fully nonparametric; (3) defining variable importance as a summary of the true, underlying data-generating mechanism; and (4) studying the properties of this parameter to create an efficient estimator of population variable importance, researchers may use possibly flexible estimation procedures for the conditional mean while still obtaining valid inference for the variable importance of measured covariates. In this dissertation, we propose a unified framework for model-agnostic variable importance. In this framework, a simple estimator of importance is nonparametric efficient, and we can obtain valid confidence intervals and p-values even when using flexible techniques to estimate importance.

While our proposed approach is broadly applicable, one area in which we have employed the procedure is in the HIV-1 vaccine trial pipeline, where machine learning-based approaches are increasingly used to cope with the large amount of data generated by expensive clinical trials. Variable importance may be a helpful step to consider when planning for future trials, which in turn may help to reduce the time to a broadly efficacious vaccine.

The remainder of this dissertation is organized as follows. In Chapter 2, we discuss the role of variable importance in HIV-1 vaccine development (introduced in the *Example* above) in greater detail. This leads into our proposed statistical methods work. Our initial proposal is based on population R^2 (Chapter 3). We then extend these results to a large class of population-level variable importance measures in Chapter 4. These results allow us to study nonparametric extensions of classification accuracy and the area under the receiver operating characteristic curve. The results of Chapters 3 and 4 allow us to measure the importance of a subgroup of covariates X_s relative to either the full covariate vector X, the null model that does not use covariate information, or a vector of potential confounding variables. In Chapter 5, we propose an intermediate measure based on the Shapley value [Shapley, 1953]. This measure has a meaningful interpretation for each individual feature even when some features are correlated. Finally, we provide some concluding remarks and future directions of research in Chapter 6.

Chapter 2

VARIABLE IMPORTANCE IN THE CONTEXT OF HIV-1 VACCINE DEVELOPMENT

2.1 Introduction

Developing an effective vaccine that prevents HIV-1 infection has proven challenging [McMichael and Hanke, 2003], in part due to the genetic and antigenic diversity of the HIV-1 envelope (Env) glycoprotein. The Env glycoprotein spike mediates viral entry into cells, and is the major target of neutralizing antibodies. For this reason, host immune pressures have led to Env having up to 30% variation across different genetic subtypes [see, e.g., Wibmer et al., 2015]. One promising route forward for preventing HIV-1 infection is passive antibody administration, where antibodies are delivered directly by injection or infusion [Kwong et al., 2013]. Human antibodies that effectively neutralize a large fraction of genetic variants of HIV-1 are known as broadly neutralizing antibodies (bnAbs). These bnAbs have demonstrated promise in preventing HIV-1 infection by targeting a wide spectrum of the genetic and antigenic diversity in the Env glycoprotein [see, e.g., Wu et al., 2010, Zhou et al., 2010, Walker et al., 2011, Klein et al., 2013, Burton and Hangartner, 2016], and typically target conserved elements in one of five regions of the Env gp160 protein: the V2 variable region, the N332 region in the V3 region, the CD4 binding sites, the gp120–gp41 interface, and the membrane proximal external region [Wibmer et al., 2015]. Information about the Env amino acid (AA) signature patterns that are associated with a neutralization phenotype of interest [Gnanakaran et al., 2010] may help to design bnAb regimens and HIV-1 vaccines.

A number of bnAbs that neutralize large proportions of HIV-1 strains have been isolated from individuals with chronic HIV-1 infection [Klein et al., 2013, Walker et al., 2011, Wu et al., 2010, Zhou et al., 2010]. For instance, the bnAb VRC01, an IgG1 monoclonal an-

tibody that targets a region in the CD4 binding sites [Wu et al., 2010, Zhou et al., 2010], neutralizes more than 90% of all viral strains tested in vitro, including all major subtypes of HIV-1 [Wu et al., 2010]. Three phase 1 clinical trials using VRC01 have been completed [Ledgerwood et al., 2015, Lynch et al., 2015a, Mayer et al., 2016]. The Antibody Mediated Prevention (AMP) trials [NCT02716675, NCT02568215, Gilbert et al., 2017] are the first proof-of-concept efficacy trials in humans to determine whether passive administration of a bnAb can prevent HIV-1 acquisition.

After the AMP trials conclude, we will conduct a *sieve analysis*, which investigates the extent to which intervention-mediated protection from HIV-1 infection varies with characteristics of the exposing viruses [Gilbert et al., 2001]. As part of this investigation, we will compare AA sequence features of breakthrough founding Env sequences from infected VRC01 recipients versus infected placebo recipients. An AA sequence sieve effect for a particular AA sequence feature is defined as statistically significant variation in prevention efficacy of VRC01 against viruses with different levels of this feature. For example, an analysis may consider the presence or absence of a residue at a single amino acid site, where the absence of the residue influences whether VRC01 can neutralize the virus. This sieve analysis approach has been used in other preventative vaccine efficacy trials [see, e.g., Rolland et al., 2012]. However, there are a large number of ways to define an HIV-1 genotype based on the HIV-1 amino acid sequence. Performing an exhaustive search over all genotypic characteristics results in low statistical power to detect effects after adjusting for multiple comparisons. Thus, there is interest in defining a small set of AA features to be used in the primary AMP sieve analysis.

The remainder of this chapter is organized as follows. We discuss how HIV-1 neutralization outcomes are measured in Section 2.2. In Section 2.3, we detail how we defined groups of AA features for our analysis of VRC01 neutralization [Magaret et al., 2019]. We highlight how variable importance will be used in creating the AMP statistical analysis plan in Section 2.4, and provide concluding remarks in Section 2.5.

2.2 Measuring HIV-1 neutralization outcomes

The dataset used in Magaret et al. [2019] and considered here consists of Env gp160 viral AA sequences, neutralization values for VRC01, and associated annotations retrieved from the Compile, Neutralize, and Tally Antibody Panels (CATNAP) database [Yoon et al., 2015]. These neutralization values are assessed by the TZM-bl assay [Sarzotti-Kelsoe et al., 2014]. The TZM-bl assay is based on TZM-bl cells, which are CXCR4-positive HeLa cells designed to express CD4 and CCR5, and are permissive to infection by a wide range of HIV viruses and Env pseudoviruses. The output of this assay is a standard curve based on 5 parameters, which can be used to determine the inhibitory concentration of the antibody, i.e., the concentration necessary to neutralize a percentage of viruses *in vitro* [Montefiori, 2009]. The CATNAP database provides the 50% inhibitory concentration (IC_{50}) value and IC_{80} value, which are the concentration of antibody necessary to neutralize 50% and 80% of viruses in vitro, respectively.

However, in some cases the VRC01 IC_{50} or IC_{80} values were right-censored. This right-censoring limit varied between the studies that contributed data to CATNAP: for some studies, the upper limit of detection of was 50 μg/ml, while for others, it was lower than 50 μg/ml. In these cases, we assumed that neutralization was not detected. Based on this, we defined a binary variable "IC_{50} censored" that denoted whether or not the IC_{50} value for a particular virus was censored based on the detection limit in the study that submitted the IC_{50} value. This outcome describes whether the virus was resistant (IC_{50} value was censored) versus sensitive (IC_{50} value not censored) to neutralization by VRC01.

2.3 Defining amino acid feature groups

We included AA sites with potential relevance to VRC01-mediated neutralization of HIV-1. For a given AA site to be included in the analysis, it had to pass a minimum variability filter: the residue had to differ from the consensus residue in at least three sequences in the entire analysis dataset derived from CATNAP. Furthermore, indicators for the presence of a

residue at a specific site were only included if that residue was present in at least three viral sequences at that site across the entire dataset. After applying this minimum variability filter, we selected any AA site that fell into any of the following pre-defined groups (these groups are additionally defined in Magaret et al. [2019]):

Group 1 **(VRC01 binding footprint sites)** the 35 AA positions in gp120 identified as contact sites between VRC01 and HIV-1 Env;

Group 2 **(CD4 binding sites)** all AA positions that constitute the CD4 binding site;

Group 3 **(Sites with sufficient exposed surface area)** all AA positions with sufficient exposed surface area, as calculated using the DSSP program [Joosten et al., 2011] using crystal structures of VRC01;

Group 4 **(Sites identified as important for glycosylation)** AA positions related to the glycan fence or identified as sites where VRC01 interacts with the Env trimer;

Group 5 **(Sites with residues that covary with the VRC01 binding footprint)** all gp120 AA positions (excluding those in the signal peptide) not included in Groups 1–4 that are outside the VRC01 binding footprint and that covary with at least one footprint position;

Group 6 **(Sites associated with VRC01-specific potential N-linked glycosylation)** (PNG sites) all AA positions with VRC01-specific PNG; and

Group 7 **(Sites in gp41 associated with VRC01 sensitivity or resistance)** all AA sites in gp41 associated with VRC01 sensitivity or resistance.

We additionally included the following groups of features:

Group 8 **(Indication of potential N-linked glycosylation sites)** a binary indicator for sites in all of Env that featured the canonical N-linked glycosylation motif in at least three of the analysis sequences and was absent from at least three of the analysis sequences;

Group 9 **(Majority virus subtypes)** including CRF01_AE, CRF02_AG, CRF07_BC, A1, A1C, A1D, B, C, D, O;

Group 10 **(Region-specific counts of PNG sites)** the number of PNG sites as defined by the canonical N-linked glycosylation motif;

Group 11 **(Viral geometry)** the total lengths of five different regions within the Env sequence known to be important for VRC01 binding: the entire Env protein, the gp120 protein, the V5 region, Loop D, and Loop E;

Group 12 **(Cysteine counts)** numbers of cysteines present within five different regions of the Env sequence: the entire Env protein, the gp120 protein, the V5 region, Loop D, and Loop E;

Group 13 **(Steric bulk at critical locations)** corresponding to the total number of small residues in the V5 region, Loop D, and in the CD4 binding loop; and

Group 14 geographic confounder variables.

2.4 Variable importance and the AMP statistical analysis plan

In Section 2.1, we noted that the AMP statistical analysis plan will include a pre-specified sieve analysis to further investigate differences in HIV-1 prevention efficacy based on characteristics of the exposing viruses. Thus, one of the goals of the analysis in Magaret et al. [2019] was to rank genotypic features by their importance for predicting TZM-bl neutralization sensitivity to VRC01. Based on this ranking, we will select the most important features to advance to the primary sieve analysis in AMP.

A rigorous definition of variable importance is required to meet this goal. In Chapter 3, we propose the nonparametric R^2-based variable importance measure that we used in Magaret et al. [2019]. In Chapter 4, we show that the measure studied in Chapter 3 lies within a class of useful variable importance measures that also includes the difference in area under the

receiver operating curve and classification accuracy. We replicate the analysis of Magaret et al. [2019] using these variable importance measures. In Chapter 5, we further extend the results of Chapters 3 and 4 to a new formulation of importance. Each of the proposed importance measures allows valid inference on the true, population-level variable importance of the subgroup of features. This is important for designing the statistical analysis plan for the AMP trials, since we can base our cutoff for inclusion in the pre-specified sieve analysis based on inferential results from the CATNAP data. In particular, if we identify groups of AA features that have importance estimated to be statistically significantly greater than zero, then including the superset of all such features in the sieve analysis appears to be a reasonable route forward.

2.5 Conclusion

In this dissertation, we focus on the bnAb VRC01 to prepare for the AMP trials. We develop methods for statistical inference on the true, population-level variable importance of subgroups of amino acid sequence features in predicting HIV-1 neutralization sensitivity. These methods allow flexible, machine learning-based algorithms to be used for prediction, while still yielding valid statistical inference. This appears to be a promising method towards yielding feature sets that can be tested for sieve effects in the AMP trials.

However, it is likely that combining multiple bnAbs in a regimen will yield improved prevention efficacy, similar to other combination regimens in HIV-1 treatment and prevention [see, e.g., Cohen et al., 2011]. Indeed, current research focuses on both combination regimens [see, e.g., Wagh et al., 2016] and bispecific antibodies [see, e.g., Wagh et al., 2018]. Thus, there is need for an analysis approach that handles these different regimens. We are currently developing a computational framework using combination outcomes [Wagh et al., 2016] that allows any antibody with data in CATNAP to be analyzed, and returns variable importance results.

Chapter 3

NONPARAMETRIC R^2-BASED VARIABLE IMPORTANCE ASSESSMENT

3.1 Introduction

Suppose that we have independent observations $O_1, \ldots, O_n$ drawn from an unknown distribution P_0, known only to lie in a potentially rich class of distributions $\mathcal{M}$. We refer to $\mathcal{M}$ as our model. Further suppose that each observation O_i consists of (X_i, Y_i), where $X_i := (X_{i1}, \ldots, X_{ip}) \in \mathbb{R}^p$ is a covariate vector and $Y_i \in \mathbb{R}$ is the outcome of interest. It is often of interest to understand the association between Y and X under P_0. To do this, we generally consider the conditional mean function μ_{P_0}, where for each $P \in \mathcal{M}$ we define

$$\mu_P(x) := E_P(Y \mid X = x) \ . \tag{3.1}$$

Estimation of μ_{P_0} is the canonical 'predictive modeling' problem. There are many tools for estimating μ_{P_0}: classical parametric techniques (e.g., linear regression), and more flexible nonparametric or semiparametric methods, including random forests [Breiman, 2001a], generalized additive models [Hastie and Tibshirani, 1990], loess smoothing [Cleveland, 1979], and artificial neural networks [Barron, 1989], among many others. Once a good estimate of μ_{P_0} is obtained, it is often of scientific interest to identify the features that contribute most to the variation in μ_{P_0}. For any given set $s \subseteq \{1, \ldots, p\}$ and distribution $P \in \mathcal{M}$, we may define the reduced conditional mean

$$\mu_{P,-s}(x) := E_P\left(Y \mid X_{-s} = x_{-s}\right) \ , \tag{3.2}$$

where for any vector v and set r of indices the symbol v_{-r} denotes the vector of all components of v with index not in r. Here, the set s can represent a single element or a group of elements. The importance of the elements in s can be evaluated by comparing μ_{P_0} to $\mu_{P_0,-s}$. This strategy will be leveraged in this paper.

The ANOVA decomposition is the main classical tool for evaluating variable importance. There, μ_{P_0} is assumed to have a simple parametric form. While this facilitates the task at hand considerably, the conclusions drawn can be misleading in view of the high risk of model misspecification. For this reason, it is increasingly common to use either nonparametric or machine learning-based regression methods, or both, to estimate μ_{P_0}; in such cases, classical ANOVA results do not apply.

Recent work on evaluating variable importance without relying on overly strong modeling assumptions can generally be categorized as being either (i) intimately tied to a specific estimation technique for the conditional mean function or (ii) agnostic to the estimation technique used. The former category includes the variable importance measures for random forests [Breiman, 2001a, Strobl et al., 2007, Ishwaran, 2007, Grömping, 2009, Nicodemus et al., 2010] and neural networks [see, e.g., Olden et al., 2004], and ANOVA in linear models. Among these, ANOVA alone appears to allow valid statistical inference. Additionally, it is generally not possible to directly compare the importance assessment stemming from different methods: they usually measure different quantities and thus have different interpretations. The latter category includes nonparametric extensions of R^2 for kernel-based estimators, local polynomial regression, and functional regression [Doksum and Samarov, 1995, Yao et al., 2005, Huang and Chen, 2008]; the marginalized mean difference, $E_{P_0}\{E_{P_0}(Y \mid X = x^1, W) - E_{P_0}(Y \mid X = x^0, W)\}$ [van der Laan, 2006, Chambaz et al., 2012, Sapp et al., 2014] where x^1 and x^0 are two meaningful reference levels of X; and the mean absolute difference, $E_{P_0}\{|Y - E_{P_0}(Y \mid X)| - |Y - E_{P_0}(Y \mid X_{-j})|\}$ [Lei et al., 2017]. Methods in this latter category allow valid inference and have broad potential applicability. The appropriate measure to use depends on the scientific context.

In our view, an ideal variable importance measure should (i) be entirely agnostic to the

estimation technique, (ii) allow valid inference, and (iii) provide an interpretation that is well suited to scientific applications. In this work, we study a variable importance measure that satisfies each of these criteria, adding to the class of technique-agnostic measures referenced above. In particular, we consider the ANOVA-based variable importance measure

$$\psi_{0,s} := \frac{\int \{\mu_{P_0}(x) - \mu_{P_0,-s}(x)\}^2 \, dP_0(x)}{var_{P_0}(Y)} \ . \tag{3.3}$$

For a vector v and a subset r of indices, we denote by v_r the vector of all components of v with index in r. Then, we may interpret (3.3) as the additional proportion of variability in the outcome explained by including X_s in the conditional mean. This follows from the fact that we can express $\psi_{0,s}$ as

$$\left(1 - \frac{E_{P_0}\left[\{Y - \mu_{P_0}(X)\}^2\right]}{var_{P_0}(Y)}\right) - \left(1 - \frac{E_{P_0}\left[\{Y - \mu_{P_0,-s}(X)\}^2\right]}{var_{P_0}(Y)}\right) \ ,$$

the difference in the population R^2 obtained using the full set of covariates as compared to the reduced set of covariates only. Thus, the parameter we focus on is a simple generalization of the classical R^2 measure of importance to a nonparametric model and is useful in any setting in which the mean squared error is a scientifically relevant population measure of predictiveness. This parameter is a function of P_0 alone, in that it describes a property of the true data-generating mechanism and not of any particular estimation method. In this work, we provide a framework for building a nonparametric efficient estimator of $\psi_{0,s}$ that permits valid statistical inference.

The purpose of the variable importance measure we study here is *not* to offer insight into the performance of any particular algorithm, but rather to describe the importance of variables in explaining the outcome in the population. This is in contrast to common algorithm-specific measures of variable importance. If an algorithm-specific diagnostic is desired, other approaches to variable importance may be preferred, as referenced above.

Care must be taken in building point and interval estimators for $\psi_{0,s}$ when μ_{P_0} and $\mu_{P_0,-s}$

are not known to belong to simple parametric families. In particular, when μ_{P_0} and $\mu_{P_0,-s}$ are estimated using flexible methods, simply plugging estimates of these regression functions into (3.3) will not yield a regular and asymptotically linear, let alone efficient, estimator of $\psi_{0,s}$. In this chapter, we propose a simple method that, given sufficiently accurate estimators of μ_{P_0} and $\mu_{P_0,-s}$, yields an efficient point estimator for $\psi_{0,s}$ and a confidence interval with asymptotically correct coverage. The approach we employ is based on ideas from the theory of semiparametric estimation and inference. We show this corrected estimator is equivalent to a plug-in estimator based on the difference in R^2 values. The population R^2 can be estimated using a sample splitting approach [see, e.g., Hastie et al., 2009].

While variable importance is related to variable selection, we stress that these two paradigms have different goals. The goal in variable selection is typically to create the best predictive model based on the current data; this model may include only a subset of the available variables. There are many contributions in both technique-specific selection [see, e.g., Breiman et al., 1984, Breiman, 2001a, Friedman, 2001, Loh, 2002] and in nonparametric selection [see, e.g., Doksum et al., 2008]. The goal in variable importance is to assess the extent to which (subsets of) features contribute to improving the population-level predictive power of the best possible outcome predictor based on all available features. To highlight the distinction between importance and selection, it is useful to consider a scenario in which two perfectly correlated covariates X_1 and X_2 are available. Neither covariate has importance relative to the other, but the pair of variables is important. A variable importance procedure considering individual and grouped features will identify this; a variable selection procedure will likely choose only one of X_1 or X_2 for use in prediction.

We present some properties of our parameter of interest and give our proposed estimator in Section 3.2. In Section 3.3, we provide empirical evidence that our proposed estimator outperforms both the naive plug-in ANOVA-based estimator and an ordinary least squares-based estimator in settings where the covariate vector is low- or moderate-dimensional and the data-generating mechanism is nonlinear. In Section 3.4, we apply our method on data from a retrospective study of heart disease in South African men. We provide concluding

remarks in Section 3.5. Technical details and an illustration of our method in the context of the landmark Boston housing data are provided in Appendix A.

3.2 Variable importance in a nonparametric model

3.2.1 Parameter of interest

We work in a fully unrestricted, and thus nonparametric, model $\mathcal{M}$. For given $s \subseteq \{1, \ldots, p\}$ and $P \in \mathcal{M}$, we use the conditional means (3.1) and (3.2) to define the statistical functional

$$\Psi_s(P) := \frac{\int \{\mu_P(x) - \mu_{P,-s}(x)\}^2 \, dP(x)}{var_P(Y)}; \tag{3.4}$$

this is the nonparametric measure of variable importance we focus on. The value of $\Psi_s(P)$ measures the importance of variables in the set $\{X_j\}_{j \in s}$ relative to the entire covariate vector. Using observations $O_1, \ldots, O_n$ independently drawn from the true, unknown joint distribution $P_0 \in \mathcal{M}$, we aim to make efficient inference about the true value $\psi_{0,s} = \Psi_s(P_0)$.

This parameter is a nonparametric extension to the usual ANOVA-derived measure of variable importance in parametric models. We first note that $\psi_{0,s} \in [0, 1]$. Furthermore, $\psi_{0,s} = 0$ if and only if Y is conditionally uncorrelated with every transformation of X_s given X_{-s}. In addition, $\psi_{0,s}$ is a ratio: the numerator of (3.4) is the *extra sum of squares*, averaged over the marginal distribution of the features; the denominator is the *total sum of squares*. Thus, the value of $\psi_{0,s}$ is precisely the improvement in predictive performance, in terms of standardized mean squared error, that we can expect if we build a model using all of X versus X_{-s}. If we assume simple linear regression models for μ_{P_0} and $\mu_{P_0,-s}$, then $\psi_{0,s}$ is precisely the usual difference in R^2 between nested models. The parameter $\psi_{0,s}$ is also invariant to linear transformations of the outcome and invertible transformations of each feature. We provide a proof of this result and discuss its implications in Appendix A.

We want to reiterate here that, in contrast to simple parametric approaches to variable importance, our functional Ψ_s simply maps any candidate data-generating mechanism to a positive number. This definition does not require a parametric specification of μ_P or $\mu_{P,-s}$.

While this is usual for non- or semiparametric inference problems, it is quite different from classical approaches to variable importance.

We now discuss some properties of Ψ_s that are relevant to building an efficient estimator of $\psi_{0,s}$. Specifically, functional (3.4) is pathwise differentiable [see, e.g., Bickel et al., 1998]. Pathwise differentiable functionals generally admit a convenient functional Taylor expansion that can be used to characterize the asymptotic behavior of plug-in estimators based on the functional. An analysis of the pathwise derivative allows us to determine the efficient influence function (EIF) of the functional relative to the statistical model [Bickel et al., 1998]. The EIF plays a key role in establishing efficiency bounds for regular and asymptotically linear estimators of the true parameter value, and most importantly, in the construction of efficient estimators, as we will highlight below. For convenience, we will denote the numerator of $\Psi_s(P)$ by $\Phi_s(P) := \int \{\mu_P(x) - \mu_{P,-s}(x)\}^2 dP(x)$. The EIF of Φ_s and of Ψ_s relative to $\mathcal{M}$ are provided in the following lemma.

Lemma 1. *The parameters Φ_s and Ψ_s are pathwise differentiable at each $P \in \mathcal{M}$ relative to $\mathcal{M}$, with efficient influence functions $D_{P,s}$ and $D^*_{P,s}$ relative to $\mathcal{M}$ respectively given by*

$$\begin{aligned} o \mapsto D_{P,s}(o) &:= 2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,-s}(x)\} + \{\mu_P(x) - \mu_{P,-s}(x)\}^2 - \Phi_s(P) \ , \\ o \mapsto D^*_{P,s}(o) &:= \frac{2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,-s}(x)\} + \{\mu_P(x) - \mu_{P,-s}(x)\}^2}{var_P(Y)} \\ &\quad - \Phi_s(P) \left\{ \frac{y - E_P(Y)}{var_P(Y)} \right\}^2 . \end{aligned}$$

The evaluation of Φ_s at $P \in \mathcal{M}$ can be expressed as

$$\Phi_s(P) = \Phi_s(P_0) + \int D_{P,s}(o) d(P - P_0)(o) + R_s(P, P_0) \ , \tag{3.5}$$

where $R_s(P, P_0)$ is a remainder term from this first-order expansion around P_0. The explicit form of $R_s(P, P_0)$ is provided in Section 3.2.3 and can be used to algebraically verify this

representation. For any given estimator $\widehat{P}_n \in \mathcal{M}$ of P_0, we can write

$$\begin{aligned}
\Phi_s(\widehat{P}_n) - \Phi_s(P_0) &= \int D_{\widehat{P}_n,s}(o) d(\widehat{P}_n - P_0)(o) + R_s(\widehat{P}_n, P_0) \\
&= \int D_{\widehat{P}_n,s}(o) d(\mathbb{P}_n - P_0)(o) + R_s(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n} D_{\widehat{P}_n,s}(O_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} D_{P_0,s}(O_i) + \int \left\{ D_{\widehat{P}_n,s}(o) - D_{P_0,s}(o) \right\} d(\mathbb{P}_n - P_0)(o) \qquad (3.6) \\
&+ R_s(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n} D_{\widehat{P}_n,s}(O_i),
\end{aligned}$$

where $\mathbb{P}_n$ is the empirical distribution based on $O_1, \ldots, O_n$, and we have made repeated use of the fact that $D_{P,s}(O)$ has mean zero under P for any $P \in \mathcal{M}$. This representation is critical for characterizing the behavior of the plug-in estimator $\Phi_s(\widehat{P}_n)$. The four terms on the right-hand side in (3.6) can be studied separately. The first term is an empirical average of mean-zero transformations of $O_1, \ldots, O_n$. The second term is an empirical process term, and the third term is a second-order remainder term. Both can be shown to be asymptotically negligible under certain conditions on $\widehat{P}_n$. The fourth term can be thought of as the bias incurred from flexibly estimating the conditional means (3.1) and (3.2), and in general, it will tend to zero slowly. This bias term motivates our choice of estimator for $\psi_{0,s}$ in Section 3.2.2. We will choose one particular method of correcting for this bias term, and the large-sample properties of our proposed estimator will be determined by the first term in (3.6).

3.2.2 Estimation procedure

Writing the numerator Φ_s of the parameter of interest as a statistical functional suggests a natural estimation procedure. If we have estimators $\hat{\mu}$ and $\hat{\mu}_s$ of μ_{P_0} and $\mu_{P_0,-s}$, respectively – obtained through any method that we choose, including machine learning techniques – a

natural plug-in estimator of $\phi_{0,s} := \Phi_s(P_0)$ is given by

$$\hat{\phi}_{\text{naive},s} := \int \{\hat{\mu}(x) - \hat{\mu}_{-s}(x)\}^2 \, d\mathbb{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \{\hat{\mu}(X_i) - \hat{\mu}_{-s}(X_i)\}^2 \ . \tag{3.7}$$

In turn, this suggests using, with $\bar{Y}_n$ denoting the empirical mean of $Y_1, \ldots, Y_n$,

$$\hat{\psi}_{\text{naive},s} := \frac{\hat{\phi}_{\text{naive},s}}{var_{\mathbb{P}_n}(Y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(X_i) - \hat{\mu}_{-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}$$

as a simple estimator of $\psi_{0,s}$. We refer to this as the *naive* estimator, because this simple estimator involves hidden tradeoffs. On one hand, it is easy to construct given the estimators $\hat{\mu}$ and $\hat{\mu}_{-s}$. On the other hand, the naive estimator does not generally enjoy good inferential properties. If a flexible technique is used to estimate μ_{P_0} and $\mu_{P_0,-s}$, constructing $\hat{\mu}$ and $\hat{\mu}_{-s}$ usually entails selecting tuning parameter values to achieve an optimal bias-variance tradeoff for μ_{P_0} and $\mu_{P_0,-s}$, respectively. This is generally not the optimal bias-variance tradeoff for estimating the parameter of interest $\psi_{0,s}$, a key fact from non- and semiparametric theory. The plug-in estimator $\hat{\psi}_{\text{naive},s}$ inherits much of the bias from $\hat{\mu}$ and $\hat{\mu}_s$, but its variance has a parametric rate with little sensitivity to the tuning of $\hat{\mu}$ and $\hat{\mu}_s$ because of the involved marginalization over the feature distribution. Some form of debiasing is thus needed, and the one-step correction achieves this by using the empirical average of the estimated EIF as a first-order approximation of this excess bias. In particular, the estimator $\hat{\psi}_{\text{naive},s}$ is generally overly biased, in the sense that its bias does not tend to zero sufficiently fast to allow consistency at rate $n^{-1/2}$, let alone efficiency. This is problematic, in particular, because it renders the construction of valid confidence intervals extremely difficult, if not impossible.

We propose to use the simple corrected estimator

$$\hat{\phi}_{n,s} := \hat{\phi}_{\text{naive},s} + \frac{1}{n}\sum_{i=1}^{n} D_{\widehat{P}_n,s}(O_i)$$

of $\phi_{0,s}$, which, in view of (3.6), will be asymptotically efficient under certain regularity conditions. This estimator, which is often referred to as the one-step estimator [see, e.g., Pfanzagl, 1982], is obtained by correcting for the excessive bias of the naive plug-in estimator $\hat{\phi}_{\text{naive},s}$. We note that to compute $\hat{\phi}_{n,s}$ it is not necessary to obtain an estimator $\widehat{P}_n$ of the entire distribution P_0. Instead, one only needs estimators $\hat{\mu}$ and $\hat{\mu}_{-s}$ of μ_{P_0} and $\mu_{P_0,-s}$. As indicated before, the variance of Y under P_0 may simply be estimated using the empirical variance. It is easy to verify algebraically that the resulting estimator of $\psi_{0,s}$ simplifies to

$$\hat{\psi}_{n,s} \;=\; \frac{\hat{\phi}_{n,s}}{var_{\mathbb{P}_n}(Y)} \;=\; \hat{\psi}_{\text{naive},s} + \frac{\sum_{i=1}^n 2\{Y_i - \hat{\mu}(X_i)\}\{\hat{\mu}(X_i) - \hat{\mu}_{-s}(X_i)\}}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} \,. \tag{3.8}$$

This estimator adjusts for the inadequate bias-variance tradeoff performed when flexible estimators $\hat{\mu}$ and $\hat{\mu}_{-s}$ are tuned to be good estimators of μ_{P_0} and $\mu_{P_0,-s}$ rather than being tuned for the end objective of estimating $\psi_{0,s}$. Simple algebraic manipulations yield that $\hat{\psi}_{n,s}$ is equivalent to the plug-in estimator

$$\hat{\psi}_{n,s,*} := \left[1 - \frac{\frac{1}{n}\sum_{i=1}^n \{Y_i - \hat{\mu}(X_i)\}^2}{var_{\mathbb{P}_n}(Y)}\right] - \left[1 - \frac{\frac{1}{n}\sum_{i=1}^n \{Y_i - \hat{\mu}_{-s}(X_i)\}^2}{var_{\mathbb{P}_n}(Y)}\right] \tag{3.9}$$

obtained by viewing $\psi_{0,s}$ as a difference in population R^2 values. Semiparametric theory indicates that plug-in estimators based on flexible regression algorithms typically require correction if the latter are not tuned towards the target of inference, as in (3.8); interestingly, this is not needed for $\hat{\psi}_{n,s,*}$.

While we are not constrained to any particular estimation method to construct $\hat{\mu}$ and $\hat{\mu}_{-s}$, we have found one particular strategy to work well in practice. Using any specific predictive modeling technique to regress the outcome Y on the full covariate vector X and then on the reduced vector X_{-s} of covariates does not take into account that the two conditional means are related, and will generally result in incompatible estimates. Specifically, we have that $E_{P_0}(Y \mid X_{-s}) = E_{P_0}\{E_{P_0}(Y \mid X) \mid X_{-s}\}$, which we can take advantage of to produce the following sequential regression estimating procedure: (1) regress Y on X to obtain an

estimate $\hat{\mu}$ of μ_{P_0}; then (2) regress $\hat{\mu}(X)$ on X_{-s} to obtain an estimate $\hat{\mu}_{-s}$ of $\mu_{P_0,-s}$.

The final estimation procedure we recommend for $\psi_{0,s}$ consists of estimator $\hat{\psi}_{n,s}$, or equivalently $\hat{\psi}_{n,s,*}$, where the conditional means involved are estimated using flexible regression estimators and this sequential regression approach; see Algorithm 1 for more details. This may also be embedded in a split-sample validation scheme: we first create a training and validation set; then obtain $\hat{\mu}$ and $\hat{\mu}_{-s}$ on the training set as outlined above; and finally, obtain an estimator of $\psi_{0,s}$ by using the validation data along with predictions from the conditional mean estimators on the validation data. This can be extended to a cross-validated procedure given in Algorithm 2, which is discussed more extensively in Appendix A.

Algorithm 1 Estimate $\psi_{0,s}$

1: Choose a technique to estimate the conditional means μ_{P_0} and $\mu_{P_0,-s}$, e.g., ensemble learning with various predictive modeling algorithms [Wolpert, 1992];
2: $\hat{\mu} \leftarrow$ Regress Y on X using the technique from step (1) to estimate μ_{P_0};
3: $\hat{\mu}_{-s} \leftarrow$ Regress $\hat{\mu}(X)$ on X_{-s} using the technique from step (1) to estimate $\mu_{P_0,-s}$;
4: $\hat{\psi}_{n,s} \leftarrow \frac{\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(X_i)-\hat{\mu}_{-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2} + \frac{\sum_{i=1}^{n} 2\{Y_i-\hat{\mu}(X_i)\}\{\hat{\mu}(X_i)-\hat{\mu}_{-s}(X_i)\}}{\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2}$, as in Equation (3.8).

Algorithm 2 Estimate $\psi_{0,s}$ using V-fold sample splitting

1: Choose a technique to estimate the conditional means μ_{P_0} and $\mu_{P_0,-s}$;
2: **for** $v = 1, \ldots, V$ **do**
3: Generate a random vector $\mathcal{I} := \{0,1\}^n$; $\mathcal{T} \leftarrow \{i : \mathcal{I}_i = 0\}$ and $\mathcal{V} \leftarrow \{i : \mathcal{I}_i = 1\}$;
4: $\hat{\mu}_v \leftarrow$ Regress Y on X in $\mathcal{T}$ using the technique from step (1) to estimate μ_{P_0};
5: $\hat{\mu}_{-s,v} \leftarrow$ Regress $\hat{\mu}_v(X)$ on X_{-s} in $\mathcal{T}$ to estimate $\mu_{P_0,-s}$;
6: $\hat{\psi}^v_{n,s,*} \leftarrow \frac{\sum_{i\in\mathcal{V}}\{Y_i-\hat{\mu}_{-s,v}(X_i)\}^2-\sum_{i\in\mathcal{V}}\{Y_i-\hat{\mu}_v(X_i)\}^2}{\sum_{i\in\mathcal{V}}(Y_i-\bar{Y}_n)^2}$, as in Equation (3.9);
7: **end for**
8: $\hat{\psi}^{\mathrm{cv}}_{n,s,*} \leftarrow \frac{1}{V}\sum_{v=1}^{V}\hat{\psi}^v_{n,s,*}$.

3.2.3 Asymptotic behavior of the proposed estimator

By studying the remainder term $R_s(\widehat{P}_n, P_0)$ and the empirical process term from (3.6), we can establish appropriate conditions on $\hat{\mu}$ and $\hat{\mu}_{-s}$ as estimators of μ_{P_0} and $\mu_{P_0,-s}$ under

which the proposed estimator $\hat{\psi}_{n,s}$ is asymptotically efficient. This allows us to determine the asymptotic distribution of the proposed estimator, and therefore, to propose procedures for performing valid inference on $\psi_{0,s}$. The first result we present establishes the explicit form of $R_s(P, P_0)$ for any $P \in \mathcal{M}$ and sufficient conditions on $\hat{\mu}$ and $\hat{\mu}_{-s}$ that guarantee that $R_s(\widehat{P}_n, P_0)$ is asymptotically negligible.

Lemma 2. *Linearization* (3.5) *holds with second-order remainder term given by*

$$R_s(P, P_0) = \int \{\mu_{P_0,-s}(x) - \mu_{P,-s}(x)\}^2 \, dP_0(x) - \int \{\mu_{P_0}(x) - \mu_P(x)\}^2 \, dP_0(x)$$

. Also, $R_s(\widehat{P}_n, P_0) = o_P(n^{-1/2})$ *if* $\int\{\hat{\mu}(x)-\mu_{P_0}(x)\}^2 dP_0(x)$ *and* $\int\{\hat{\mu}_{-s}(x)-\mu_{P_0,-s}(x)\}^2 dP_0(x)$ *are* $o_P(n^{-1/2})$.

The remainder term is a sum of several terms, each of which tend to zero as sample size grows. Each of these second-order terms can feasibly be made to be $o_P(n^{-1/2})$, even while using flexible regression techniques, including generalized additive models [Hastie and Tibshirani, 1990], to estimate the conditional mean functions.

The second result we present establishes conditions under which the empirical process term appearing in (3.6) is asymptotically negligible. One condition involves a P_0-Donsker class [van der Vaart, 2000], a fundamental object in semiparametric efficiency theory.

Lemma 3. *Provided* $\int\{\hat{\mu}(x) - \mu_{P_0}(x)\}^2 dP_0(x)$ *and* $\int\{\hat{\mu}_{-s}(x) - \mu_{P_0,-s}(x)\}^2 dP_0(x)$ *both tend to zero in probability, and* $o \mapsto D_{\widehat{P}_n,s}(o)$ *falls in a* P_0*-Donsker class with probability tending to one, it holds that* $\int\{D_{\widehat{P}_n,s}(o) - D_{P_0,s}(o)\}d(\mathbb{P}_n - P_0)(o) = o_P(n^{-1/2})$.

This empirical process term is negligible under rather weak conditions. Uniform consistency of $\hat{\mu}$ and $\hat{\mu}_{-s}$ suffices without the need for minimal rates of convergence. The additional Donsker class condition requires that the set of possible realizations of $\hat{\mu}$ and $\hat{\mu}_{-s}$ become sufficiently restricted with probability tending to one as sample size grows. This condition is satisfied if, for example, the uniform sectional variation norm [Gill et al., 1995] of $D_{\widehat{P}_n,s}$ is bounded with probability tending to one. When using flexible machine learning-based

regression estimators, there may be reason for concern regarding the validity of the Donsker class condition. In such cases, using the cross-validated estimator $\hat{\psi}^{\text{cv}}_{n,s,*}$ may circumvent this condition. While this cross-validated estimator is only marginally more complex than the estimator proposed here, we restrict attention here to studying the simpler estimator, and leave study of the cross-validated estimator to Appendix A.

The following theorem builds upon these two lemmas to describe the asymptotic behavior of the proposed estimator. Even if the cross-validated estimator is used, the regression estimators must satisfy the requirements of Lemma 2 for this result to hold.

Theorem 1. *Suppose that both* $\int\{\hat{\mu}(x)-\mu_{P_0}(x)\}^2 dP_0(x)$ *and* $\int\{\hat{\mu}_{-s}(x)-\mu_{P_0,-s}(x)\}^2 dP_0(x)$ *are* $o_P(n^{-1/2})$*, and that* $o \mapsto D_{\widehat{P}_{n,s}}(o)$ *falls in a* P_0*-Donsker class with probability tending to one. Then, the proposed estimator* $\hat{\psi}_{n,s}$ *is asymptotically linear with influence function* $D^*_{P_0,s}$*. In particular, this implies that (a)* $\hat{\psi}_{n,s}$ *tends to* $\psi_{0,s}$ *in probability, (b)* $\hat{\psi}_{n,s}$ *is regular, and if* $\psi_{0,s} \in (0,1)$*, (c)* $n^{1/2}(\hat{\psi}_{n,s}-\psi_{0,s})$ *tends in distribution to a mean-zero Gaussian random variable with variance* $\sigma^2_{0,s} := \int\{D^*_{P_0,s}(o)\}^2 dP_0(o)$*.*

A natural plug-in estimator of $\sigma_{0,s}$ is given by

$$\hat{\sigma}_{n,s} := \left[\frac{1}{n}\sum_{i=1}^{n}\{\widehat{D}^*_{P_0,s}(O_i)\}^2\right]^{1/2},$$

where $\widehat{D}^*_{P_0,s}$ is any consistent estimator of $D^*_{P_0,s}$. For example, $\widehat{D}^*_{P_0,s}$ may be taken to be $D^*_{P_0,s}$ with μ_{P_0}, $\mu_{P_0,-s}$, $E_{P_0}(Y)$, $var_{P_0}(Y)$ and $\phi_{0,s}$ replaced by $\hat{\mu}$, $\hat{\mu}_{-s}$, $\bar{Y}_n$, $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2$ and $\hat{\phi}_{n,s}$, respectively. In view of the asymptotic normality of $n^{1/2}(\hat{\psi}_{n,s}-\psi_{0,s})$, an asymptotically valid $(1-\alpha)\times 100\%$ Wald-type confidence interval for $\psi_{0,s}$ can be obtained as $\hat{\psi}_{n,s} \pm q_{1-\alpha/2}\hat{\sigma}_{n,s}n^{-1/2}$, where q_β is the β-quantile of the standard normal distribution.

When flexible (potentially machine learning-based) estimators of the involved regression are used, the naive estimator is generally not asymptotically linear: it will usually be overly biased, resulting in a rate of convergence slower than $n^{-1/2}$. Constructing valid confidence intervals based on the naive estimator may therefore be extremely difficult, if not impossible.

It may be tempting to adopt a bootstrap approach as remedy. However, this would not be advisable since, besides the computational burden of such an approach, there is little theoretical justification for using the nonparametric bootstrap in this context [Shao, 1994]. Alternatively, a more sophisticated subsampling approach (e.g., m out of n bootstrap) may be valid [Politis et al., 1999], though such algorithms can be difficult to appropriately tune in practice [see, e.g., Samworth, 2003].

3.2.4 Behavior under the zero-importance null hypothesis

This work primarily focuses on developing an efficient estimator of a variable importance measure that is a property of the true data-generating mechanism using flexible estimation techniques, and on describing how valid inference may be drawn when the set s of features under evaluation does not have degenerate importance. Specifically, we have restricted our attention to cases in which $\psi_{0,s} \in (0, 1)$ strictly. It may be of interest, however, to test the null hypothesis $\psi_{0,s} = 0$ of zero importance. Developing valid inference under this particular null hypothesis appears very difficult. Because the null hypothesis is on the boundary of the parameter space, $D_{P_0,s}$ is identically zero under this null, and it is likely that a higher-order expansion must be used to construct and characterize the behavior of an appropriately-regularized estimator of $\phi_{0,s}$ and thus of $\psi_{0,s}$. However, the parameters Φ_s and Ψ_s are generally not even second-order pathwise differentiable, and so, higher-order expansions cannot easily be constructed. There may be hope in using approximate second-order gradients, as outlined in Carone et al. [2018], though this remains an open problem. To highlight the difficulties that arise under this particular null hypothesis, we conducted a simulation study for a setting in which one of the variables has zero importance. The results from this study are provided in the next section.

3.3 Experiments on simulated data

We now present empirical results describing the performance of the proposed estimator (3.8) compared to that of the naive plug-in estimator (3.7). In all implementations, we use

the sequential regression estimating procedure described in Algorithm 1 for each feature or group of interest to compute compatible estimates of the required regression functions, and we compute nominal 95% Wald-type confidence intervals as outlined in Section 3.2.3.

3.3.1 Low-dimensional vector of features

We consider here data generated according to the following specification:

$$X_1, X_2 \overset{iid}{\sim} \text{Uniform}(-1, 1) \text{ and } \epsilon \sim N(0, 1) \text{ independent of } (X_1, X_2)$$
$$Y = X_1^2 \left(X_1 + \tfrac{7}{5}\right) + \tfrac{25}{9} X_2^2 + \epsilon \ .$$

We generated $1,000$ random datasets of size $n \in \{100, 300, 500, 700, 1000, 2000, \ldots, 8000\}$, and considered in each case the importance of X_j for $j \in \{1, 2\}$. The true value of the variable importance measures implied by this data-generating mechanism can be shown to be $\psi_{0,1} \approx 0.158$ and $\psi_{0,2} \approx 0.342$. This nonlinear setting helps to highlight the drawbacks of relying on a simple parametric model to estimate the conditional means.

To obtain $\hat{\mu}$ and each $\hat{\mu}_{-j}$, we fit locally-constant loess smoothing using the `R` function `loess` with tuning selected to minimize a five-fold cross-validated estimate of the empirical risk based on the squared error loss function. Loess smoothing was chosen because it is a data-adaptive algorithm with an efficient implementation, and it satisfies the minimum convergence rate condition outlined in Section 3.2.3, allowing us to numerically verify our theoretical results. Because we obtained the same trends using locally-constant kernel regression, we do not report summaries from these additional simulations here. This fact nevertheless highlights the ease of comparing results from two different estimation techniques.

We computed the naive and proposed estimators and respective confidence intervals for each replication, and compared these to a parametric difference in R^2 based on simple linear regression using ordinary least squares (OLS). Because a simple asymptotic distribution for the naive estimator is unavailable, a percentile bootstrap approach with 1,000 bootstrap samples was used in an attempt to obtain approximate confidence intervals based on $\hat{\psi}_{\text{naive},j}$.

For each estimator, we then computed the empirical bias scaled by $n^{1/2}$ and the empirical variance scaled by n. Our output for the estimated bias includes confidence intervals for the true bias based on the resulting draws from the bootstrap sampling distribution. Finally, we computed the empirical coverage of the nominal 95% confidence intervals constructed.

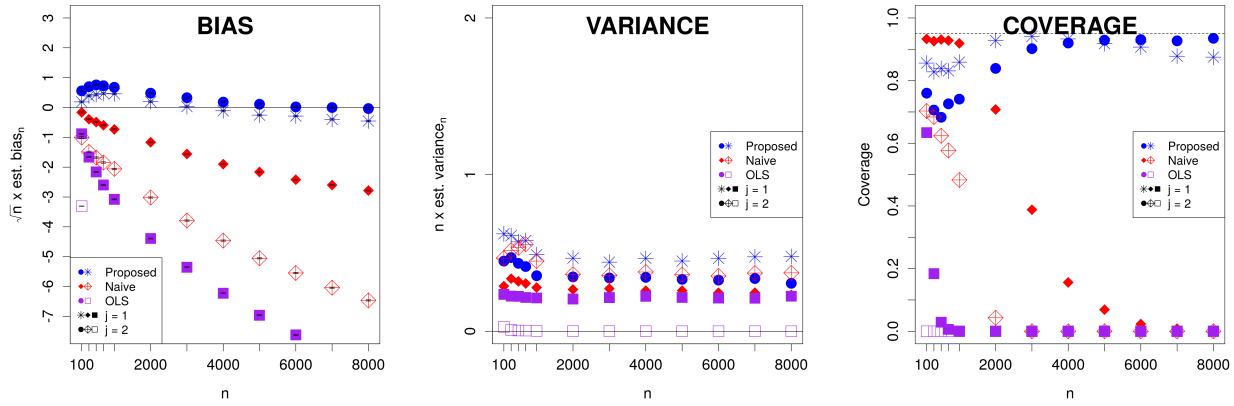

Figure 3.1: Empirical bias scaled by $\sqrt{n}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed, naive, and OLS estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to estimate (3.1) and (3.2) in the case of the proposed and naive estimators. Circles, filled diamonds, and filled squares denote that we have removed X_1; stars, crossed diamonds, and empty squares denote that we have removed X_2. This figure appears in color in the electronic version of this chapter.

Figure 3.1 displays the results of this simulation. In the left panel, we note that the scaled empirical bias of the proposed estimator decreases towards zero as n tends to infinity, regardless of which feature we remove. Also, we see that both the naive estimator and the OLS estimator have substantial bias that does not tend to zero faster than $n^{-1/2}$. This coincides with our expectations: the naive estimator involves an inadequate bias-variance tradeoff with respect to the parameter of interest and does not access an additional quantity to correct for this fact; the OLS estimator has a misspecified mean model. Though there is very substantial bias reduction from using the proposed estimator, we see that its scaled bias appears to dip slightly below zero for large n. We expect for larger n to see this scaled bias for the proposed estimator get closer to zero; numerical error in our computations may

explain why this does not exactly happen. These results provide empirical evidence that either a correction or using formulation (3.9) is necessary to account for the slow rates of convergence in estimation of $\psi_{0,s}$ introduced because μ_{P_0} and $\mu_{P_0,-s}$ are flexibly estimated.

In the middle panel of Figure 3.1, we see that the variance of the proposed estimator is close to that of the naive estimator – we have thus not suffered much at all from removing excess bias in our estimation procedure. The variance of the OLS estimator is the smallest of the three: using a parametric model tends to result in a smaller variance estimate. The ratio of the variance of the naive estimator to the variance of the proposed estimator is near one for all n considered, and ranges between approximately 0.8 and 1.2 in our simulation study. Finally, in the right-hand panel, we see that as sample size grows, coverage increases for the confidence interval based on the proposed estimator and approaches the nominal level. In contrast, the coverage of intervals based on both the naive estimator and the OLS estimator decreases instead and very quickly becomes completely unsatisfactory.

3.3.2 Testing the zero-importance null hypothesis

We now consider data generated according to the following specification:

$$X_1, X_2 \overset{iid}{\sim} \text{Uniform}(-1, 1) \text{ and } \epsilon \sim N(0, 1) \text{ independent of } (X_1, X_2);\ Y = \tfrac{25}{9} X_1^2 + \epsilon.$$

We generated 1,000 random datasets of size $n \in \{100, 300, 500, 700, 1000, 2000, 3000\}$ and again considered in each case the importance of X_j for $j \in \{1, 2\}$. The true value of the variable importance measures implied by this data-generating mechanism can be shown to be $\psi_{0,1} \approx 0.407$ and $\psi_{0,2} = 0$. We estimated the conditional means and summarized the results of this simulation as in the previous simulation.

Figure 3.2 displays the results of this simulation. In the left-hand panel, we observe that the proposed estimator has smaller scaled bias in magnitude than the naive estimator when we remove the feature with nonzero importance. However, when we remove the feature with zero importance, the proposed estimator has slightly higher bias. While this is somewhat

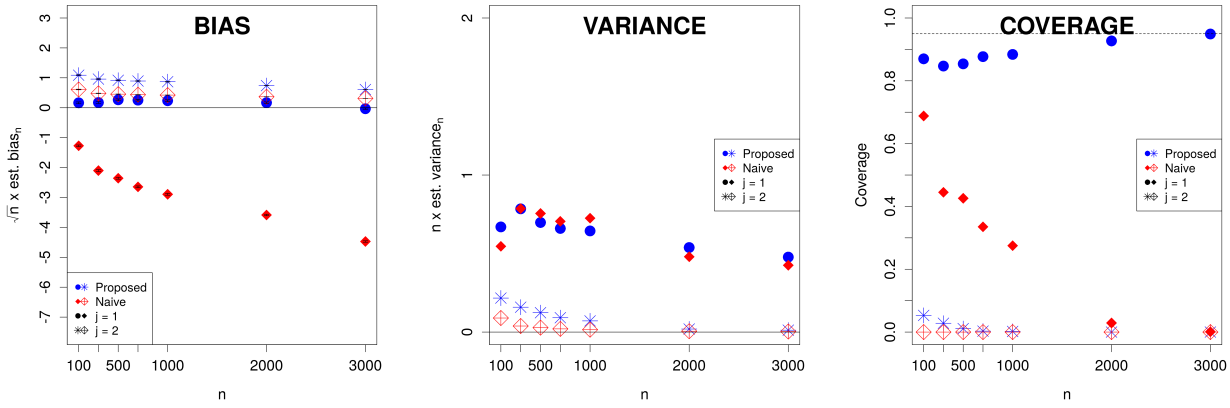

Figure 3.2: Empirical bias scaled by $\sqrt{n}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to estimate (3.1) and (3.2). Circles and filled diamonds denote that we have removed X_1, while stars and crossed diamonds denote that we have removed X_2. We operate under the null hypothesis for X_2; $\psi_{0,2} = 0$. This figure appears in color in the electronic version of this chapter.

surprising, it likely is due to the additive correction in the one-step construction being slightly too large. The scaled bias of the proposed estimator, regardless of j, tends to zero as n increases, which is not true of the naive estimator. In the middle panel, we see that we have not incurred excess variance by using the proposed estimator. In the right-hand panel, we see that both estimators have close to zero coverage for the parameter under the null hypothesis, but that the proposed estimator has higher coverage than the naive estimator for the predictive feature. These results highlight that more work needs to be done for valid testing and estimation under this boundary null hypothesis. While our current proposal yields valid results for the predictive feature, even in the presence of a null feature, ensuring valid inference for null features themselves remains an important challenge.

3.3.3 Moderate-dimensional vector of features

We consider one setting in which the features are independent and a second in which groups of features are correlated. In setting A, we generate data according to the following specifi-

cation:

$$X_1, X_2, \ldots, X_{15} \overset{iid}{\sim} N(0,4) \text{ and } \epsilon \sim N(0,1) \text{ independent of } (X_1, X_2, \ldots, X_{15})$$
$$Y = I_{(-2,+2)}(X_1) \cdot \lfloor X_1 \rfloor + I_{(-\infty,0]}(X_2) + I_{(0,+\infty)}(X_3) + \left|\tfrac{X_6}{4}\right|^3 + \left|\tfrac{X_7}{4}\right|^5 + \tfrac{7}{3}\cos\left(\tfrac{X_{11}}{2}\right) + \epsilon \ .$$

In setting B, the covariate distribution was modified to include clustering. Specifically, we generated $(X_1, X_2, \ldots, X_{15}) \sim MVN_{15}(\mu, \Sigma)$, where the mean vector is

$$\mu = 3 \times (0,0,0,0,0,1,1,1,1,1,0,0,0,0,0) - 2 \times (0,0,0,0,0,0,0,0,0,0,1,1,1,1,1)$$

and the variance-covariance matrix is given by

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} \end{bmatrix},$$

where we have set

$$\Sigma_{11} = \begin{bmatrix} 1 & 0.15 & 0.15 \\ 0.15 & 1 & 0.15 \\ 0.15 & 0.15 & 1 \end{bmatrix}, \quad \Sigma_{22} = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \quad \text{and} \quad \Sigma_{33} = \begin{bmatrix} 1 & 0.85 & 0.85 \\ 0.85 & 1 & 0.85 \\ 0.85 & 0.85 & 1 \end{bmatrix}$$

and each of Σ_{12}, Σ_{13} and Σ_{23} are three-by-three zero matrices. The random error ϵ and the outcome Y are then generated as in setting A. In both settings, we generated 500 random datasets of size $n \in \{100, 300, 500, 1000\}$, and considered the importance of the features included in the sets $\{1,2,3,4,5\}$, $\{6,\ldots,10\}$, and $\{11,\ldots,15\}$ for each sample size. The true value of the variable importance measure corresponding to each of the considered groups in both settings is given in Table 3.1. Results for the analysis of additional groupings for both settings are provided in Appendix A.

For each scenario considered, we estimated the conditional mean functions using gradient

Table 3.1: Approximate values of $\psi_{0,s}$ for each simulation setting and group considered for effect size in the moderate-dimensional simulations in Section 3.3.3.

Group	Setting A	B
$(X_1, X_2, \ldots, X_5)$	0.295	0.281
$(X_6, X_7, \ldots, X_{10})$	0.240	0.314
$(X_{11}, X_{12}, \ldots, X_{15})$	0.242	0.179

boosted trees [Friedman, 2001] fit using the `GradientBoostingRegressor` function in the `sklearn` module in Python. Gradient boosted trees were used due to their generally favorable prediction performance and large degree of flexibility, with full knowledge that they are not guaranteed to satisfy the minimum rate condition outlined in Section 3.2.3. We used five-fold cross-validation to select the optimal number of trees with one node as well as the optimal learning rate for the algorithm. We summarized the results of these simulations in the same manner as in the low-dimensional simulations.

The results for setting A are presented in Figure 3.3. From the top row, we note that as n increases, the scaled empirical bias of the proposed estimator approaches zero whereas that of the naive estimator increases in magnitude across all three groupings s considered. From the bottom row, we observe that the empirical coverage of intervals based on the proposed estimator increases towards the nominal level as n increases, and is uniformly higher than the empirical coverage of the bootstrap intervals based on the naive estimator.

The results for setting B are presented in Figure 3.4. From the top row, we notice some residual bias in the proposed estimator for $s = \{11, \ldots, 15\}$. It is possible that larger samples may be needed to observe more thorough bias reduction – indeed, this group of features is that with the highest within-group correlation. Nevertheless, the scaled empirical bias of the proposed estimator approaches zero as n increases for both $s = \{1, \ldots, 5\}$ and $s = \{6, \ldots, 10\}$. In all cases, the scaled empirical bias of the naive estimator increases in magnitude as n increases. In the bottom row, we again see that intervals based on the

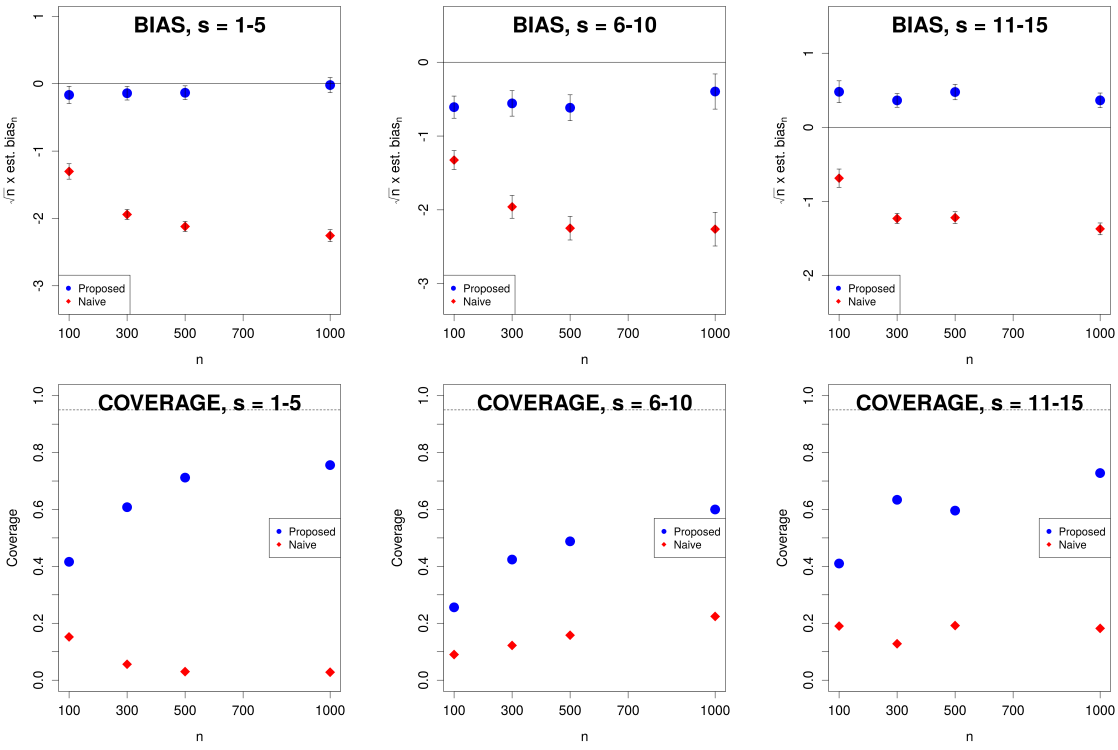

Figure 3.3: Top row: empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting A, using gradient boosted trees to estimate (3.1) and (3.2). Bottom row: empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting B, using gradient boosted trees to estimate (3.1) and (3.2). We consider all s combinations from Table 3.1. Diamonds denote the naive estimator, and circles denote the proposed estimator. Monte Carlo error bars are displayed vertically. This figure appears in color in the electronic version of this chapter.

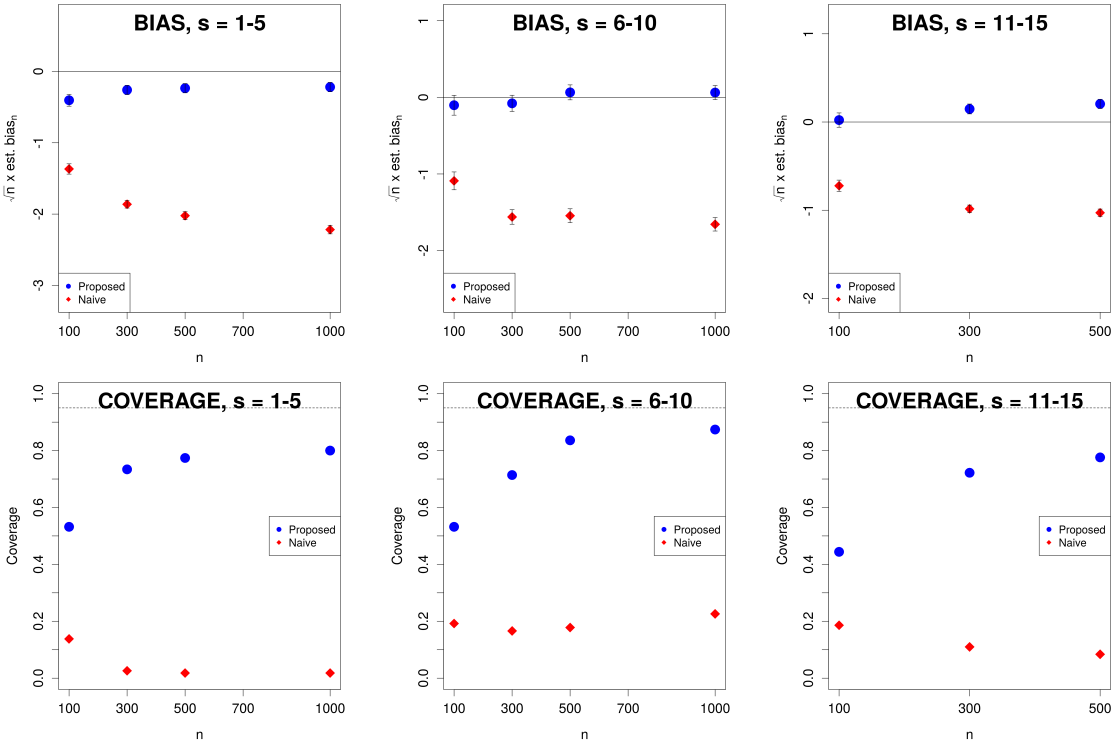

Figure 3.4: Top row: empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting B, using gradient boosted trees to estimate (3.1) and (3.2). Bottom row: empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting B, using gradient boosted trees to estimate (3.1) and (3.2). We consider all s combinations from Table 3.1. Diamonds denote the naive estimator, and circles denote the proposed estimator. Monte Carlo error bars are displayed vertically. This figure appears in color in the electronic version of this chapter.

proposed estimator have uniformly higher coverage than those based on the naive estimator.

The proposed estimator performs substantially better than the naive estimator in these simulations, though higher levels of correlation appear to be associated with relatively poorer point and interval estimator performance. This suggests that it may be wise to consider in practice the importance of entire groups of correlated predictors rather than that of individual features. This is a sensible approach for dealing with correlated features, which necessarily render variable importance assessment challenging. In our simulations the empirical coverage of proposed estimator-based intervals for the importance of a group of highly correlated features ($s = \{11, \ldots, 15\}$, Figure 3.4) approaches the nominal level with increasing sample size, indicating that the proposed approach does yield good results in such cases.

Use of the proposed estimator results in better point and interval estimation performance than the naive estimator in the presence of null features. For example, when evaluating the importance of the group $(X_1, \ldots, X_5)$, the group $(X_8, X_9, X_{10}, X_{12}, \ldots, X_{15})$ has null importance. However, as before, we expect the behavior of point and interval estimators for the variable importance of null features to be not as good. Future work on valid estimation and testing under this null hypothesis is necessary.

3.4 Results from the South African heart disease study data

We consider a subset of the data from the Coronary Risk Factor Study [Rosseauw et al., 1983], a retrospective cross-sectional sample of 462 white males aged 15–64 in a region of the Western Cape, South Africa; these data are publicly available in Hastie et al. [2009]. The primary aim of this study was to establish the prevalence of ischemic heart disease risk factors in this high incidence region. For each participant, the presence or absence of myocardial infarction (MI) at the time of the survey is recorded, yielding 160 cases of MI. In addition, measurements on two sets of features are available: behavioral features, including cumulative tobacco consumption, current alcohol consumption, and type A behavior, a behavioral pattern linked to stress [Friedman and Rosenman, 1971]; and biological features, including systolic blood pressure, low-density lipoprotein (LDL) cholesterol, adiposity (similar to body

mass index), family history of heart disease, obesity, and age.

We considered the importance of each feature separately, as well as that of these two groups of features, when predicting the presence or absence of MI. We estimated the conditional means using the sequential regression estimating procedure outlined in Section 3.2.2 and using the Super Learner [van der Laan et al., 2007] via the `SuperLearner` R package. The Super Learner is a particular implementation of stacking [Wolpert, 1992], and the resulting estimator is guaranteed to have the same risk as the oracle estimator, asymptotically, along with finite-sample guarantees [van der Laan et al., 2007]. Our library of candidate learners consists of boosted trees, generalized additive models, elastic net, and random forests implemented in the R packages `gbm`, `gam`, `glmnet`, and `randomForest` respectively, each with varying tuning parameters. Ten-fold cross-validation was used to determine the optimal convex combination of these learners chosen to minimize the cross-validated mean squared error. This process allowed the Super Learner to determine the optimal tuning parameters for the individual algorithms as part of its optimal combination, and our resulting estimator of the conditional means is the optimal convex combination of the individual algorithms. Finally, we produced confidence intervals based on the proposed estimator alone, since as we have seen earlier, intervals based on the naive estimator are generally invalid.

The results are presented in Figure 3.5. The ordering is slightly different in the two plots; this is not surprising, since the one-step procedure should eliminate excess bias in the naive estimator introduced by estimating the conditional means using flexible learners. We find that biological factors are more important than behavioral factors. The most important individual feature is family history of heart disease; family history has been found to be a risk factor of MI in previous studies. It appears scientifically sensible that both groups of features are more important than any individual feature besides family history.

We compared these results to a logistic regression model fit to these data. Based on the absolute values of z-statistics, logistic regression picks age as most important, followed by family history. This slight difference is captured in our uncertainty estimates (Figure 3.5): there, we see that the point estimates for age and family history are close, and their

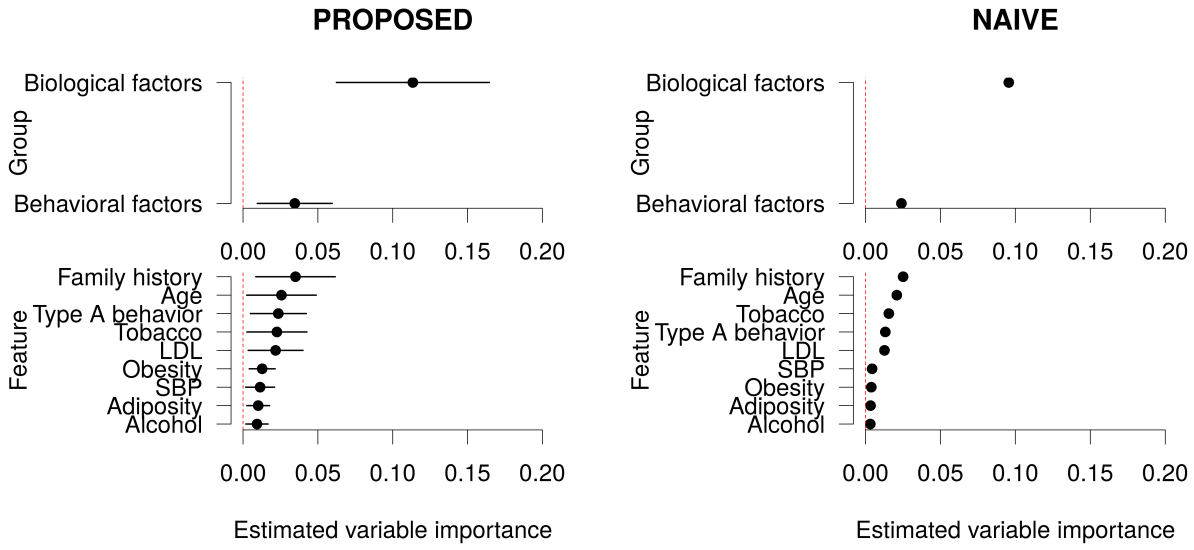

Figure 3.5: Estimates from the South African heart disease study, for the proposed and naive estimators of the standardized variable importance parameter, on left and right respectively. We estimate (3.1) and (3.2) using the Super Learner with the elastic net, generalized additive models, gradient boosted trees, and random forests in its library.

confidence intervals largely overlap. We find the same pattern for LDL cholesterol and tobacco consumption, the third- and fourth-ranked variables by logistic regression. While our results match closely with the simplest approach to analyzing variable importance in these data, our proposed method is not dependent on a single estimation technique, such as logistic regression. The use of more flexible learners to estimate $\psi_{0,s}$, as we have done in this analysis, renders our findings less likely to be driven by potential model misspecification.

3.5 Conclusion

We have obtained novel results for a familiar measure of variable importance, interpreted as the additional proportion of variability in the outcome explained by including a subset of the features in the conditional mean outcome relative to the entire covariate vector. This parameter can be readily seen as a nonparametric extension of the classical R^2 measure, and it provides a description of the true relationship between the outcome and covariates rather than an algorithm-specific measure of association. We have also studied the properties of this parameter and derived its nonparametric efficient influence function. Leveraging tools from semiparametric and nonparametric efficiency theory, we have described the construction of an asymptotically efficient estimator of the true variable importance measure built upon flexible, data-adaptive learners. We found that the corrected estimator $\hat{\psi}_{n,s}$ is equivalent to the plug-in estimator $\hat{\psi}_{n,s,*}$; further study is needed for plug-in estimators of other variable importance functionals. We have studied the properties of this estimator, notably the distributional limit of a suitably normalized version of the estimator, and described the construction of asymptotically valid confidence intervals. In simulations, we have found the proposed estimator to have good practical performance, particularly when comparing to a naive estimator of the proposed variable importance measure. We did find this performance to depend very much on whether or not the true variable importance measure equals zero. When it does, a limiting distribution is not readily available, and significant theoretical innovation then seems to be needed in order to perform valid inference. However, for those features with true importance, the behavior of point and interval estimates is not influ-

enced by the presence of null features. In practice, some judgment is necessary to determine whether there is a sensible cutoff for designating a feature as null, but if it exists, the value of this cutoff would likely be close to zero. While the parameter we have studied has broad interpretability, alternative measures of variable importance may also be useful in certain settings (e.g., difference in the area under the receiver operating characteristic curve in the context of a binary outcome). We will study such measures in future work.

For each candidate set of variables, the estimation procedure we proposed requires estimation of two conditional mean functions. To guarantee the good statistical properties of our estimator, these conditional means must be estimated well. For this reason, and as was illustrated in our work, we recommend the use of model stacking with a wide range of candidate learners, ranging from the very parametric to the fully nonparametric. This flexibility mitigates concerns regarding model misspecification. Recent work suggests that if the Highly Adaptive Lasso estimator is included in the library of learners, then under minimal conditions, the aggregate learner is guaranteed to satisfy the minimum rate of convergence condition [Benkeser and van der Laan, 2016]. Additionally, we suggest the use of sequential regressions to minimize any incompatibility between the two conditional means estimated.

A multiple testing issue arises when inference is desired on many feature subsets; here, a Bonferroni-type approach may be easily implemented. Alternatively, we can use a consistent estimator of the variance-covariance matrix for the importance of all subsets of features under study, obtained using the influence functions exhibited in this paper. This alternative multiple-testing adjustment has improved power over a Bonferroni-type approach. Strategies based on this approach are described, for example, in Dudoit and van der Laan [2007].

Supporting Information

Technical details, additional results from the moderate-dimensional simulations, and an analysis of the Boston housing data are available in Appendix A. Code to reproduce all results is available on the first author's GitHub page. We implement the methods discussed above in the R package `vimp` and the Python package `vimpy`, both freely available on CRAN and

PyPI, respectively.

Acknowledgements

The authors wish to thank Professor Antoine Chambaz for insightful comments that improved this chapter. This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under award numbers F31AI140836, R01AI029168, and U.S. Public Health Service Grant UM1AI068635 [SDMC: HIV Vaccine Trials Network]; and by the NIH under award number DP5OD019820. The opinions expressed in this chapter are those of the authors and do not necessarily represent the official views of the NIAID or the NIH.

Chapter 4

A UNIFIED APPROACH TO MODEL-AGNOSTIC VARIABLE IMPORTANCE

4.1 Introduction

In predictive modeling applications, it is often of interest to assess the relative contribution of subsets of features towards predicting a response, a concept often referred to as variable importance. Based on the scientific question of interest, multiple notions of variable importance may be useful. One such notion describes the population-level relationships between the features and the response; we refer to this as *intrinsic* variable importance. A second notion describes how a given fitted algorithm uses the features to make predictions; we refer to this as *extrinsic* variable importance. Traditionally, intrinsic variable importance has been considered in the context of simple population models (e.g., linear models). For such models, both the prediction algorithm and the associated variable importance measure (VIM) are easy to compute from model fit 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]. 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 instead of simple population models. Extrinsic VIMs for interpreting these complex algorithms have been reviewed by Fisher et al. [2018] and Murdoch et al. [2019], and include, for example, the VIM for random forests [Breiman, 2001a]. However, extrinsic VIMs often do not provide information about the population of interest. While some recently proposed VIMs directly incorporate complex prediction algorithms [van der Laan, 2006, Williamson et al., 2017, Lei et al., 2017], these

VIMs are not as simple to obtain. Thus, to date, intrinsic variable importance has involved a trade-off between good prediction performance and population relevance.

In this chapter, we seek to circumvent the limitations of model-based intrinsic variable importance. Rather than considering variable importance as a function of a given predictive algorithm or simple population model, we provide a unified approach to reformulate 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 of interest, as we illustrate in several examples. In this framework, once a measure of predictiveness has been selected, the task of estimating 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, it is comparable across methods for estimating the oracle prediction functions. 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 Chapter 3, we focused on an application of the proposed framework to infer about a nonparametric R^2-based variable importance, for which we described a nonparametric efficient estimator. We also presented the construction of valid confidence intervals based on flexibly estimated prediction functions, but found formal testing of null importance to be challenging. Here, we extend our results to arbitrary 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 nonparametric 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 chapter is organized as follows. In Section 4.2, we define variable importance as a contrast in population-level oracle predictiveness and provide simple examples. In Section 3.2.2, we construct a nonparametric 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 4.4, we explore an extension to deal with more complex predictiveness measures. In Section 3.3, 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 antibody resistance in Section 4.6, and provide concluding remarks in Section 4.7. All technical details can be found in Appendix B.

4.2 Variable importance

4.2.1 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. Further, 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 ℓ and not in ℓ as w_ℓ and $w_{-\ell}$, 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 $\|\cdot\|_{\mathcal{F}}$, and 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 in s. In all examples we consider, we will take $\mathcal{F}$ to be essentially unrestricted up to regularity conditions. Common choices include the class of all P_0-square-integrable functions from $\mathcal{X}$ to $\mathcal{Y}$ 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 $\mathcal{X}$ to $\mathcal{Y}$ endowed with the supremum norm $f \mapsto \|f\|_{\infty,\mathcal{X}} := \sup_{x \in \mathcal{X}} |f(x)|$.

4.2.2 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 $f \in \mathcal{F}$ when P is the true data-generating distribution, with large values of $V(f, P)$ implying high predictiveness. Examples of predictiveness measures — including R^2, deviance, the area under the ROC curve, and classification accuracy — are discussed in detail in Section 4.2.3. Were the true data-generating mechanism P_0 known, a natural candidate prediction function would be any P_0-population maximizer of predictiveness over the class $\mathcal{F}$:

$$f_0 \in \underset{f \in \mathcal{F}}{\operatorname{argmax}} V(f, P_0) \ . \tag{4.1}$$

This population maximizer can be viewed as the oracle prediction function within $\mathcal{F}$ 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 function class considered, although in contexts we consider this is not the case as long as $\mathcal{F}$ 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 $f \in \mathcal{F}_s$, 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

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0) \ . \tag{4.2}$$

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 $\mathcal{F}$, and if $\mathcal{F}$ is large enough, then this definition is agnostic to this choice.

4.2.3 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 $\mu_0 : x \mapsto E_0(Y \mid X = x)$ under P_0 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 $V(f, P_0) := 1 - E_0 \left\{Y - f(X)\right\}^2 / \sigma_0^2$, where $\sigma_0^2 := E_0 \left\{Y - E_0 \left(Y\right)\right\}^2 = E_0 \left[Y - E_0 \left\{\mu_0(X)\right\}\right]^2$ is 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 $f \mapsto E_0 \left\{Y - f(X)\right\}^2$, 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

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

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. 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 Appendix B, 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 $V(f, P_0) = P_0\{f(X_1) < f(X_2) \mid Y_1 = 0, Y_2 = 1\}$, where (X_1, Y_1) and (X_2, Y_2) represent independent draws from P_0. As shown in Appendix B, 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 but those with index in s is obtained similarly but with μ_0 replaced by $\mu_{0,s} : x \mapsto E_0(Y \mid X_{-s} = x_{-s})$.

4.3 Estimation and inference

4.3.1 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 efficient estimator of $\psi_{0,s}$ using independent observations $Z_1, \ldots, Z_n$ from P_0. Definition (4.2) suggests considering the plug-in estimator

$$\psi_{n,s} := V(f_n, P_n) - V(f_{n,s}, P_n) \ , \tag{4.3}$$

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

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

where μ_n and $\mu_{n,s}$ are estimates of the conditional means μ_0 and $\mu_{0,s}$, respectively. We provide the explicit form of our proposed estimator for all examples in Appendix B.

The simplicity of the plug-in construction makes it particularly appealing. However, the

literature on semiparametric inference and targeted learning suggests that such an estimator may fail to be consistent at rate $n^{-1/2}$ — and thus is inefficient — if f_0 and $f_{0,s}$ are flexibly estimated. Generally, this 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]. Nevertheless, in Chapter 3, 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. This combination of simplicity and optimality makes these plug-in estimators attractive.

4.3.2 Large-sample properties

We now study conditions under which $\psi_{n,s}$ is a nonparametric efficient estimator of the VIM value $\psi_{0,s}$ and 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

$$v_n - v_0 \;=\; \{V(f_0, P_n) - V(f_0, P_0)\} + \{V(f_n, P_0) - V(f_0, P_0)\} + r_n \;, \tag{4.4}$$

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 (4.4) 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 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 $\mathcal{F}$, we may

reasonably expect that

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

for any smooth path $\{f_{0,\epsilon} : -\infty < \epsilon < +\infty\} \subset \mathcal{F}$ 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 $\mathcal{R} := \{c(P_1 - P_2) : c \in \mathbb{R}, P_1, P_2 \in \mathcal{M}\}$ of finite signed measures generated by $\mathcal{M}$. For any $R \in \mathcal{R}$, say $R = c(P_1 - P_2)$, we refer to the supremum norm $\|R\|_\infty := |c| \sup_z |F_1(z) - F_2(z)|$, 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 $h \in \mathcal{R}$, 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 $\{z\}$. Finally, we define the following set of conditions:

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

(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

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

(A3) (*minimum rate of convergence*) $\|f_n - f_0\|_{\mathcal{F}} = o_P(n^{-1/4})$;

(A4) (*weak consistency*) $\int \{g_n(z)\}^2 dP_0(z) = o_P(1)$;

(A5) (*limited complexity*) there is a P_0-Donsker class $\mathcal{G}_0$ such that $P_0(g_n \in \mathcal{G}_0) \to 1$.

Theorem 2. *If conditions (A1)–(A5) hold, then v_n is an asymptotically linear estimator of v_0 with influence function equal to the nonparametric efficient influence function (EIF) $\phi_0 : z \mapsto \dot{V}(f_0, P_0; \delta_z - P_0)$.*

This result implies, in particular, that the plug-in estimator v_n is a consistent, asymptotically normal, and nonparametric efficient estimator of v_0. 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 $\mathcal{F}$ by $f_{n,s}$, $f_{0,s}$ and $\mathcal{F}_s$ in the conditions above, and denoting the resulting EIF 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 $\varphi_{0,s} : z \mapsto \dot{V}(f_0, P_0; \delta_z - P_0) - \dot{V}(f_{0,s}, P_0; \delta_z - P_0)$. If $0 < \tau_{0,s}^2 := E_0\{\varphi_{0,s}^2(Z)\} < \infty$ and $\psi_{0,s} > 0$, 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 $z_{1-\alpha}$ denotes the $(1-\alpha)^{th}$ quantile of the standard normal distribution. We discuss the setting where $\psi_{0,s} = 0$ in Section 4.3.5.

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 $\mathcal{F}$ must be determined for this condition to hold. We have provided details for Examples 1–4 in Appendix B, 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 $\mu_0 \in \mathcal{F}$, and so, condition (A1) holds with $C = 1/\sigma_0^2$ and $\|\cdot\|_{\mathcal{F}}$ taken to be either the $L_2(P_0)$ or supremum norm. In Example 2, provided that all elements of $\mathcal{F}$ are bounded between γ and $1-\gamma$ for some $\gamma \in (0, 1)$, and that $\mu_0 \in \mathcal{F}$, then condition (A1) holds with $C = \{\gamma \log(1-\gamma)\}^{-1}$ and $\|\cdot\|_{\mathcal{F}}$ taken to be either the $L_2(P_0)$ or supremum norm. In Example 3, condition (A1) holds for $C = \kappa$ and $\|\cdot\|_{\mathcal{F}}$ 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 $C = 2\kappa$ and $\|\cdot\|_{\mathcal{F}}$ 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 $\delta > 0$. Details are provided in Appendix B. Condition (A3) requires that f_0 be estimated at a sufficiently fast rate in order for second-order terms to be asymptotically negligible, while condition (A4) 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 (A5) restricts the complexity of the algorithm used to generate f_n. Conditions (A3)–(A5) depend not only on the predictiveness measure chosen and on the true data-generating mechanism but also on properties of the prediction function estimator.

4.3.3 Special case: standardized V-measures

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

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

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 seen so far is a standardized V-measure, namely defined by:

(ex. 1) $a = 1$, $G((u, v)) = -(u - v)^2$, $V_2(P) = var_P(Y)$, $m = 1$;

(ex. 2) $a = 1$, $G((u, v)) = -\{u \log v + (1 - u) \log(1 - v)\}$, $V_2(P) = P(Y = 1) \times \log P(Y = 1) + P(Y = 0) \log P(Y = 0)$, $m = 1$;

(ex. 3) $a = 0$, $G((u, v)) = I(u = v)$, $V_2(P) = 1$, $m = 1$;

(ex. 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 as:

(ex. 1) $\phi_0(z) = -\{y - \mu_0(x)\}^2 / \sigma_0^2 + v_0 \{2 - (y - \mu_0)^2 / \sigma_0^2\}$;

(ex. 2) $\phi_0(z) = -2[y \log \mu_0(x) + (1 - y) \log \{1 - \mu_0(x)\}] / \overline{\pi}_0 + v_0 [2 \log \{\pi_0/(1 - \pi_0)\} (y - \pi_0)/\overline{\pi}_0 - 1]$;

(ex. 3) $\phi_0(z) = yI \{\mu_0(x) > 0.5\} + (1 - y) I \{\mu_0(x) \leq 0.5\} - v_0$;

(ex. 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 we have used the shorthand notation $\mu_0(x) := E_0(Y \mid X = x)$, $\mu_0 := E_0(Y)$, $\sigma_0^2 := var_0(Y)$, $\pi_0 := P_0(Y = 1)$, and $\overline{\pi}_0 := \pi_0 \log \pi_0 + (1 - \pi_0) \log(1 - \pi_0)$. Furthermore, for standardized V-measures, condition (A2) is often easier to verify. 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. If instead $m \geq 2$, condition (A2) is satisfied, for example, if $|G((Y_1, f(X_1)), \ldots, (Y_m, f(X_m)))|$ is bounded almost surely under P_0.

4.3.4 Implementation based on cross-fitting

Condition (A5) 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 the use of sample-splitting, whereby f_0 is estimated using the training data, and the predictiveness measure is then evaluated on the test data. This readily extends to K-fold cross-fitting. The resulting cross-fit estimator v_n^* is nonparametric efficient under weaker conditions than those imposed on v_n, as the theorem below states.

Theorem 3. *If conditions (A1)–(A4) hold, then v_n^* is an asymptotically linear estimator of v_0 with influence function corresponding to the nonparametric EIF ϕ_0.*

As before, this theorem also readily provides conditions under which $v_{n,s}^*$ is an asymptotically linear estimator of $v_{0,s}$ with influence function equal to the nonparametric EIF $\phi_{0,s}$, and so, under which $\psi_{n,s}^* := v_n^* - v_{n,s}^*$ is a nonparametric efficient estimator of the VIM value $\psi_{0,s}$. We describe the construction of the cross-fitted estimator $\psi_{n,s}^*$ of $\psi_{0,s}$ in Algorithm 3, and provide the explicit form of $\psi_{n,s}^*$ for Examples 1–4 in Appendix B. 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 3 Estimation of VIM value $\psi_{0,s}$ using K-fold cross-fitting

1: generate a random vector $B_n \in \{1, \ldots, K\}^n$ by sampling uniformly from $\{1, \ldots, K\}$ with replacement, and denote by D_j the subset of observations with index in $\{i : B_{n,i} = j\}$ for $j = 1, \ldots, K$.
2: **for** $k = 1, \ldots, K$ **do**
3: using the data in $\cup_{j \neq k} D_j$, construct estimates f_n^k of f_0 and $f_{n,s}^k$ of $f_{0,s}$;
4: using the data in D_k, construct empirical distribution estimate P_n^k of P_0;
5: compute $\psi_{n,s}^k := V(f_n^k, P_n^k) - V(f_{n,s}^k, P_n^k)$;
6: **end for**
7: set $\psi_{n,s}^* := \frac{1}{K} \sum_{k=1}^K \psi_{n,s}^k$.

4.3.5 Testing the null VIM 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 a valid test of the null hypothesis $\psi_{0,s} = 0$ is difficult. In such cases, standard Wald-type testing based on $\tau_{n,s}^2$ will fail to be calibrated, as illustrated in numerical simulations reported in Chapter 3. 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}.

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 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 is true for the corresponding cross-fitted procedures — this is the approach we consider below.

In practice, a group of variables may be considered unimportant even when $\psi_{0,s}$ is nonzero but small, yet such grouping would be deemed important in large enough samples. For this reason, given a threshold $\delta > 0$, it may be more scientifically appropriate to consider testing the δ-null $H_0 : \psi_{0,s} \in [0, \delta]$ versus its complement alternative $H_1 : \psi_{0,s} > \delta$. The δ-null approaches the strict null as $\delta \to 0$. The idea of sample-splitting also allows us to tackle δ-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 $\eta_0 := E_0\{\phi_0(Z)\}^2$ and $\eta_{0,s} := E_0\{\phi_{0,s}(Z)\}^2$, respectively. Then, we may consider rejecting the δ-null hypothesis H_0 in favor of its complement H_1 if and only if

$$t_n := \frac{v_n^* - v_{n,s}^* - \delta}{\sqrt{\frac{1}{n-n_s}\eta_n^2 + \frac{1}{n_s}\eta_{n,s}^2}} > z_{1-\alpha} \ , \tag{4.5}$$

where $z_{1-\alpha}$ 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 4.

Algorithm 4 Cross-fitted test of $H_0 : \psi_{0,s} \in [0, \delta]$ vs $H_1 : \psi_{0,s} > \delta$ at level $1 - \alpha$ for $\delta \geq 0$

1: Generate a random vector $B_n \in \{1, 2\}^n$ by sampling uniformly from $\{1, 2\}$ with replacement, and build datasets D_1 and D_2 with observations indexed in $\{i = 1, 2, \ldots, n : B_{n,i} = 1\}$ and $\{i = 1, 2, \ldots, n : B_{n,i} = 2\}$, respectively.
2: Compute v_n^* and η_n using observations in D_1 alone and K-fold cross-fitting.
3: Compute $v_{n,s}^*$ and $\eta_{n,s}$ using observations in D_2 alone and K-fold cross-fitting.
4: Compute t_n according to the definition provided in (4.5).
5: Set $p_n := 1 - \Phi(t_n)$ with Φ the standard normal distribution function.
6: Reject H_0 in favor of H_1 if and only if $p_n < \alpha$.

In the above procedure, the use of distinct subsets of the data is critical for constructing v_n^* and $v_{n,s}^*$, but not for η_n and $\eta_{n,s}$, which could be estimated using the entire dataset instead.

4.4 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 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 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 the larger the values of Y, the better the clinical outcome, 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 be observed 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 $\mathbb{V}(f, \mathbb{P}_0) := E_{\mathbb{P}_0}[\{1 - f(X)\}Y(0) + f(X)Y(1)]$. In contrast, the observed data structure is $Z := (X, A, Y) \sim P_0$, and we must identify some observed-data predictiveness measure V such that $V(f, P_0) = \mathbb{V}(f, \mathbb{P}_0)$. Defining the outcome regression $Q_P(a, x) := E_P(Y \mid A = a, X = x)$, it is not difficult to verify that

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

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 a key requirement from Section 4.3.

The simple plug-in approach described in Section 4.3 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 $f \mapsto V(f, P)$, and define the parameter mapping $V^* : P \mapsto V(f_P, P)$ so that v_0 can be expressed as $V^*(P_0)$. If $\widehat{P} \in \mathcal{M}$ is an estimator of P_0, the plug-in estimator $V^*(\widehat{P}_n)$ typically fails to be asymptotically linear: it suffers from excessive bias whenever flexible learning techniques have been used 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], for example. This fact renders the use of debiasing approaches necessary. In contrast, the one-step estimator

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

where ϕ_n is the nonparametric EIF of V^* at $\widehat{P}_n$, is nonparametric efficient [Pfanzagl, 1982] under regularity conditions. Alternatively, the framework of targeted minimum loss-based estimation describes how to convert $\widehat{P}_n$ into a revised estimator $\widehat{P}_n^*$ such that $V^*(\widehat{P}_n^*)$ is itself nonparametric efficient without further debiasing [van der Laan and Rose, 2011]. Similarly as in Section 4.3, 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 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, the optimality of f_P implies that f_P does not contribute to the nonparametric EIF of $P \mapsto V(f_P, P)$, as the theorem below states.

Theorem 4. *If both $P \mapsto V(f_P, P)$ and $P \mapsto V(f_0, P)$ are pathwise differentiable at P_0 relative to the nonparametric model $\mathcal{M}$, they must have the same nonparametric EIF at P_0.*

This result is simple yet powerful. It indicates that provided regularity conditions for pathwise differentiability hold, 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 novel derivations. Armed with this observation, we revisit the motivating example we presented in this section, and also consider an additional example involving data missingness.

Example 5: mean outcome under a binary intervention rule
As described above, in this example, the full data parameter $\mathbb{V}(f, \mathbb{P}) := E_{\mathbb{P}}[Y(f(X))]$ 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 $\mathcal{F}$ 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 $Q_{0,s}(a, x) := E_0\{Q_0(a, X) \mid X_{-s} = x_{-s}\}$. 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 $P \mapsto V(f_0, P)$ — and thus also of $P \mapsto V(f_P, P)$ — 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 $g_0(a, x) := P_0(A = a \mid X = x)$ for each $a \in \{0, 1\}$.

Thus, under regularity conditions (provided in Appendix B), 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 we have defined f_n 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 an appropriate estimator of $f_{0,s}$, say $f_{n,s}(x) := I\{Q_{n,s}(1,x) > Q_{n,s}(0,x)\}$ 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 for 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 we have defined $U := \Delta Y$. The observed-data predictiveness measure

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

equals the ideal-data accuracy measure provided that (a) Δ and Y are independent given X, and (b) $P(\Delta = 1 \mid X = x) > 0$ 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 (existing) subpopulation of patients (as defined by the value of X) for which the outcome can never be observed. Defining $\pi_0(x) := P_0(U = 1 \mid \Delta = 1, X = x)$, the unrestricted optimizers f_0 and $f_{0,s}$ are given pointwise by $f_0(x) = I\{\pi_0(x) > 0.5\}$ and $f_{0,s}(x) = I\{E_0\{\pi_0(X) \mid X_{-s} = x_{-s}\} > 0.5\}$. Finally, the nonparametric EIF of $P \mapsto V(f_0,P)$ at P_0

is given by

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

where we now have defined the nuisance parameters $g_0(x) := P_0 \left(\Delta = 1 \mid X = x \right)$ and $Q_0(x) := P_0 \left\{ Y = f_0(x) \mid \Delta = 1, X = x \right\} = f_0(x)\pi_0(x) + \{1 - f_0(x)\}\{1 - \pi_0(x)\}$, so that $Q_0(x)$ is no more than a simple transformation of $\pi_0(x)$.

4.5 Numerical experiments

4.5.1 Simulation setup

We now present empirical results describing the performance of our proposed plug-in VIM estimator. We consider in all cases two covariates and a binary outcome generated as

$$Y \sim Bernoulli(0.6) \ \text{ and } \ X \mid Y = y \sim N_2(\mu_y, \Sigma) \ ,$$

where Σ is the 2×2 identity matrix and $\mu_0 = (0, 0)^\top$. We consider $\mu_1 = (1.5, 2)^\top$ (Scenario 1) and $\mu_1 = (1.5, 0)^\top$ (Scenario 2), which reflect both features versus one feature (X_1) having nonzero importance, respectively. In each experiment, we generated 1,000 random datasets of size $n \in \{100, 500, 1000, \ldots, 4000\}$, and considered in each case the importance of both X_1 and X_2. All analyses were performed using our R package `vimp` and may be reproduced using code available online (see details in Appendix B).

To obtain f_n and each $f_{n,j}$, we used the Super Learner [van der Laan et al., 2007], a particular implementation of stacking [Wolpert, 1992]; the resulting estimator is guaranteed to have the same risk as the oracle estimator, asymptotically, and also enjoys finite-sample performance guarantees [van der Laan et al., 2007]. We used as candidate estimators gradient boosted trees [Friedman, 2001] and random forests [Breiman, 2001a], each with the default tuning parameters, in addition to logistic regression and the sample mean. We used five-fold cross-validation to determine the optimal convex combination of these learners cho-

Table 4.1: Approximate values of $\psi_{0,s}$ for the numerical experiments in Section 4.5. The values are different for the different VIMs.

Scenario 1		Scenario 2		
X_1	X_2	X_1	X_2	Importance measure
0.051	0.116	0.181	0	Accuracy
0.040	0.106	0.356	0	AUC

sen to minimize the cross-validated negative log-likelihood. The resulting optimal convex combination of the individual algorithms is our estimator of the conditional means involved in these simulations.

We considered here 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 are provided in Table 4.1. We computed the relevant plug-in estimator using Algorithm 3 with five-fold cross-fitting. We additionally computed nominal 95% Wald-type confidence intervals and obtained p-values for the null hypothesis for each replication using the sample-splitting procedure of Algorithm 4. We then computed the empirical mean squared error (MSE) scaled by n, the empirical coverage of nominal 95% confidence intervals, and the empirical power of our hypothesis test for the δ-null with $\delta = 0$ (strict null) and $\delta = 0.05$. The latter value indicates a scenario in which only an absolute decrease in classification accuracy or AUC of 5%, i.e., $\psi_{0,s} > 0.05$, is deemed to reflect scientifically meaningful importance.

4.5.2 Primary empirical results

Figure 4.1 displays the results of the experiment of Scenario 1, where both features have nonzero importance. In the top-left panel, we observe that the MSE of the proposed estimators is approximately proportional to n^{-1}, as expected, regardless of sample size or which feature we remove. The top-right panel shows that the coverage of nominal 95% confidence intervals increases to the nominal level with increasing sample size. In the bottom row, we

display the proportion of tests rejected for the strict null ($\delta = 0$, bottom-left panel) or the δ-null with $\delta = 0.05$ (bottom-right panel). Regardless of the value of δ, for $j = 2$, the power of our proposed cross-validated hypothesis testing procedure increases with increasing sample size, reflecting the fact that the true VIM value for this feature is larger than δ in both cases. However, a large sample size is necessary for high power. We expect that with a large enough sample size, even for features close to the null, there will be power to detect an effect. For $j = 1$, the proportion of tests rejected approaches one when $\delta = 0$, but is approximately controlled at 0.05 when $\delta = 0.05$. This is expected since in the latter case the true VIM value is within the null region for AUC and only very slightly outside the null region for accuracy.

In Figure 4.2, we display the results of the experiment in Scenario 2. We see the same results as in Figure 4.1 for the top row. Encouragingly, in the bottom row, we see that for $j = 1$ power increases fairly rapidly, reflecting the large effect size for this variable. For $j = 2$, the type I error is controlled slightly below the 0.05 level. This holds true regardless of the value of δ.

This simulation study suggests that the estimation and inferential procedures proposed, including our sample-split hypothesis test, have good performance, as suggested by theory.

4.5.3 Additional empirical results

We ran several additional simulation studies. We investigated the behavior of our procedures for the deviance-based VIM (*Example 2*), and found very similar results as reported here for accuracy and AUC. Details are provided in Appendix B. We also investigated the importance of cross-fitting when constructing VIM estimators. Our findings suggest that cross-fitting is critical when flexible algorithms are used, in which case the simple estimator performs poorly and its cross-fitted counterpart instead shows good performance. When correctly-specified parametric model-based estimators are used, both estimators perform similarly well, reflecting that fact that for parametric estimators the Donsker class condition (A5) is automatically satisfied and cross-fitting is not needed.

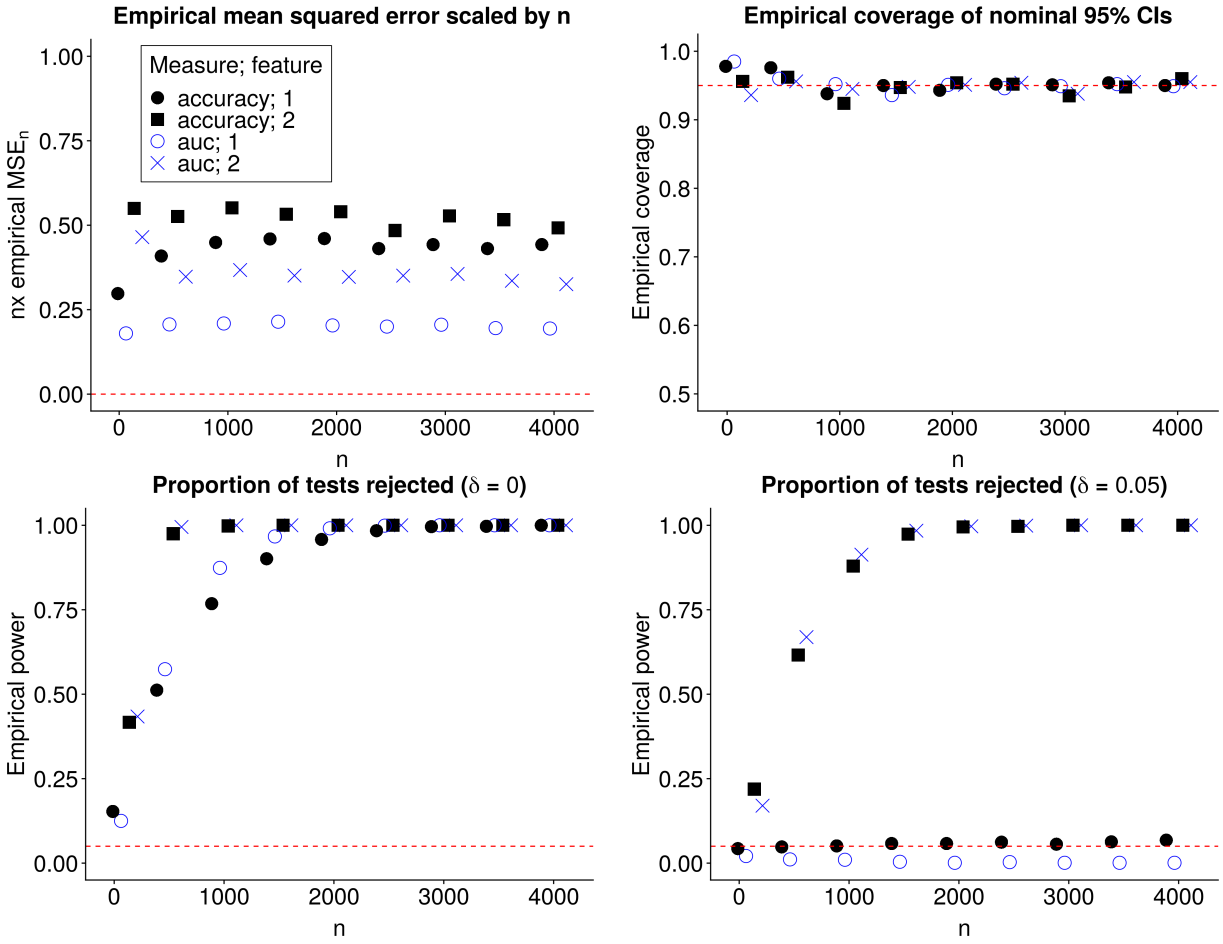

Figure 4.1: Performance of plug-in estimators for estimating importance via the accuracy and AUC. Clockwise from top left: empirical mean squared error for the proposed plug-in estimator scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; power of the split-sample hypothesis testing procedure vs n when $\delta = 0.05$; and power of the split-sample hypothesis testing procedure vs n when $\delta = 0$. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively. This figure appears in color in the electronic version of this chapter.

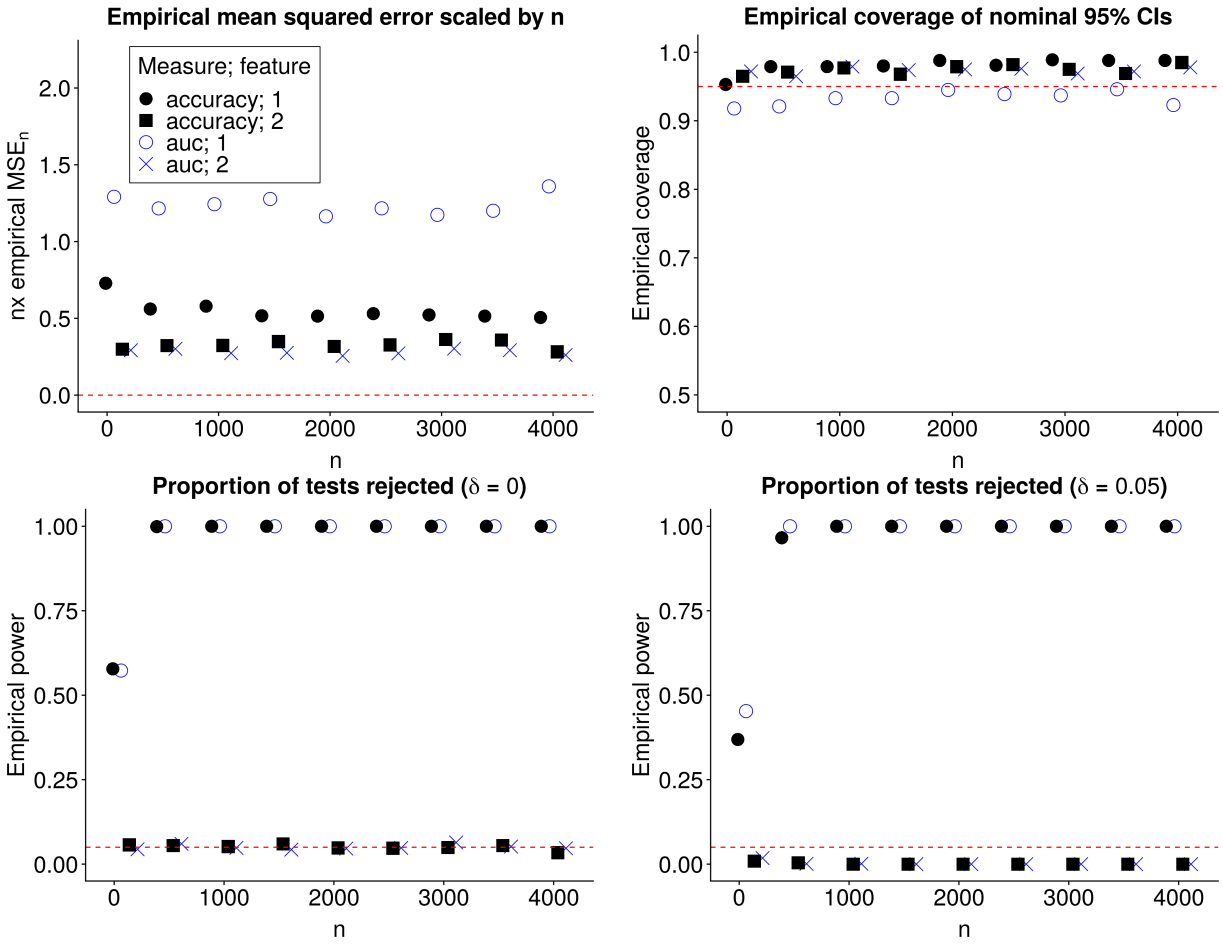

Figure 4.2: Performance of plug-in estimators for estimating importance via the accuracy and AUC. Clockwise from top left: empirical mean squared error for the proposed plug-in estimator scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; power of the split-sample hypothesis testing procedure vs n when $\delta = 0.05$; and power of the split-sample hypothesis testing procedure vs n when $\delta = 0$. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively. We operate here under the null hypothesis for X_2: $\psi_{0,s} = 0$. This figure appears in color in the electronic version of this chapter.

4.6 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 are evaluating one promising bnAb, VRC01, for its ability to prevent HIV-1 infection [Gilbert et al., 2017]. A secondary objective will 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 is being designed to restrict to a subset of AA features that putatively affect prevention efficacy. Given the hypothesis undergirding the trials that VRC01 will prevent infection via *in vivo* neutralization, a useful approach ranks AA features by their estimated VIM for predicting *in vitro* neutralization — e.g., whether or not an HIV-1 virus is sensitive to neutralization by VRC01, denoted by "IC_{50} Censored" below — and selects the top-ranked features.

In an effort to determine these important AA features, Magaret et al. [2019] analyzed HIV-1 envelope (Env) AA sequence features of 611 publicly-available HIV-1 Env pseudoviruses made from blood samples of HIV-1 infected individuals and 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]. The authors estimated variable importance using the difference in nonparametric R^2. Here, we replicate the results of Magaret et al. [2019] and compare results based on R^2 with a variable importance analysis based on classification accuracy and AUC.

We used the Super Learner with a large library of candidate learners to estimate the involved conditional mean functions. These learners included the lasso, random forests, and boosted decision trees, each with varying tuning parameters. Our library of learners is described more fully in Appendix B. Our resulting estimator is the convex combination of the candidate estimators, where we used ten-fold cross-validation to determine the convex

combination that minimized the negative log-likelihood loss. We additionally describe the performance of our estimator in Appendix B.

In Figure 4.3, 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 in an estimate that the CD4 binding sites and the VRC01 binding footprint are the two most important groups. This matches our expectations from basic science experiments that 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 (i.e., $\delta = 0$) for each group. We found that AA features in the CD4 binding sites (group 2) and the VRC01 binding footprint (group 1) had p-values of 7×10^{-3} and 0.015, respectively, based on AUC (denoted by stars in Figure 4.3). Based on these analyses, AA features in the CD4 binding sites and VRC01 binding footprint 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 AMP analysis.

4.7 Discussion

We have proposed a unified 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

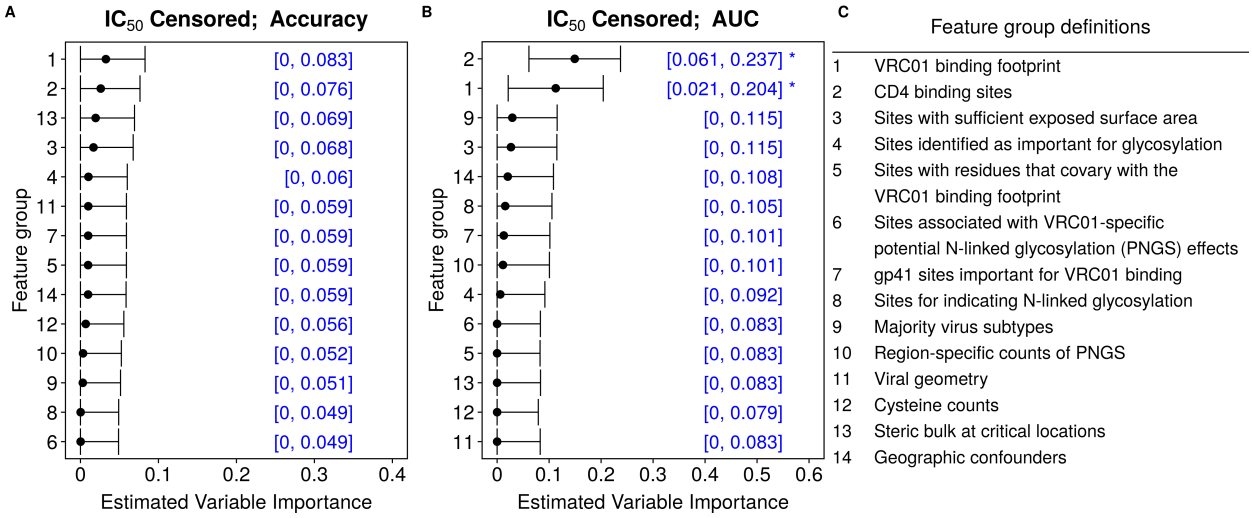

Figure 4.3: Variable importance measured by accuracy (panel A) and AUC (panel B) for the groups defined in panel C. Stars denote importance deemed statistically significantly different from zero at the 0.05 level. This figure appears in color in the electronic version of this chapter.

cases typically require more effort, including the computation of an efficient influence function. Interpretation of the estimated VIMs and p-values depends on the application, and may include considering the ranked VIM values, considering features with VIM values above some scientifically meaningful threshold, considering features based on a p-value threshold, or some combination of these.

We also proposed a simple strategy for performing tests of the null 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. Developing a more powerful test of the null importance hypothesis is therefore of great interest. 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. Additionally, because the proposed confidence intervals are based on the

entire dataset and the proposed hypothesis test is based on sample splitting, there is not a one-to-one correspondence between the two procedures in terms of the null hypothesis being rejected with p-value $< \alpha$ if and only if the confidence interval excludes δ. Harmonizing these inferential procedures is thus of interest. These ideas are being pursued in ongoing research.

Software

We implement the methods discussed above in the R package `vimp` and the Python package `vimpy`, both freely available on CRAN and PyPI, respectively.

Supporting Information

Technical details are available in Appendix B. All results may be reproduced using code hosted on the first author's GitHub page. The data from Section 4.6 are available at https://github.com/benkeser/vrc01/tree/1.0.

Acknowledgements

This work was supported by NIH awards F31 AI140836, R01 AI029168, UM1 AI068635 and DP5 OD019820. The opinions expressed in this chapter are those of the authors and do not necessarily represent the official views of the NIAID or the NIH.

Chapter 5

ASSESSING POPULATION FEATURE IMPORTANCE USING SHAPLEY VALUES

5.1 Introduction

Variable importance describes the contribution of subsets of features towards predicting the response. Understanding this importance may make machine learning-based algorithms more interpretable [see, e.g., Guidotti et al., 2018, Murdoch et al., 2019]. Estimating and obtaining valid statistical inference on the true population variable importance measure (VIM) when flexible estimation techniques are used is increasingly of interest. In particular, valid statistical inference on the true VIM is crucial when high-impact decisions will be made based on the results of machine learning algorithms.

Recently proposed VIMs may be broadly categorized as being *extrinsic* (describing the fitted algorithm) or *intrinsic* (describing the population of interest). Extrinsic VIMs include both individual prediction-level and dataset-level measures. Many extrinsic VIMs are defined with respect to a particular algorithm; examples of these VIMs include the VIMs for trees [Breiman, 2001a] and neural networks [see, e.g., Garson, 1991]. This may lead to difficulty in interpreting the output of the variable importance procedure; thus, recent work has focused on defining extrinsic VIMs that may be used with many algorithms. One such group of VIMs extends the Shapley values [Shapley, 1953], a useful concept from game theory [see, e.g., Lundberg and Lee, 2017, Lundberg et al., 2018, Aas et al., 2019, Frye et al., 2019]. While these measures are useful for describing how a fitted algorithm used the features to make an individual prediction, statistical inference on these measures appears challenging. Additionally, Lundberg and Lee [2017], Aas et al. [2019], and Frye et al. [2019] use a particular predictiveness measure, namely the conditional expectation, that may not be useful in all

contexts. Examples of intrinsic VIMs include the difference in nonparametric R^2 [Doksum and Samarov, 1995, Feng et al., 2018] and the mean absolute difference [Lei et al., 2017]. The VIMs proposed and validated in Chapters 3 and 4 are also intrinsic, population importance measures. However, these measures are all interpreted as the importance of a subset of features relative to the remaining features.

In this chapter, we propose a unified approach for assessing population feature importance using Shapley values based on a chosen measure of predictiveness, e.g., R^2 or area under the receiver operating characteristic curve (AUC). The population Shapley value may be viewed as an intermediate measure between the *marginal* importance of a subgroup of features relative to the null model that does not use covariate information and the *conditional* importance of a subgroup of features explored in Chapters 3 and 4. Our contributions in this chapter are: (i) the population Shapley value describing the importance of a subset of features relative to a given measure of predictiveness, (ii) a simple procedure for estimating the population Shapley value and obtaining a confidence interval, and (iii) a procedure for obtaining a p-value for testing the zero-importance null hypothesis. We demonstrate empirically that our method estimates variable importance accurately, yields confidence intervals with asymptotically correct coverage, and provides type I error control. We also analyze data from a study of mortality prediction in intensive care unit (ICU) patients.

Since it is computationally intractable to sum over all possible subsets, there are generally two approaches to estimating the Shapley value: sampling-based [see, e.g., Castro et al., 2009, Štrumbelj and Kononenko, 2014] and algorithm-specific approaches. Algorithm-specific approaches to estimating the Shapley value have been pioneered by Lundberg and Lee [2017]. Obtaining prediction functions is simplified for some algorithms through using the structure of the algorithm. However, these algorithm-specific methods currently rely on a particular choice of measure of predictiveness, namely the conditional expectation of the outcome given the features.

5.2 Variable importance

5.2.1 Data structure and notation

Consider the random vector $O = (X, Y)$ drawn according to a joint probability distribution P_0. We assume that P_0 belongs to a rich (nonparametric) class of distributions $\mathcal{M}$. Suppose that observations $O_1, \ldots, O_n$ are drawn independently from P_0. Further, suppose that $X_i = (X_{i1}, \ldots, X_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$ is a feature vector and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is the outcome of interest. Here, $\mathcal{X}$ and $\mathcal{Y}$ denote the sample spaces of X and Y, respectively. For a set of indices ℓ and a vector a, we refer to the elements of a with index in ℓ and not in ℓ as a_ℓ and $a_{-\ell}$, respectively. We will use the shorthand notation E_0 below to refer to expectation under P_0.

We denote the power set of $N := \{1, \ldots, p\}$ by $\mathcal{S}$. For any $s \in \mathcal{S}$, we denote by $\mathcal{X}_s$ and $\mathcal{X}_{-s}$ the sample spaces of X_s and X_{-s}, respectively. We also define 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 $\|\cdot\|_{\mathcal{F}}$, and for any $s \in \mathcal{S}$ 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 will take $\mathcal{F}$ to be essentially unrestricted up to regularity conditions.

5.2.2 Oracle predictiveness and the population Shapley value

We now detail how we define variable importance as a population parameter. As in Chapter 4, suppose that $V(f, P)$ is a measure of the predictiveness of a given candidate prediction function $f \in \mathcal{F}$ when P is the true data-generating distribution, and that large values of $V(f, P)$ imply high predictiveness. Examples of predictiveness measures — including R^2, deviance, area under the ROC curve, and classification accuracy — are provided in Section 5.2.4.

We consider any maximizer of predictiveness over the class $\mathcal{F}$ under P_0,

$$f_0 \in \underset{f \in \mathcal{F}}{\operatorname{argmax}} V(f, P_0), \tag{5.1}$$

as a natural candidate prediction function if P_0 were known. Thus, f_0 can be viewed as the oracle prediction function within $\mathcal{F}$ under P_0. So long as $\mathcal{F}$ is sufficiently rich, then the definition of f_0 depends on the chosen predictiveness measure and on the data-generating mechanism. Often, f_0 (or some transformation thereof) is the underlying target of estimation for machine learning-based prediction algorithms. The *total 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 $f \in \mathcal{F}_{-s}$, the *residual oracle predictiveness* $V(f_{0,-s}, P_0)$ quantifies the remaining prediction potential after exclusion of features with index in s. In Chapter 4, we defined the importance of the variable (or subgroup of variables) X_s relative to the full covariate vector X as $V(f_0, P_0) - V(f_{0,-s}, P_0)$: the amount of oracle predictiveness lost by excluding X_s from X. However, this definition of variable importance will assign null importance to collinear features. To address this issue, it is useful to consider instead the oracle prediction function $f_{0,s}$ that maximizes $V(f, P_0)$ over all $f \in \mathcal{F}_s$. Then the *marginal oracle predictiveness* $V(f_{0,s}, P_0)$ quantifies the prediction potential of features with index in s compared to the null model $f_{0,\emptyset}$ that does not access any feature information. In some applications, there is a set of covariates $W := X_c$ that are only included as potential confounders, where c denotes the indices of these confounding variables. Here, $f_{0,c}$ may be used in place of $f_{0,\emptyset}$ to define the marginal oracle predictiveness.

We define the *population-level Shapley value importance* of the variable X_j as the average gain in marginal oracle predictiveness from including feature X_j. In other words, we consider the VIM value defined as

$$\psi_{0,j} := \frac{1}{p} \sum_{s \in \mathcal{S}} \binom{p-1}{|s|}^{-1} \{V(f_{0,s \cup j}, P_0) - V(f_{0,s}, P_0)\}. \tag{5.2}$$

Here, we use the classical form of the Shapley value [see, e.g., Shapley, 1953, Charnes et al., 1988] with an arbitrary measure of predictiveness. By construction, we note that $\psi_{0,j} \geq 0$. Since the definition of $\psi_{0,j}$ involves the oracle prediction function within $\mathcal{F}$, is is important to choose $\mathcal{F}$ to be large enough so that its choice does not restrict f_0. As above, if there is an index set c denoting potential confounders that must always be adjusted for, then we can define $\mathcal{S}$ accordingly so that the minimum subset is c rather than the empty set.

The Shapley values have several desirable properties in the context of feature importance. First, $\sum_{j=1}^{p} \psi_{0,j} = V(f_0, P_0) - V(f_{0,\emptyset}, P_0)$. In game-theoretic applications, $v_{0,\emptyset} := V(f_{0,\emptyset}, P_0)$ is often assumed to be zero; however, in the context of predictiveness, this is not always the case. Thus, each individual Shapley value can be viewed as the contribution of the corresponding feature to the total oracle predictiveness $v_0 := V(f_0, P_0)$. Second, the Shapley values are symmetric: for marginal oracle predictiveness $v_{0,s} := V(f_{0,s}, P_0)$, if $v_{0,s\cup i} = v_{0,s\cup j}$ for every subset $s \subseteq (N \setminus \{i, j\})$, then $\psi_{0,i} = \psi_{0,j}$. Finally, if $v_{0,s\cup j} = v_{0,s}$ for all $s \subseteq (N \setminus j)$, then $\psi_{0,j} = 0$.

The weighted average in (5.2) suggests an equivalent formulation of the population Shapley values. Let v_0 denote the 2^p-dimensional vector of measures of predictiveness for all possible subsets of $\mathcal{S}$. We define $W \in \mathbb{R}^{2^p \times 2^p}$ to be a matrix of weights, where $W_{11} = W_{2^p,2^p} = 1$, and for any $j \in 2, \ldots, 2^p - 1$, $W_{jj} = \binom{p-2}{j-1}^{-1}$. 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 $\psi_{0,\emptyset} := V(f_{0,\emptyset}, P_0)$, we denote by ψ_0 the $(p+1)$-dimensional vector of population Shapley values. We may then write that

$$\psi_0 := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{ \| \sqrt{W}(Z\psi - v_0) \|_2^2 + \lambda^\top (G\psi - c_0) \}, \tag{5.3}$$

where $\lambda = [\lambda_1, \lambda_2]^\top \geq 0$, $G = [z(\emptyset)^\top, z(N)^\top]^\top \in \mathbb{R}^{2 \times (p+1)}$, and $c_0 = [v_{0,\emptyset}, v_{0,N}]^\top \in \mathbb{R}^2$, a result that we prove in Appendix C. This weighted least squares formulation of the Shapley values is related to previous formulations, but explicitly codes the equality constraints and allows $v_{0,\emptyset}$ to be nonzero [cf. Charnes et al., 1988, Lundberg and Lee, 2017, Aas et al., 2019]. If we define the distribution Q_0 as assigning weight $\binom{p-2}{|S|-1}^{-1}$ for $S \in \mathcal{S} \setminus \{\emptyset, N\}$ and weight

1 for $S \in \{\emptyset, N\}$, then (5.3) is equivalent to a population average:

$$\psi_0 \equiv \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{E_{Q_0}(z(S)\psi - v_{0,S})^2 + \lambda^\top (G\psi - c_0)\}. \tag{5.4}$$

5.2.3 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 (5.2) (and similarly (5.3)) allows us to define a local version of variable importance: for a subpopulation $A \subseteq \mathcal{X}$,

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

where we have simply plugged the conditional distribution $P_{0|X\in A}$ into (5.2). Taken to the extreme, where the subpopulation A consists only of a single observation, this definition of local feature importance is equivalent to the Shapley values considered by Lundberg and Lee [2017], though here we allow an arbitrary measure of predictiveness in place of the conditional expectation. We emphasize here that valid statistical inference on this individual-observation-level importance appears difficult, if not impossible, without the aid of computationally expensive procedures such as the bootstrap [Lundberg et al., 2018].

We have also focused exclusively on the Shapley values for each individual feature. However, it may be of interest to determine the importance of a subgroup of features. To do so, we must define a meaningful partition $\mathcal{P}_k := \{s_1, \ldots, s_k \in \mathcal{S} : \bigcup_{i=1}^k s_i = N \text{ and } s_i \bigcap s_j = \emptyset \text{ for any } (i,j) \text{ pair}\}$ of the features into k disjoint groups. Then the group Shapley values may be determined as in (5.2), though the sum is taken over the groups in the partition $\mathcal{P}_k$.

5.2.4 Examples of predictiveness measures

We now list several common VIMs that fall within our proposed framework. The conditional mean functions $\mu_0 : x \mapsto E_0(Y \mid X = x)$ and $\mu_{0,s} : x \mapsto E_0(Y \mid X_s = x_s)$ under P_0 play

a prominent role in the examples below. This is convenient, because μ_0 is the target of estimation for many standard machine learning algorithms, and $\mu_{0,s}$ may be viewed as the target of these algorithms restricted to only X_s.

Example 1: R^2

The R^2 predictiveness measure is defined as $V(f, P_0) := 1 - E_0\left\{Y - f(X)\right\}^2 / \sigma_0^2$, where $\sigma_0^2 := E_0\left\{Y - E_0\left(Y\right)\right\}^2 = E_0\left[Y - E_0\left\{\mu_0(X)\right\}\right]^2$ is the variance of Y under P_0. The optimizer of $V(f, P_0)$ is given by $f_0 = \mu_0$ as long as $\mu_0 \in \mathcal{F}$. For this predictiveness measure, $v_{0,\emptyset} = 0$.

Example 2: deviance

When Y is binary, the deviance predictiveness measure is defined as

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

where $\pi_0 := P_0\left(Y = 1\right)$ is the marginal success probability of Y under P_0. This measure is a scaled version of cross-entropy. Again, we find the optimizer of $f \mapsto V(f, P_0)$ to be $f_0 = \mu_0$ for any rich enough $\mathcal{F}$. For this predictiveness measure, $v_{0,\emptyset} = 0$.

Example 3: classification accuracy

Classification accuracy, defined as $V(f, P_0) = P_0\left\{Y = f(X)\right\}$, is an alternative predictiveness measure for binary outcomes. The Bayes classifier $b_0 : x \mapsto I\{\mu_0(x) > 1/2\}$ maximizes $f \mapsto V(f, P_0)$, and so, $f_0 = b_0$ as long as $b_0 \in \mathcal{F}$. For this predictiveness measure, $v_{0,\emptyset} = P_0(Y = 1)$.

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 $V(f, P_0) = P_0\{f(X_1) < f(X_2) \mid Y_1 = 0, Y_2 = 1\}$, where (X_1, Y_1) and (X_2, Y_2) represent independent draws from P_0. Once more, $f_0 = \mu_0$ provided $\mu_0 \in \mathcal{F}$. For this predictiveness measure, $v_{0,\emptyset} = 0.5$.

5.3 Estimation and inference

5.3.1 Plug-in estimation

In our framework, the variable importance of all available features under P_0, denoted by ψ_0, is a population parameter. Thus, inferring about ψ_0 from the available data provides an assessment of variable importance. More formally, our goal is to construct a nonparametric efficient estimator of ψ_0 using independent observations $O_1, \ldots, O_n$ from P_0. Such an estimator admits the construction of valid confidence intervals and hypothesis tests that have correct type I error control.

Definition (5.3) suggests considering the plug-in estimator

$$\psi_{0,n} := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{ \| \sqrt{W}(Z\psi - v_n) \|_2^2 + \lambda^\top (G\psi - c_n) \},$$

where $v_n \in \mathbb{R}^{2^p}$ is a vector of estimated predictiveness measures $v_{n,s} := V(f_{n,s}, P_n)$ for each $s \in \mathcal{S}$; here, P_n is the empirical distribution based on $O_1, \ldots, O_n$ and $f_{n,s}$ is an estimator of the population optimizer $f_{0,s}$. Often, $f_{n,s}$ is obtained by building a predictive model for the outcome Y using only those features in X_s. It is natural to use machine learning algorithms for this task, which may involve tuning via cross-validation. However, obtaining all 2^p estimators is typically an intractable task. Instead, definition (5.4) suggests considering the plug-in estimator

$$\psi_{m,n} := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{ E_{Q_m}(z(S)\psi - v_{n,S})^2 + \lambda^\top (G\psi - c_n) \}, \tag{5.5}$$

where Q_m is the empirical distribution based on sampling $S_1, \ldots, S_m$ from $\mathcal{S}$. Additionally, as shown in Chapter 4, using cross-fitting to obtain an estimator $v^*_{n,S}$ and using v^*_n and $v^*_{n,S}$ in place of v_n and $v_{n,S}$ above often results in improved performance. This plug-in estimator may be obtained using stochastic gradient descent and the Karush-Kuhn-Tucker conditions for (5.3); we describe the construction of this estimator in detail in Algorithm 5. We have

found that using AMSGrad [Reddi et al., 2019], a modification of Adam [Kingma and Ba, 2014], to perform stochastic gradient updates results in improved empirical performance over classical stochastic gradient descent. In the algorithm, we use element-wise operations unless otherwise specified. The algorithm relies on the fact that obtaining a sample from Q_0 is equivalent to sampling a subset size Q from the distribution on subset sizes with probability given by the normalized weights in Q_0, and then sampling a subset S with size q uniformly from $\mathcal{S}$. The algorithm scales linearly with m, and scales in p according to the tools used to estimate the optimizers $f_{0,s}$. We have found that using initial parameters $\alpha = 0.05$, $\beta_{1t} = 0.9$ for all $t = 1, \ldots, m$, $\beta_2 = 0.99$, and $\epsilon = 1 \times 10^{-16}$ works well in simulated examples.

This plug-in estimator is appealing due to its simplicity. In general, however, such an estimator may fail to be consistent at rate $n^{-1/2}$ if the optimizers $f_{0,s}$ are flexibly estimated. In Chapter 4, we described conditions under which plug-in estimators of the conditional variable importance $v_{0,N} - v_{0,N\setminus s}$ are efficient even if flexible tools are used to estimate $\{f_{0,s}\}_{s\in\mathcal{S}}$. Below, we show that under an additional condition on the number of subsets sampled from $\mathcal{S}$, the same holds true for the plug-in estimator $\psi_{m,n}$.

5.3.2 Large-sample inferential properties

We now study conditions under which $\psi_{m,n}$ is a nonparametric estimator of the Shapley values ψ_0 and describe how to conduct valid inference on ψ_0. The behavior of $\psi_{m,n}$ can be studied by first decomposing

$$\psi_{m,n} - \psi_0 = (\psi_{0,n} - \psi_0) + (\psi_{m,0} - \psi_0) + r_{m,n}, \tag{5.6}$$

where $\psi_{m,0}$ is obtained by replacing $v_{n,S}$ with $v_{0,S}$ in (5.5) and $r_{m,n} := (\psi_{m,n} - \psi_{m,0}) - (\psi_{0,n} - \psi_0)$. Each term on the right-hand side of (5.6) can then be studied separately to determine the large-sample behavior of $\psi_{m,n}$. The first term is the contribution from having had to estimate the optimizers $\{f_{0,s} : s \in S_1, \ldots, S_m\}$ and the empirical distribution P_n, and may be controlled using results from Chapter 4. The second term is the contribution from sampling

Algorithm 5 Estimation of VIM value ψ_0 using K-fold cross-fitting and AMSGrad

input initial parameters α, β_1, β_2, ϵ
choose a step size, e.g., $\alpha_t = \alpha/\sqrt{t}$ for $t = 1, \ldots, m$.
compute $v^*_{n,\emptyset}$ and $v^*_{n,N}$ based on K-fold cross-fitting.
initialize $\psi^{(0)}_{m,n} = [v^*_{n,\emptyset}, (v^*_{n,N}/p)\mathbf{1}_p]$, where $\mathbf{1}_p \in \mathbb{R}^p$ is a vector of all ones.
initialize $\lambda^{(0)}_{m,n} = G\psi^{(0)}_{m,n} - h$.
for $t = 1, \ldots, m$ **do**
  initialize $m_\lambda, m_\psi, v_\lambda, v_\psi, \hat{v}_\lambda, \hat{v}_\psi$ as zero-vectors of the appropriate dimension.
  sample a subset size Q from the distribution on subset sizes.
  sample S with size Q uniformly from $\mathcal{S}$.
  compute $v^*_{n,S}$ using K-fold cross-fitting.
  $g_\lambda \leftarrow G\psi^{(i-1)}_{m,n} - c_n$.
  $m^{(t)}_\lambda \leftarrow \beta_{1t} m^{(t-1)}_\lambda + (1-\beta_{1t}) g_\lambda$, $v^{(t)}_\lambda \leftarrow \beta_2 v^{(t-1)}_\lambda + (1-\beta_2) g^2_\lambda$, $\hat{v}_\lambda = max(\hat{v}^{(t-1)}_\lambda, v^{(t)}_\lambda)$.
  $\lambda^{(t)}_{m,n} \leftarrow \lambda^{(t-1)}_{m,n} - \alpha_t m^{(t)}_\lambda / \left(\sqrt{\hat{v}^{(t)}_\lambda} + \epsilon\right)$.
  $g_\psi \leftarrow (-2) z(S)^\top [v^*_{n,S} - z(S)\psi^{(t-1)}_{m,n}] - G^\top \lambda^{(t)}_{m,n}$.
  $m^{(t)}_\psi \leftarrow \beta_{1t} m^{(t-1)}_\psi + (1-\beta_{1t}) g_\psi$, $v^{(t)}_\psi \leftarrow \beta_2 v^{(t-1)}_\psi + (1-\beta_2) g^2_\psi$, $\hat{v}_\psi = max(\hat{v}^{(t-1)}_\psi, v^{(t)}_\psi)$.
  $\psi^{(t)}_{m,n} \leftarrow \psi^{(t-1)}_{m,n} - \alpha_t m^{(t)}_\psi / \left(\sqrt{\hat{v}^{(t)}_\psi} + \epsilon\right)$.
end for
set $\psi_{m,n} := \frac{1}{m} \sum_{t=1}^{m} \psi^{(t)}_{m,n}$.

subsets from $\mathcal{S}$. The third term is a difference-in-differences remainder term that we can expect to tend to zero at rate $n^{-1/2}$ under some regularity conditions.

Our first result makes use of several conditions from Chapter 4 that we restate here. These conditions require additional notation. Below, 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 $R \in \mathcal{R}$, say $R = c(P_1 - P_2)$, we refer to 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. 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 $h \in \mathcal{R}$, and define the random function $g_n : o \mapsto \dot{V}(f_{n,N}, P_0; \delta_o - P_0) - \dot{V}(f_0, P_0; \delta_o - P_0)$, where δ_o is the degenerate distribution on $\{o\}$. We first define a set of conditions for the estimator $v_{n,N}$:

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

(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

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

(A3) (*minimum rate of convergence*) $\|f_{n,N} - f_0\|_{\mathcal{F}} = o_P(n^{-1/4})$;

(A4) (*weak consistency*) $\int \{g_n(o)\}^2 dP_0(o) = o_P(1)$;

A collection of conditions may be defined by replacing all instances of $f_{n,N}$, f_0, and $\mathcal{F}$ by $f_{n,s}$, $f_{0,s}$, and $\mathcal{F}_s$ in the conditions above. We define the vector-valued function $\dot{V}_0 : o \mapsto [\dot{V}(f_{0,\emptyset}, P_0; \delta_o - P_0), \ldots, \dot{V}(f_{0,N}, P_0; \delta_o - P_0)]^\top$, and the dual solution λ^* of the minimization problem (5.3). Additionally, we set $A := (Z^\top W Z)$ and $e_{\emptyset,2^p} := [e_1^\top, e_{2^p}^\top]^\top$, where e_j is the unit vector corresponding to position j. The final condition specifies the number of subsets to sample:

(A5) (*minimum number of subsets*) For each sequence $\gamma, \gamma_1, \ldots \in \mathbb{R}$ satisfying that $|\gamma_j - \gamma| \to 0$, $m = \gamma_n n$.

Theorem 5. *If the collection of conditions implied by (A1)–(A4) for each subset $s \in \mathcal{S}$ holds and (A5) holds, then $\psi_{m,n}$ is an asymptotically linear estimator of ψ_0 with influence function equal to the nonparametric efficient influence function*

$$\begin{aligned}\phi_0 : (o, s) \mapsto & A^{-1}\left[Z^\top W - G^\top\{GA^{-1}G^\top\}^{-1}\{GA^{-1}Z^\top W - e_{\emptyset,2^p}\}\right]\dot{V}_0(o) \\ & + \gamma^{-1}A^{-1}[z(s)^\top\{v_{0,s} - z(s)\psi_0\} + G^\top\lambda^*].\end{aligned}$$

This result implies, in particular, that $\psi_{m,n}$ is a consistent, asymptotically normal, and nonparametric efficient estimator of ψ_0. If $0 < \sigma^2_{0,j} := E_0\{\phi_0(O,S)\phi_0(O,S)^\top\}_{jj} < \infty$ for each $j = 1, \ldots, p$, this suggests that the asymptotic variance of $n^{1/2}(\psi_{m,n,j} - \psi_{0,j})$ can be estimated by $\sigma^2_{n,j}$, an estimator resulting from plugging in consistent estimators of each component. These estimators may be constructed even if not all elements $S \in \mathcal{S}$ are sampled. In this case, we use consistent estimators A_m, Z_m, and W_m of A, Z, and W above. The explicit form of $\sigma^2_{n,j}$ is then given by

$$\begin{aligned}\sigma^2_{n,j} := & \frac{1}{n}\sum_{i=1}^{n}\{A_m^{-1}\left[Z_m^\top W_m - G^\top\{GA_m^{-1}G\}^{-1}\{GA_m^{-1}Z_m^\top W_m - e_{\emptyset,2^p}\}\right]\dot{V}_0(O_i)\}_j^2 \\ & + \frac{1}{m}\sum_{\ell=1}^{m}\{\gamma_n^{-1}A_m^{-1}\left[v_{n,S_\ell} - z(S_\ell)\psi_{m,n} + G^\top\lambda^*\right]\}_j^2\end{aligned}$$

Conditions (A1)–(A4) are required to control the first-order contribution from estimation of $f_{0,s}$ for each $s \in \mathcal{S}$. These conditions are satisfied for R^2, deviance, accuracy, and AUC (proven in Chapter 4). Finally, condition (A5) is necessary to control the contribution from having had to estimate Q_0. In particular, this condition implies that $m = o(n)$. To our knowledge, this is the first result specifying the number of subsets that it is necessary to sample to construct an efficient estimator of the Shapley values.

5.3.3 Testing the null VIM hypothesis

Our inferential results thus far pertain to the case $\psi_{0,j} > 0$ strictly for each $j = 1, \ldots, p$. When $\psi_{0,j} = 0$, then the efficient influence function $\phi_0(o)_j$ is identically zero. Thus, Wald-type intervals based on $\sigma^2_{n,j}$ will fail to be properly calibrated and tests based on $\sigma^2_{n,j}$ will fail to appropriately control type I error.

The goal of any hypothesis testing procedure is to derive a test that is consistent and controls the type I error rate. In Chapter 4, we found that it is challenging to achieve proper type I error control without sample splitting. In this approach, $v^*_{n,N}$ was estimated using part of the data, and $v^*_{n,-s}$ was estimated using the remaining data. This resulted in a valid test of the δ-null hypothesis $H_0 : v_{0,N} - v_{0,-s} \in [0, \delta]$, where $\delta \geq 0$. Generalizing this approach to test each individual Shapley value would result in low power, since this approach requires splitting the sample into m different groups.

A valid test of the δ-null hypothesis $H_0 : \psi_{0,j} \in [0, \delta]$ may nevertheless be devised based on the comparison of the jth Shapley value to the null Shapley value $\psi_{0,\emptyset}$. This is complicated by the fact that the null Shapley value is nonzero for some measures of predictiveness, leading to the following procedure. Based on one half of the data, we obtain an estimator $\psi_{m,n,j}$ of $\psi_{0,j}$ and an estimator $\sigma^2_{n,j}$ of the variance $\sigma^2_{0,j}$. Based on the other half of the data, we again split the sample and obtain estimators $\psi_{n,\emptyset,1}$ and $\psi_{n,\emptyset,2}$ of $\psi_{0,\emptyset}$, with corresponding variance estimators $\sigma^2_{n,\emptyset,1}$ and $\sigma^2_{n,\emptyset,2}$. Then, we create a test statistic $T_n := \frac{(\psi_{m,n,j} + \psi_{n,\emptyset,1} - \psi_{n,\emptyset,2}) - \delta}{\sqrt{n_1^{-1}\sigma^2_{n,j} + n_2^{-1}\sigma^2_{n,\emptyset,1} + n_3^{-1}\sigma^2_{n,\emptyset,2}}}$, resulting in p-value $p_n := 1 - \Phi(T_n)$, where n_1, n_2, and n_3 denote the respective sample sizes of the split dataset and Φ denotes the standard Normal cumulative distribution function. Finally, we reject H_0 if and only if $p_n < \alpha$ for some pre-specified level α. Under conditions (A1)–(A5), for any $\alpha \in (0, 1)$, the proposed test is consistent and has type I error equal to α. This approach easily extends to the case where the minimal subset is c, consisting of the indices of all potential confounders that are always adjusted for.

5.4 Numerical experiments

In this section, we present empirical results of our proposed plug-in estimator of variable importance. First, we consider three covariates and a continuous outcome generated as

$$\begin{aligned}
X = (X_1, X_2, X_3) &\sim N_3(0, \Sigma) \text{ and } Y \mid X = x \sim N(f(x), 1), \text{ where} \\
f(x) &= \sum_{j=1}^{p} f_j(x_j), f_1(x) = \text{ sign}(x), \\
f_2(x) &= (-6)I(x \leq -4) + (-4)I(-4 < x \leq -2) + (-2)I(0 \leq x < -2) \\
&\quad + 2I(2 < x \leq 4) + 4I(x > 4), \text{ and} \\
f_3(x) &= (-1)I(x \leq -4 \text{ or } -2 < x \leq 0 \text{ or } 2 < x \leq 4) \\
&\quad + I(-4 < x \leq -2 \text{ or } 0 < x \leq 2 \text{ or } x > 4);
\end{aligned}$$

and Σ is a 3×3 identity matrix. We considered here VIMs based on R^2. The true Shapley values implied by this data-generating mechanism are 0.229, 0.312, and 0.229. We generated 1,000 random datasets of size $n \in \{100, 500, 1000, \ldots, 5000\}$, and considered in each case the importance of all three covariates. All analyses were performed using Python version 3.6.7 and may be reproduced using code available online.

To obtain each $f_{n,S}$, we used five-fold cross-validation over a library of candidate estimators. We used as candidate estimators boosted trees from the package `xgboost` with maximum tree depth equal to three, learning rate equal to 10^{-3}, and a number of estimators varying among $\{2000, 4000, \ldots, 10000\}$. Our resulting estimator is the boosted tree algorithm with number of estimators that minimized the cross-validated mean squared error.

We computed the relevant VIM estimators using Algorithm 5 with five-fold cross-fitting for each predictiveness measure. We used initial parameters $\alpha = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1 \times 10^{-16}$; we set $\alpha_t = \alpha/\sqrt{t}$ and $\beta_{1t} = \beta_1$ for $t = 1, \ldots, m$. We then computed the empirical mean squared error (MSE) scaled by n, the empirical coverage of nominal 95% confidence intervals, and the empirical power of our proposed hypothesis test. Finally, for

each feature we computed the mean absolute SHAP value [Lundberg and Lee, 2017] based on fitting the boosted tree algorithm alone.

In Figure 5.1, we display the results of this experiment. In the top-left panel, we see that as n increases, the scaled empirical MSE of our estimator decreases to a fixed level — namely, the scaled empirical variance — regardless of the feature under consideration. This matches our expectations from Section 5.3.2: the scaled empirical bias of our proposed estimator should tend to zero with increasing sample size, while the scaled empirical variance should be constant. The top-right panel shows that the coverage of nominal 95% confidence intervals is at the nominal level. The bottom-left panel shows that our proposed hypothesis test is consistent: for sample size greater than 100, the empirical power is near one. Finally, the bottom-right panel shows the mean absolute SHAP value. While the point estimates are difficult to interpret, we see that regardless of sample size the SHAP procedure correctly identifies that feature two has the largest importance, but estimates the wrong ordering of features one and three.

We plan to run a second scenario of this experiment where rather than Σ being the 3×3 identity matrix, instead the off-diagonal entries are all 0.5. This should highlight the benefits of the Shapley value approach over a conditional importance approach.

Finally, we plan to run an experiment with six variables and a complex, non-additive conditional mean function. This should showcase our proposed stochastic gradient descent algorithm in a scenario where it might be less feasible to directly sample all possible subsets of the power set.

5.5 Predicting mortality of patients in the intensive care unit

We consider here data on patients' stays in the intensive care unit (ICU) [Silva et al., 2012]. These data contain 4000 records on several features: five general descriptors collected upon admission to the ICU; and 37 features — including the Glasgow Coma Score (GCS), blood urea nitrogen (BUN), and heart rate — measured over the course of the first 48 hours after admission to the ICU.

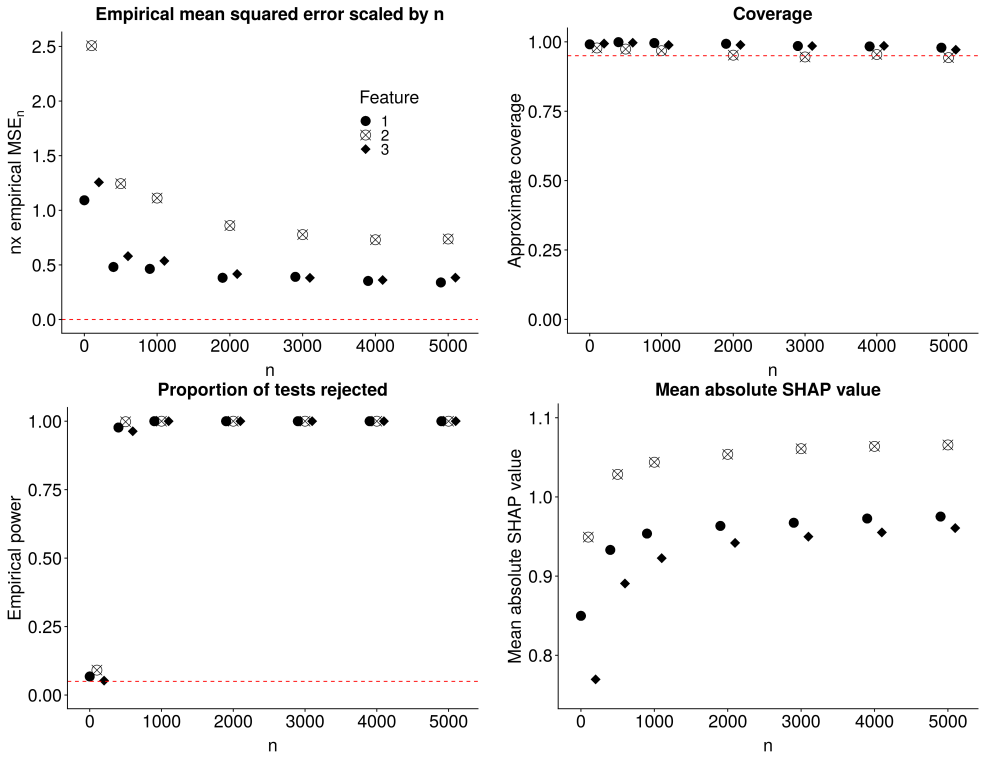

Figure 5.1: Performance of plug-in estimators for estimating the population Shapley values via the difference in R^2. Clockwise from top-left: empirical mean squared error for the proposed plug-in estimator scaled by n vs n for $j \in \{1, 2, 3\}$; approximate coverage of nominal 95% confidence intervals for the true importance vs n; mean absolute SHAP value vs n; and power of the split-sample hypothesis testing procedure vs n when $\delta = 0$. Filled circles denote X_1, crossed circles denote X_2, and filled diamonds denote X_3.

Based on these 42 features, we extracted 55 summary features for prediction, following Feng et al. [2018]. These summary features include the general descriptors mentioned above and the min/max, mean, and last measured value of variables used in the simplified acute physiology (SAPS) I or II scores. We provide a full list of extracted features in Appendix C. These SAPS scores are pre-established scores for estimating the mortality of ICU patients. We estimate the population Shapley values for each variable using AUC to measure predictiveness. Additionally, we provide the mean absolute SHAP value.

We obtained estimates of the conditional means using five-fold cross-validation to select among boosted tree algorithms with maximum tree depth equal to four, learning rate equal to 10^{-3}, and a number of estimators varying among $\{2000, 4000, \ldots, 12000\}$; our resulting estimator is the boosted tree algorithm with number of estimators that minimized the cross-validated negative log likelihood. We again use Algorithm 5 with five-fold cross-fitting for each predictiveness measure, and initial parameters $\alpha = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1 \times 10^{-16}$; we set $\alpha_t = \alpha/\sqrt{t}$ and $\beta_{1t} = \beta_1$ for $t = 1, \ldots, m$.

This analysis was complicated by missing data in some features. The majority of features had less than 2.5% missing data. However, some features — including those related to respiration, lactate, and systolic blood pressure — had between 50% and 75% missing data. This missing data poses challenges to any estimation procedure. The `xgboost` implementation of boosted trees that we used here handles missing data by learning the direction to choose (left or right) at a given split when the feature is missing that minimizes the training loss [Chen and Guestrin, 2016]. This strategy relies on the features being missing at random relative to the measured features, and is appealing since no separate imputation step must be performed prior to fitting the boosting algorithm. However, in some contexts performing the imputation first may be desirable. Finally, the `xgboost` approach makes use of all of the data, rather than keeping only observations with complete information — which would drastically reduce the sample size, and may introduce bias — or keeping only features with complete information — which may ignore important features.

In Figure 5.2, we display the results of this analysis. We display only the top 20 features

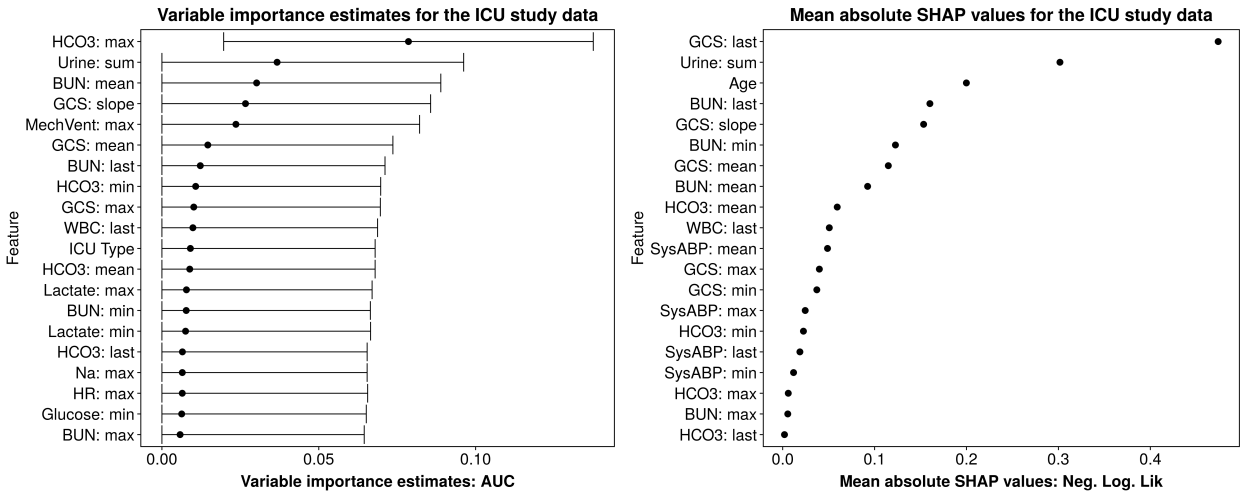

Figure 5.2: Variable importance measured by the difference in AUC (left-hand panel) and the mean absolute SHAP value when the negative log likelihood was used to fit the boosted tree (right-hand panel) for the ICU study data. We display only the top 20 features selected by each method.

selected by each method. The left-hand panel displays the variable importance estimated using the difference in AUCs, while the right-hand panel displays the mean absolute SHAP value. Since the scales for each VIM are different, it is easiest to compare the VIM rankings. The AUC-based VIM results in an estimate that the maximum of the bicarbonate ($HC0_3$) values is the most important; a high value of bicarbonate may result from dehydration or other fluid loss. The next most important variables are estimated to be the sum of the urine-based variables, the average blood urea nitrogen (BUN) level, and the slope of the Glasgow Coma Score (GCS). Each of these variables is computed over the first 48 hours after admission to the ICU. The SHAP-based VIM also estimates that variables based on the GCS, urine, and BUN have large importance in predicting in-hospital death. However, the two methods differ in which summary of these variables is estimated to be most important. This analysis suggests that a follow-up study where measurements of metabolic variables, including bicarbonate and blood urea nitrogen; the GCS; and urine variables are prioritized in triage may be fruitful.

5.6 Discussion

We propose the population Shapley values, defined using an arbitrary predictiveness measure that is relevant for the task at hand, as a global measure of variable importance. Examples of potentially relevant predictiveness measures that fall within our framework are R^2, classification accuracy, the area under the ROC curve, and the deviance. We show how to perform valid statistical inference on the population Shapley values. We also show how to perform a valid hypothesis test of null importance. Our proposed stochastic gradient descent algorithm provides a method for efficiently estimating all Shapley values simultaneously, and provides a bound on the number of feature subsets that must be sampled.

The variable importance measure considered here is a summary of the data-generating mechanism, and as such is a global measure of importance. However, more local importance measures may be obtained by restricting to a smaller subpopulation. This does not extend to single-observation-level importance, which may be of interest in some tasks. There, our procedure can be viewed as an extension of the SHAP values of Lundberg and Lee [2017] to a setting with an arbitrary measure of predictiveness. In these settings, a resampling procedure (e.g., a boostrap) appears necessary for developing intervals.

Supporting Information

Technical details are available in Appendix C. All results may be reproduced using code hosted on the first author's GitHub page.

Acknowledgments

The authors wish to thank Jessica Perry for insightful comments that improved this manuscript. This work was supported by NIH awards F31 AI140836, R01 AI029168, UM1 AI068635 and DP5 OD019820. The opinions expressed in this chapter are those of the authors and do not necessarily represent the official views of the NIAID or the NIH.

Chapter 6
CONCLUDING REMARKS

6.1 Summary

In this dissertation, we have proposed a unified framework for model-agnostic variable importance assessment. In Chapter 3, we focused on an application of the proposed framework to infer about a nonparametric R^2-based variable importance. We used this approach in two additional manuscripts: first, we developed a novel neural network architecture to simultaneously estimate all required conditional means [Feng et al., 2018]; and second, we assessed the importance of various DNA sequence features in predicting HIV-1 neutralization sensitivity [Magaret et al., 2019]. Then, in Chapter 4, we proposed and validated the full framework. We are currently implementing this framework in a follow-up analysis to Magaret et al. [2019], where we develop a general framework to allow prediction of neutralization based on arbitrary combinations of antibodies rather than a single antibody (so long as data on these antibodies are publicly available). Finally, in Chapter 5, we extended the results of Chapter 4 to define the population Shapley values.

6.2 Future work

As described in Chapter 4, our framework allows us to tackle cases involving complex predictiveness measures (e.g., defined in terms of counterfactual outcomes or involving missing data). We studied several common VIMs explicitly in this dissertation. However, there are multiple important areas of future work that we plan to pursue.

We are actively interested in explicitly studying other useful VIMs. One such class is VIMs for time-to-event and censored outcomes, including the difference in the restricted mean survival time and the time-varying area under the receiver operating characteristic curve

[see, e.g., Heagerty et al., 2000]. Both of these measures fall within our proposed framework. However, it is necessary to rigorously evaluate these measures and define new measures that reasonably capture prediction performance in time-to-event and censored outcome settings. Another interesting measure is the *δ-clinical accuracy* $P\{|Y - f(X)| < \delta\}$. This measure quantifies how often the prediction $f(X)$ is within δ of Y, and may be useful in some applications. The δ-clinical accuracy is particularly interesting because its optimizer does not have a closed form.

A second avenue of future work will involve extending our results to longitudinal settings. In this case, we obtain observations $Z_1, \ldots, Z_n$ measured over k time points, resulting in outcome values $Y_{i1}, \ldots, Y_{ik}$, baseline covariates $X_i \in \mathbb{R}^p$, and potentially time-varying covariates $W_{i1}, \ldots, W_{ik} \in \mathbb{R}^q$. It may be of interest to assess either time-varying variable importance or the importance over time using algorithms that can handle longitudinal data structures. It is thus important to develop general results for longitudinal data structures and also specific VIMs that are relevant to longitudinal data.

Another interesting direction involves further exploration of our proposed hypothesis testing procedure in Chapters 4 and 5. It is clear that we are losing power by using sample splitting without data reuse. Thus, developing a more powerful test of the null hypothesis is of great interest. One possible method for improving the test is to perform the proposed procedure many times, each time generating an independent set of data splits, and then average the results. Another is to choose split sizes more judiciously: for example, in Chapter 5, we have not made use of the fact that estimating the predictiveness of the null model $v_{0,\emptyset}$ that does not access any feature information requires a much smaller sample size than estimating $v_{0,\{j\}}$ for any $j \in \{1, \ldots, p\}$. However, both of these procedures involve splitting the data first and then running the analyses on the separate splits, and averaging over many splits. Instead, we might consider running V-fold cross-validated regression analyses on the full dataset, and then averaging over all possible combinations of the V folds into two independent groups. This reduces computation time, and should be valid since estimation of the regression functions has a second-order contribution to estimating the VIM value. Finally,

more complex analytical tools, including approximate higher-order influence functions, may help in developing more powerful tests.

BIBLIOGRAPHY

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

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

AR Barron. Statistical properties of artificial neural networks. In *Proceedings of the 28th IEEE Conference on Decision and Control*, pages 280–285. IEEE, 1989.

D Benkeser and M van der Laan. The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics*, pages 689–696, 2016.

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

J Bi, K Bennett, M Embrechts, C Breneman, and M Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

PJ Bickel, CAJ Klaasen, Y Ritov, and JA Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

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

L Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001b.

L Breiman, J Friedman, CJ Stone, and RA Olshen. *Classification and Regression Trees*. CRC press, 1984.

DR Burton and L Hangartner. Broadly neutralizing antibodies to HIV and their role in vaccine design. *Annual Review of Immunology*, 34:635–659, 2016.

Marco Carone, Ivan Diaz, and Mark J van der Laan. Higher-order targeted loss-based estimation. In Mark J van der Laan and Sherri Rose, editors, *Targeted learning in data science: causal inference for complex longitudinal studies*, chapter 26, pages 483–510. Springer, 2018.

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 Chambaz, P Neuvial, and MJ van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6:1059–1099, 2012.

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.

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

WS Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

MS Cohen, YQ Chen, M McCauley, T Gamble, MC Hosseinipour, et al. Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 365(6): 493–505, 2011.

LE Dodd and MS Pepe. Partial AUC estimation and regression. *Biometrics*, 59(3):614–623, 2003.

K Doksum and A Samarov. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23(5): 1443–1473, 1995.

K Doksum, S Tang, and K-W Tsui. Nonparametric variable selection: the EARTH algorithm. *Journal of the American Statistical Association*, 103(484):1609–1620, 2008.

S Dudoit and MJ van der Laan. *Multiple testing procedures with applications to genomics*. Springer Science & Business Media, 2007.

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.

A Fisher, C Rudin, and F Dominici. 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*, 2018.

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

M Friedman and RH Rosenman. Type A behavior pattern: its association with coronary heart disease. *Annals of Clinical Research*, 3(6):300–312, 1971.

C Frye, I Feige, and C Rowat. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv:1910.06358*, 2019.

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

PB Gilbert, S Self, M Rao, A Naficy, and J Clemens. Sieve analysis: methods for assessing from vaccine trial data how efficacy varies with genotypic and phenotypic pathogen variation. *Journal of Clinical Epidemiology*, 54:68–85, 2001.

PB Gilbert, M Juraska, AC deCamp, S Karuna, S Edupuganti, et al. Basis and statistical design of the passive HIV-1 Antibody Mediated Prevention (AMP) test-of-concept efficacy trials. *Statistical Communications in Infectious Diseases*, 9(1), 2017.

RD Gill, MJ van der Laan, and JA Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 31(3):545–597, 1995.

S Gnanakaran, MG Daniels, T Bhattacharya, AS Lapedes, A Sethi, M Li, , et al. Genetic signatures in the envelope glycoproteins of HIV-1 that associate with broadly neutralizing antibodies. *PLoS Computational Biology*, 6(10):e1000955, 2010.

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

R Guidotti, A Monreale, S Ruggieri, F Turini, F Giannotti, and D Pedreschi. A survey of methods for explaining black box models. *ACM Computer Surveys*, 51(5):93:1–93:42, August 2018. doi: 10.1145/3236009.

D Guo, X Shi, KC Arledge, D Song, L Jiang, et al. A single residue within the V5 region of HIV-1 envelope facilitates viral escape from the broadly neutralizing monoclonal antibody VRC01. *The Journal of Biological Chemistry*, 287(51):43170–43179, 2012. doi: 10.1074/jbc.M112.399402.

D Harrison and DL Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer, 2009.

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

PJ Heagerty, T Lumley, and MS Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

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

LS Huang and J Chen. Analysis of variance, coefficient of determination and F-test for local polynomial regression. *The Annals of Statistics*, 36(5):2085–2109, 2008.

AE Hubbard, S Kherad-Pajouh, and MJ van der Laan. Statistical inference for data adaptive target parameters. *The International Journal of Biostatistics*, 12(1):3–19, 2016.

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

RP Joosten, TA te Beek, E Krieger, ML Hekkelman, and RH Hooft. A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 15(Database issue):D411–419, 2011. doi: 10.1093/nar/gkq1105.

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

F Klein, H Mouquet, P Dosenovic, JF Scheid, L Scharf, and MC Nussenzweig. Antibodies in HIV-1 vaccine development and therapy. *Science*, 341(6151):1199–1204, 2013. doi: 10.1126/science.1241144.

B Korber, M LaBute, and K Yusim. Immunoinformatics comes of age. *PLoS Computational Biology*, 2(6), 2006. doi: 10.1371/journal.pcbi.0020071.

PD Kwong, JR Mascola, and GJ Nabel. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nature Reviews Immunology*, 13(9):693–701, 2013. doi: 10.1038/nri3516.

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

JE Ledgerwood, EE Coates, G Yamshchikov, JG Saunders, L Holman, et al. Safety, pharmacokinetics and neutralization of the broadly neutralizing hiv-1 human monoclonal antibody VRC01 in healthy adults. *Clinical and Experimental Immunology*, 182(3):289–301, 2015. doi: 10.1111/cei.12692.

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

Y Li, LM Walker, X Wu, J Guenaga, Y Feng, et al. Mechanism of neutralization by the broadly neutralizing HIV-1 monoclonal antibody VRC01. *Journal of Virology*, 85(17): 8954–8967, 2011. doi: 10.1128/JVI.00754-11.

S Lipovetsky and M Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

W-Y Loh. Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386, 2002.

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

SM Lundberg, B Nair, MS Vavilala, M Horibe, MJ Eisses, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.

RM Lynch, E Boritz, EE Coates, A DeZure, P Madden, et al. Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. *Science Translational Medicine*, 7(319):319ra206, 2015a. doi: 10.1126/scitranslmed.aad5752.

RM Lynch, P Wong, L Tran, S O'Dell, MC Nason, et al. HIV-1 fitness cost associated with escape from the VRC01 class of CD4 binding site neutralizing antibodies. *Journal of Virology*, 89(8):4201–4213, 2015b. doi: 10.1128/JVI.03608-14.

CA Magaret, DC Benkeser, BD Williamson, BR Borate, LN Carpp, et al. Prediction of VRC01 neutralization sensitivity by HIV-1 gp160 sequence features. *PLoS Computational Biology*, 15(4):e1006952, 2019.

KH Mayer, K Seaton, Y Huang, N Grunenberg, J Hural, et al. Clinical safety and pharmacokinetics of IV and SC VRC01, a broadly neutralizing mAb. Conference on Retroviruses and Opportunistic Infections (CROI), February 2016.

AJ McMichael and T Hanke. HIV vaccines 1983–2003. *Nature Medicine*, 9(7):874–880, 2003. doi: 10.1038/nm0703-874.

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

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.

NCT02568215. Evaluating the safety and efficacy of the VRC01 antibody in reducing acquisition of HIV-1 infection in women. https://clinicaltrials.gov/ct2/show/NCT02568215. Accessed: 2017-10-05. ClinicalTrials.gov identifier NCT02568215.

NCT02716675. Evaluating the safety and efficacy of the VRC01 antibody in reducing acquisition of HIV-1 infection among men and transgender persons. https://clinicaltrials.gov/ct2/show/NCT02716675. Accessed: 2017-10-05. ClinicalTrials.gov identifier NCT02716675.

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

KK Nicodemus, JD Malley, C Strobl, and A Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):110, 2010.

JD Olden and DA Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1):135–150, 2002.

JD Olden, MK Joy, and RG Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 173(3):389–397, 2004.

MS Pepe. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56(2):352–359, 2000.

MS Pepe, T Cai, and G Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229, 2006.

J Pfanzagl. *Contributions to a general asymptotic statistical theory*. Springer, 1982.

DN Politis, JP Romano, and M Wolf. *Subsampling*. Springer, New York, 1999.

SJ Reddi, S Kale, and S Kumar. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

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.

M Rolland and P Gilbert. Evaluating immune correlates in HIV type 1 vaccine efficacy trials: What RV144 may provide. *AIDS Research and Human Retroviruses*, 28(4):400–404, 2012.

M Rolland, S Tovanabutra, AC deCamp, N Frahm, PB Gilbert, et al. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nature Medicine*, 17: 366–371, 2011. doi:10.1038/nm.2316.

M Rolland, PT Edlefsen, BB Larsen, S Tovanabutra, E Sanders-Buell, et al. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env-V2. *Nature*, 490(7420): 417–420, 2012. doi: 10.1038/nature11519. PubMed PMCID: PMC3551291.

J Rosseauw, J Du Plessis, A Benade, P Jordann, J Kotze, P Jooste, and J Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64(12): 430–436, 1983.

R Samworth. A note on methods of restoring consistency to the bootstrap. *Biometrika*, 90 (4):985–990, 2003.

S Sapp, MJ van der Laan, and K Page. Targeted estimation of binary variable importance measures with interval-censored outcomes. *The International Journal of Biostatistics*, 10 (1):77–97, 2014.

M Sarzotti-Kelsoe, RT Bailer, E Turk, CL Lin, M Bilska, et al. Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *Journal of Immunological Methods*, 409:131–146, 2014.

J Shao. Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, 122(4):1251–1262, 1994.

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

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

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.

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

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

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

ML Thompson and W Zucchini. On the statistical analysis of ROC curves. *Statistics in Medicine*, 8(10):1277–1290, 1989.

P Utachee, P Isarangkura-na ayuthaya, K Tokunaga, K Ikuta, N Takeda, and M Kameoka. Impact of amino acid substitutions in the V2 and C2 regions of human immunodeficiency virus type 1 CRF01_AE envelope glycoprotein gp120 on viral neutralization susceptibility to broadly neutralizing antibodies specific for the CD4 binding site. *Retrovirology*, 11(1): 32, 2014. doi: 10.1186/1742-4690-11-32. PubMed PMCID: PMC4003292.

MJ van der Laan. Efficient and inefficient estimation in semiparametric models, 1991.

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

MJ van der Laan and S Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

MJ van der Laan, EC Polley, and AE Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Online Article 25, 2007.

AW van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

AW van der Vaart and JA Wellner. Weak convergence and empirical processes. Springer, 1996.

K Wagh, T Bhattacharya, C Williamson, A Robles, M Bayne, et al. Optimal combinations of broadly neutralizing antibodies for prevention and treatment of HIV-1 clade C infection. *PLoS Pathogens*, 12(3):e1005520, 2016.

K Wagh, MS Seaman, M Zingg, T Fitzsimons, DH Barouch, et al. Potential of conventional & bispecific broadly neutralizing antibodies for prevention of HIV-1 subtype A, C & D infections. *PLoS Pathogens*, 14(3):e1006860, 2018.

LM Walker, M Huber, KJ Doores, E Falkowska, R Pejchal, et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477(7365):466–470, 2011. doi: 10.1038/nature10373.

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

CK Wibmer, JN Bhiman, ES Gray, N Tumba, SSA Karim, et al. Viral escape from HIV-1 neutralizing antibodies drives increased plasma neutralization breadth through sequential recognition of multiple epitopes and immunotypes. *PLoS Pathogens*, 9(10):e1003738, 2013. doi: 10.1371/journal.ppat.1003738.

CK Wibmer, PL Moore, and L Morris. HIV broadly neutralizing antibody targets. *Current Opinion in HIV and AIDS*, 10(3):135, 2015.

BD Williamson, PB Gilbert, N Simon, and M Carone. Nonparametric variable importance assessment using machine learning techniques. *UW Biostatistics Working Paper Series*, Working Paper 422, 2017.

DH Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

X Wu, ZY Yang, CM Hogerkorp, WR Schief, MS Seaman, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*, 329(5993): 856–861, 2010. doi: 10.1126/science.1187659.

F Yao, HG Müller, and JL Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

H Yoon, J Macke, AP Jr West, B Foley, PJ Bjorkman, et al. CATNAP: a tool to compile, analyze, and tally neutralizing antibody panels. *Nucleic Acids Research*, 43(W1):W213–219, 2015. doi: 10.1093/nar/gkv404.

W Zheng and MJ van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper 273*, 2010.

W Zheng and MJ van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.

T Zhou, I Georgiev, X Wu, Z-Y Yang, K Dai, et al. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*, 329(5993):811–817, 2010.

Appendix A

SUPPORTING INFORMATION FOR CHAPTER 3

A.1 Proofs of lemmas and theorems

Throughout, for brevity of notation, we take Pf to denote $\int f(x)dP(x)$ for any measure P and P-measurable function f. We define the full and reduced conditional means for a measure P as

$$\mu_P(x) := E_P(Y \mid X = x) \text{ and } \mu_{P,-s}(x) := E_P(Y \mid X_{-s} = x_{-s}) ,$$

where for any $p-$dimensional vector v and a subset $s \subseteq \{1, 2, \ldots, p\}$ the symbol v_{-s} denotes the elements in v with index not in s. The following proofs rely on a study of the statistical functionals

$$\Phi_s(P) := \int \{\mu_P(x) - \mu_{P,-s}(x)\}^2 dP(x) \text{ and } \Psi_s(P) := \frac{\Phi_s(P)}{var_P(Y)} .$$

Proof of Lemma 1. For a given distribution $P \in \mathcal{M}$, we denote by p the density of P with respect to some dominating measure ν. For bounded $h \in L_2(P)$, we can define the parametric submodel $p_\epsilon = (1 + \epsilon h)p$, which is valid for small enough ϵ and has score h at $\epsilon = 0$. Every regular parametric submodel centered at P and with score h at the origin is either of this form or can be approximated arbitrarily well by a submodel of this form. Given that the statistical model $\mathcal{M}$ considered is nonparametric, and that $D_{P,s} \in L_2(P)$ with $PD_{P,s} = 0$, if we show that for any $P \in \mathcal{M}$

$$\frac{\partial}{\partial \epsilon}\Phi_s(P_\epsilon)\bigg|_{\epsilon=0} = \int D_{P,s}(o)h(o)dP(o)$$

then we will have established that $\Phi_s(P)$ is pathwise differentiable at P with efficient influence function $D_{P,s}$ [Bickel et al., 1998].

The evaluation of Φ_s on P_ϵ equals

$$\begin{aligned}\Phi_s(P_\epsilon) &= \iint \{\mu_{P_\epsilon}(x) - \mu_{P_\epsilon,-s}(x)\}^2 dP_\epsilon(o) = \iint \theta_{s,\epsilon}(x) dP_\epsilon(o) \\ &= \iint \theta_{s,\epsilon}(x)\{1+\epsilon h(x,y)\}p(x,y)\nu(dx,dy) \\ &= \iint \theta_{s,\epsilon}(x)p(x,y)\nu(dx,dy) + \epsilon \iint \theta_{s,\epsilon}(x)h(x,y)p(x,y)\nu(dx,dy) \ ,\end{aligned}$$

where $\theta_{s,\epsilon}(x) := \{\mu_{P_\epsilon,-s}(x) - \mu_{P_\epsilon}(x)\}^2$, and so, we have that

$$\frac{\partial}{\partial\epsilon}\Phi_s(P_\epsilon)\Big|_{\epsilon=0} = \iint \frac{\partial}{\partial\epsilon}\theta_{s,\epsilon}(x)\Big|_{\epsilon=0} p(x,y)\nu(dx,dy) + \iint \theta_s(x)h(x,y)p(x,y)\nu(dx,dy) \ , \quad \text{(A.1)}$$

where $\theta_s = \theta_{s,\epsilon}|_{\epsilon=0}$. Using basic laws of probability, and with some abuse of notation, we can write $\theta_{s,\epsilon}(x)$ in terms of p and h as

$$\theta_{s,\epsilon}(x) = \left[\frac{\int y\{1+\epsilon h(x,y)\}p(x,y)\nu(dy)}{\int\{1+\epsilon h(x,y)\}p(x,y)\nu(dy)} - \frac{\iint y\{1+\epsilon h(x,y)\}p(x,y)\nu(dx_s,dy)}{\iint\{1+\epsilon h(x,y)\}p(x,y)\nu(dx_s,dy)}\right]^2$$

and we can then compute that $\frac{\partial}{\partial\epsilon}\theta_{s,\epsilon}(x)\big|_{\epsilon=0}$ equals

$$2\{\mu_P(x) - \mu_{P,-s}(x)\}\left[\frac{\int\{y-\mu_P(x)\}h(x,y)p(x,y)\nu(dy)}{\int p(x,y)\nu(dy)} - \frac{\iint\{y-\mu_{P,-s}(x)\}h(x,y)p(x,y)\nu(dx_s,dy)}{\iint p(x,y)\nu(dx_s,dy)}\right] .$$

In view of (A.1), this allows us to write that

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\Phi_s(P_\epsilon)\Big|_{\epsilon=0} &= \iint \left[2\{\mu_P(x) - \mu_{P,-s}(x)\}\{y-\mu_P(x)\} + \theta_s(x)\right] h(x,y)p(x,y)\nu(dx,dy) \\ &= \iint \left[2\{\mu_P(x) - \mu_{P,-s}(x)\}\{y-\mu_P(x)\} + \theta_s(x) - \Phi_s(P)\right] h(x,y)p(x,y)\nu(dx,dy)\end{aligned}$$

as required, where to obtain the first line we used that $\mu_P(X) - \mu_{P,-s}(X)$ has mean zero conditionally upon $X_{-s} = x_{-s}$ as a simple consequence of the law of total expectation, and

to obtain the second line we used that $\iint h(x,y)p(x,y)\nu(dx,dy) = 0$.

Because Ψ_s is the ratio of two parameters, namely Φ_s and the population outcome variance parameter, both of which are pathwise differentiable and have known efficient influence functions relative to nonparametric models, it follows that Ψ_s is itself pathwise differentiable at each $P \in \mathcal{M}$. Furthermore, its efficient influence function can readily be found using the delta method. We will use the fact that the parameter $P \mapsto var_P(Y)$ has nonparametric efficient influence function given by

$$o \mapsto D_{P,v}(o) := \{y - E_P(Y)\}^2 - var_P(Y) \ .$$

It follows then that the nonparametric efficient influence function of Ψ_s at P equals

$$\begin{aligned} o \mapsto D^*_{P,s}(o) &= \frac{D_{P,s}(o)var_P(Y) - D_{P,v}(o)\Phi_s(P)}{var^2_P(Y)} \\ &= \frac{2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,-s}(x)\} + \{\mu_P(x) - \mu_{P,-s}(x)\}^2 - \Phi_s(P)}{var_P(Y)} \\ &\quad - \frac{[\{y - E_P(Y)\}^2 - var_P(Y)]\Phi_s(P)}{\{var_P(Y)\}^2} \ . \end{aligned}$$

□

Proof of Lemma 2. We can express the expansion of interest using the Pf notation described above as

$$\Phi_s(P) - \Phi_s(P_0) \ = \ (P - P_0)D_{P,s} + R_s(P, P_0) \ = \ -P_0 D_{P,s} + R_s(P, P_0) \ ,$$

where we have used the fact that $PD_{P,s} = 0$ since, by definition, $D_{P,s}(O)$ has mean zero under P. This implies that the form of $R_s(P, P_0)$ can be derived as $\Psi_s(P) - \Psi_s(P_0) + P_0 D_{P,s}$. The explicit form provided in Lemma 2.2 can be obtained from this expression as follows:

$$
\begin{aligned}
R_s(P, P_0) &= \Phi_s(P) - \Phi_s(P_0) + P_0 D_{P,s} \\
&= \Phi_s(P) - P_0\{(\mu_{P_0} - \mu_{P_0,s})^2\} + 2P_0\{(\mu_P - \mu_{P,-s})(\mu_{P_0} - \mu_P)\} + P_0\{(\mu_P - \mu_{P,-s})^2\} \\
&\quad - \Phi_s(P) \\
&= P_0\{(\mu_P - \mu_{P,-s})^2\} - P_0\{(\mu_{P_0} - \mu_{P_0,s})^2\} + 2P_0\{(\mu_P - \mu_{P,-s})(\mu_{P_0} - \mu_P)\} \\
&= P_0\{(\mu_{P_0,s} - \mu_{P,-s})^2 - (\mu_{P_0} - \mu_P)^2\} \ ,
\end{aligned}
$$

where the last line is obtained by arithmetic manipulations. This directly implies that $R_s(\widehat{P}_n, P_0) = o_P(n^{-1/2})$ if and only if $\hat{\mu} - \mu_{P_0}$ and $\hat{\mu}_{-s} - \mu_{P_0,-s}$ are both $o_P(n^{-1/4})$ in $L_2(P_0)$ norm. □

Proof of Lemma 3. This is a direct application of Lemma 19.24 of van der Vaart [2000]. □

Proof of Theorem 1. Under the conditions of the theorem, we have that $\hat{\phi}_{n,s} - \phi_{0,s} = P_n D_{P_0,s} + o_P(n^{-1/2})$. Additionally, it is easy to verify that $var_{\mathbb{P}_n}(Y) - var_{P_0}(Y) = P_n D_{P_0,v} + o_P(n^{-1/2})$, where $D_{P_0,v}(o) = \{y - E_{P_0}(Y)\}^2 - var_{P_0}(Y)$. By the delta method, it follows then that

$$
\begin{aligned}
\hat{\psi}_{n,s} - \psi_{0,s} \ = \ \frac{\hat{\phi}_{n,s}}{var_{\mathbb{P}_n}(Y)} - \frac{\phi_{0,s}}{var_{P_0}(Y)} \ &= \ P_n \left[\frac{var_{P_0}(Y) D_{P_0,s} - \phi_{0,s} D_{P_0,v}}{var_{P_0}(Y)^2} \right] + o_P(n^{-1/2}) \\
&= \ P_n D^*_{P_0,s} + o_P(n^{-1/2}) \ .
\end{aligned}
$$

In other words, the proposed estimator $\hat{\psi}_{n,s}$ is an asymptotically linear estimator of $\psi_{0,s}$ with influence function $D^*_{P_0,s}$. By the weak law of large numbers, this implies that $\hat{\psi}_{n,s}$ is consistent for $\psi_{0,s}$. It also implies that $\hat{\psi}_{n,s}$ is a regular estimator because its influence function is given by a gradient of the pathwise derivative of Ψ_s. Finally, by the central limit theorem, it implies that $n^{1/2}(\hat{\psi}_{n,s} - \psi_{0,s})$ tends to a mean-zero Gaussian variate with variance $var_{P_0}\{D^*_{P_0,s}(O)\} = P_0 D^{*2}_{P_0,s}$.

□

A.2 Invariance to transformations

In some applications, it is common to center and standardize the features by subtracting their mean and dividing by their standard deviation prior to estimation. In other applications, it is common to transform the outcome or the features using some monotone transformation in order to achieve some form of normalization. It is therefore of interest to determine how such transformations impact the variable importance measure we have proposed. This is what the following result describes.

Theorem 6. *Suppose that* $g_Y : \mathbb{R} \to \mathbb{R}$ *is a linear function, and that* $g_X : \mathbb{R}^p \to \mathbb{R}^p$ *has the form* $(x_1, x_2, \ldots, x_p) \mapsto (g_1(x_1), g_2(x_2), \ldots, g_p(x_p))$ *for invertible functions* $g_j : \mathbb{R} \to \mathbb{R}$, $j = 1, 2, \ldots, p$. *If* $P_{0,g}$ *is the distribution of* $(g_X(X), g_Y(Y))$ *under* P_0, *then* $\Psi_s(P_{0,g}) = \Psi_s(P_0)$.

The proposed variable importance measure is therefore invariant to a wide range of transformations of the underlying data unit, namely linear transformations of the outcome and invertible transformations of each feature. In particular, this implies that the proposed parameter is invariant to univariate linear standardizations of features and the outcome.

The invariance of the proposed variable importance parameter to certain transformations of either the outcome or features ensures that the estimand remains the same after transformation. However, it does not guarantee that the estimate obtained on any particular dataset will enjoy this same invariance property. Nevertheless, variations in the variable importance estimate obtained with and without such transformation are not expected to be large if sufficiently flexible estimators are used and the data set is reasonably large, because both estimators are then consistent for the same estimand. As such, the lack of invariance of the estimator is not expected to pose any practical problem for large data sets, and may be of interest for future research for small data sets. We do note that if the estimation procedure used to obtain conditional mean estimates itself enjoys the same invariance properties as the parameter, finite-sample invariance of the point estimator will then also hold.

Proof of Theorem 6. Take $a, b \in \mathbb{R}$ and consider the transformed outcome $Y^* = a + bY$.

Denoting by $P_{0,a,b}$ the distribution of (X, Y^*) induced by P_0, we can write that

$$\begin{aligned}
\Psi_s(P_{0,a,b}) &= \frac{\int \left\{E_{P_{0,a,b}}(Y^* \mid X = x) - E_{P_{0,a,b}}(Y^* \mid X_{-s} = x_{-s})\right\}^2 dP_{0,a,b}(x)}{var_{P_{0,a,b}}(Y^*)} \\
&= \frac{\int \left\{E_{P_0}(a + bY \mid X = x) - E_{P_0}(a + bY \mid X_{-s} = x_{-s})\right\}^2 dP_0(x)}{var_{P_0}(a + bY)} \\
&= \frac{\int b^2 \left\{E_{P_0}(Y \mid X = x) - E_{P_0}(Y \mid X_{-s} = x_{-s})\right\}^2 dP_0(x)}{b^2 var_{P_0}(Y)} \;=\; \Psi_s(P_0) \;,
\end{aligned}$$

where we have used the linearity of the expectation and the fact that the marginal distribution of X is the same under P_0 and $P_{0,a,b}$.

Suppose the transformation $g_X : \mathbb{R}^p \to \mathbb{R}^p$ has the form

$$(x_1, x_2, \ldots, x_p) \mapsto (g_1(x_1), g_2(x_2), \ldots, g_p(x_p))$$

for invertible functions $g_j : \mathbb{R} \to \mathbb{R}$, $j = 1, 2, \ldots, p$, and let $X^* = g_X(X) = (g_1(X_1), g_2(X_2), \ldots, g_p(X_p))$. Denote by P_{0,g_X} the distribution of (X^*, Y) induced by P_0. For any P, the denominator of $\Psi(P)$ only involves the marginal distribution of Y under P. Because P_0 and P_{0,g_X} induce the same marginal distribution of Y, the denominators of $\Psi_s(P_0)$ and $\Psi_s(P_{0,g_X})$ are identical. This is also true of the numerators since

$$\begin{aligned}
\Phi_s(P_{0,g_X}) &= E_{P_{0,g_X}}\left[E_{P_{0,g_X}}(Y \mid X^*) - E_{P_{0,g_X}}(Y \mid X^*_{-s})\right]^2 \\
&= E_{P_{0,g_X}}\left[E_{P_0}(Y \mid X^*) - E_{P_0}(Y \mid X^*_{-s})\right]^2 \\
&= E_{P_0}\left[E_{P_0}(Y \mid X) - E_{P_0}(Y \mid X_{-s})\right]^2 \\
&= \Phi_s(P_0) \;,
\end{aligned}$$

where in the second line we have used that P_{0,g_X} and P_0 induce the same conditional distribution of Y given any transformation $g_0(X)$ of X, and where the third line follows from the invertibility of g_X. Therefore, we find, as claimed, that $\Psi_s(P_{0,g_X}) = \Psi_s(P_0)$. □

A.3 Cross-validated estimation of variable importance

In Section 3.2 of the main manuscript, we briefly mention that when using very flexible regression estimators, there may be reason for concern regarding the validity of the Donsker class condition in Lemma 3. We now present the cross-validated version of the one-step procedure involved in our proposed estimator, and perform a numerical experiment to provide some empirical evidence that a cross-validation procedure applied to the initial estimator alone is not sufficient for valid inference – in fact, the corrected estimator through a cross-validated version of the one-step procedure is necessary for the required asymptotic behavior.

A.3.1 Estimation procedure

Consider an independent sample $O_1, O_2, \ldots, O_n$ drawn from an unknown distribution P_0 known only to lie in a potentially rich model $\mathcal{M}$, and that the data unit O_i consists of (X_i, Y_i), where $X_i := (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p$ is a covariate vector and $Y_i \in \mathbb{R}$ is the outcome of interest.

Define a V-fold cross-validation scheme, where for $v \in \{1, 2, \ldots, V\}$ we

1. generate a random vector $\mathcal{I} \in \{0, 1\}^n$;

2. define a training sample $\mathcal{T} := \{i : \mathcal{I}_i = 0\}$ and a validation sample $\mathcal{V} := \{i : \mathcal{I}_i = 1\}$;

3. obtain estimators $\hat{\mu}^0$ and $\hat{\mu}^0_{-s}$ using our chosen method of estimating μ_{P_0} and $\mu_{P_0,-s}$ on the training data $\mathcal{T}$ only;

4. obtain an estimator of the parameter of interest on the validation data $\mathcal{V}$ according to

$$\hat{\psi}_{v,s,*} := \left[1 - \frac{\sum_{i\in\mathcal{V}}\{Y_i - \hat{\mu}^0(X_i)\}^2}{\sum_{i\in\mathcal{V}}(Y_i - \bar{Y}_{\mathcal{V}_1})^2}\right] - \left[1 - \frac{\sum_{i\in\mathcal{V}}\{Y_i - \hat{\mu}^0_{-s}(X_i)\}^2}{\sum_{i\in\mathcal{V}}(Y_i - \bar{Y}_{\mathcal{V}})^2}\right],$$

   where $\bar{Y}_{\mathcal{V}} := \frac{1}{\sum_{i=1}^n I(i\in\mathcal{V})} \sum_{i\in\mathcal{V}} Y_i$;

5. estimate the EIF on the validation data $\mathcal{V}$ according to

$$\hat{u}_{v,s} := \frac{\sum_{i \in \mathcal{V}} 2\{Y_i - \hat{\mu}^0(X_i)\}\{\hat{\mu}^0(X_i) - \hat{\mu}^0_{-s}(X_i)\}}{\sum_{i \in \mathcal{V}}(Y_i - \bar{Y}_{\mathcal{V}})^2};$$

6. obtain an estimator of the standard deviation of the proposed estimator on the validation data $\mathcal{V}$ according to

$$\hat{\sigma}_{v,s} := \left[\frac{1}{\sum_{i=1}^n I(i \in \mathcal{V})} \sum_{i \in \mathcal{V}} \{\widehat{D}^*_{P_0,s}(O_i)\}^2\right]^{1/2},$$

where $\widehat{D}^*_{P_0,s}$ may be taken to be $D^*_{P_0,s}$ with μ_{P_0}, $\mu_{P_0,-s}$, $E_{P_0}(Y)$, $var_{P_0}(Y)$ and $\phi_{0,s}$ replaced by $\hat{\mu}^0$, $\hat{\mu}^0_{-s}$, $\bar{Y}_{\mathcal{V}}$, $\frac{1}{\sum_{i=1}^n I(i \in \mathcal{V})} \sum_{i \in \mathcal{V}} (Y_i - \bar{Y}_{\mathcal{V}})^2$ and $\hat{\phi}^0_{n,s}$, with $\bar{Y}_{\mathcal{V}} = \frac{1}{\sum_{i=1}^n I(i \in \mathcal{V})} \sum_{i \in \mathcal{V}} Y_i$ and $\hat{\phi}^0_{n,s}$ the estimator on the validation data $\mathcal{V}$.

The cross-validated estimator is then defined as

$$\hat{\psi}^{\text{cv}}_{n,s,*} := \frac{1}{V} \sum_{v=1}^{V} \hat{\psi}_{v,s,*}; \tag{A.2}$$

as in the main manuscript, this cross-validated estimator utilizes the equivalence between the corrected ANOVA-decomposition-based estimator and the difference in R^2-based estimator. The following lemma describes the asymptotic behavior of this corrected cross-validated estimator.

Lemma 4. *Suppose that both* $\int \{\hat{\mu}(x) - \mu_{P_0}(x)\}^2 dP_0(x)$ *and* $\int \{\hat{\mu}_{-s}(x) - \mu_{P_0,-s}(x)\}^2 dP_0(x)$ *are* $o_P(n^{-1/2})$. *Then, the proposed estimator* $\hat{\psi}^{cv}_{n,s,*}$ *is asymptotically linear with influence function* $D^*_{P_0,s}$. *In particular, this implies that (a)* $\hat{\psi}^{cv}_{n,s,*}$ *tends to* $\psi_{0,s}$ *in probability, (b)* $\hat{\psi}^{cv}_{n,s,*}$ *is regular, and if* $\psi_{0,s} \in (0,1)$, *(c)* $n^{1/2}(\hat{\psi}^{cv}_{n,s,*} - \psi_{0,s})$ *tends in distribution to a mean-zero Gaussian random variable with variance* $\sigma^2_{0,s} := \int \{D^*_{P_0,s}(o)\}^2 dP_0(o)$.

This lemma follows from a straightforward extension of Theorem 1, where we relax the Donsker class conditions using Theorem 25.57 in [van der Vaart, 2000]. A consistent estimator

of the standard deviation of the corrected cross-validated estimator (A.2) is given by

$$\hat{\sigma}_{n,s}^{\text{cv}} := \frac{1}{V} \sum_{v=1}^{V} \hat{\sigma}_{v,s},$$

and an asymptotically valid $(1 - \alpha) \times 100\%$ Wald-type confidence interval for $\psi_{0,s}$ can be obtained as $\hat{\psi}_{n,s}^{\text{cv}} \pm q_{1-\alpha/2} \hat{\sigma}_{n,s}^{\text{cv}} n^{-1/2}$, where q_β is the β-quantile of the standard normal distribution. This result implies that the corrected cross-validated estimator is efficient, regular, and asymptotically linear as long as the conditional means are estimated at a rate of $n^{-1/4}$, which is achievable in practice.

A.3.2 Numerical experiments on a low-dimensional vector of covariates

We now perform an experiment to provide empirical evidence justifying our claim that cross-validation alone does not yield a regular and asymptotically linear estimator of $\psi_{0,s}$. The two settings that we consider are the same as the low-dimensional settings in the main manuscript; in the first setting, we consider data generated according to the following specification:

$$X_1, X_2 \overset{iid}{\sim} \text{Uniform}(-1, 1) \text{ and } \epsilon \sim N(0, 1) \text{ independent of } (X_1, X_2)$$
$$Y = X_1^2 \left(X_1 + \tfrac{7}{5}\right) + \tfrac{25}{9} X_2^2 + \epsilon \, .$$

We generated $1,000$ random datasets of size $n \in \{100, 300, 500, 700, 1000, 2000, \ldots, 8000\}$ and considered in each case the importance of X_j for $j \in \{1, 2\}$. The true value of the variable importance measures implied by this data-generating mechanism can be shown to be $\psi_{0,1} \approx 0.158$ and $\psi_{0,2} \approx 0.342$.

To obtain $\hat{\mu}$ and $\hat{\mu}_{-j}$, we fit locally-linear loess smoothing using the `R` function `loess` with tuning selected to minimize a five-fold cross-validated estimate of the empirical risk based on the squared error loss function. To obtain the naive and corrected cross-validated estimators, we performed an outer layer of five-fold cross validation according to the specification above. Confidence intervals based on the corrected cross-validated estimator were

computed using the cross-validated standard deviation estimator described above. We do not compute bootstrap intervals based on the naive estimator, as this would add a large amount of computation time, and the bias inherent in the naive estimator should yield poor coverage of such intervals.

Figure A.1 displays the results under this alternative hypothesis. In this case, we see a smaller bias of both the naive and the corrected cross-validated estimators than we saw in the simulations in the main manuscript – this is likely due to using locally-linear loess smoothing rather than locally-constant loess smoothing. However, we still see that the naive cross-validated estimator does not have bias going to zero faster $n^{-1/2}$, highlighting that the correction we propose is still necessary, even if sample splitting is used in estimation. The variance of the proposed corrected cross-validated estimator is similar to that of the naive cross-validated estimator, indicating that we have not suffered much from removing the excess bias in the estimation procedure. Finally, confidence intervals based on the corrected cross-validated estimator quickly approach the nominal level of 95%.

Under the null hypothesis for X_2, we generate data according to

$$X_1, X_2 \overset{iid}{\sim} \text{Uniform}(-1, 1) \text{ and } \epsilon \sim N(0, 1) \text{ independent of } (X_1, X_2)$$
$$Y = \tfrac{25}{9} X_1^2 + \epsilon \ .$$

We used the same estimation procedure as above, and computed the same summaries.

Figure A.2 displays the results of this experiment. Here we see that there is much more residual bias in the corrected CV estimator, for the non-null feature. The naive CV estimator again does not have bias that goes to zero faster than $n^{-1/2}$, again highlighting that the correction is necessary. In this situation, as in the null hypothesis simulation in the main manuscript, we see coverage of confidence intervals approaching the nominal level for the non-null feature, but worse coverage for the null feature. However, in this case, we have better coverage than with the non-cross-validated estimator, a phenomenon that could be of interest in future research.

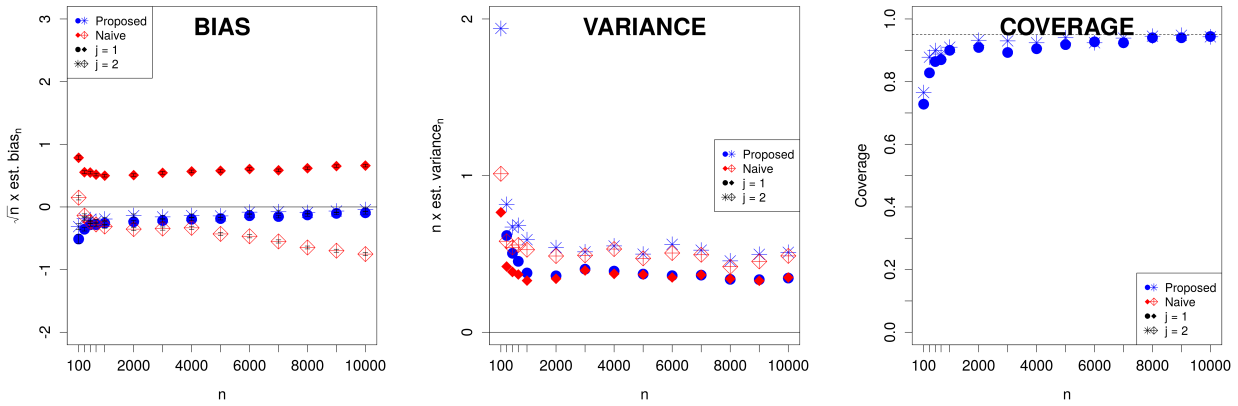

Figure A.1: Empirical bias scaled by $\sqrt{n}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive cross-validated estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to estimate the conditional means. Circles and filled diamonds denote that we have removed X_1, while stars and crossed diamonds denote that we have removed X_2.

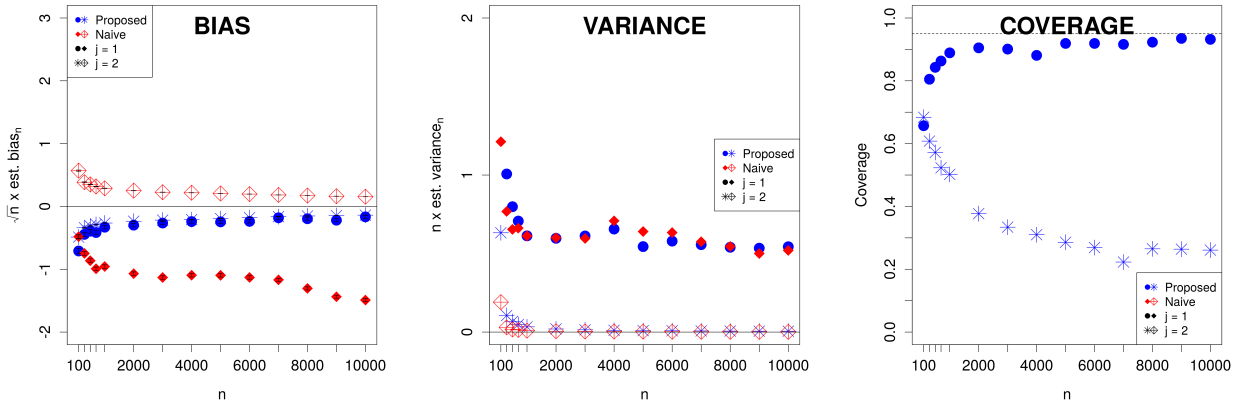

Figure A.2: Empirical bias scaled by $\sqrt{n}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive cross-validated estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to estimate the conditional means. Circles and filled diamonds denote that we have removed X_1, while stars and crossed diamonds denote that we have removed X_2.

Each of these experiments may be reproduced using code on the first author's GitHub page.

A.4 Additional simulation results: moderate-dimensional vector of features

We consider two settings: one in which all of the features are independent, and a second in which groups of features are correlated. In the first setting (setting A), we generate data according to the following specification:

$$X_1, X_2, \ldots, X_{15} \overset{iid}{\sim} N(0,4) \text{ and } \epsilon \sim N(0,1) \text{ independent of } (X_1, X_2, \ldots, X_{15})$$
$$Y = I_{(-2,+2)}(X_1) \cdot \lfloor X_1 \rfloor + I_{(-\infty,0]}(X_2) + I_{(0,+\infty)}(X_3) + \left|\tfrac{X_6}{4}\right|^3 + \left|\tfrac{X_7}{4}\right|^5 + \tfrac{7}{3}\cos\left(\tfrac{X_{11}}{2}\right) + \epsilon \ .$$

We generated 500 random datasets of size $n \in \{100, 300, 500, 1000\}$, and consider the importance of the features included in the sets $\{\{11\}$ and $\{1, 2, 3, 6, 7\}\}$ for each sample size. An analysis of additional groups is provided in the main manuscript. The truth corresponding to each of these situations is given in Table A.1.

In the second setting (setting B), the covariate distribution was modified to include clustering. Specifically, we generated $(X_1, X_2, \ldots, X_{15}) \sim MVN_{15}(\mu, \Sigma)$, where the mean vector is

$$\mu = 3 \times (0,0,0,0,0,1,1,1,1,1,0,0,0,0,0) - 2 \times (0,0,0,0,0,0,0,0,0,0,1,1,1,1,1)$$

and the variance-covariance matrix is given by

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} \end{bmatrix},$$

where we have set

$$\Sigma_{11} = \begin{bmatrix} 1 & 0.15 & 0.15 \\ 0.15 & 1 & 0.15 \\ 0.15 & 0.15 & 1 \end{bmatrix}, \quad \Sigma_{22} = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \quad \text{and} \quad \Sigma_{33} = \begin{bmatrix} 1 & 0.85 & 0.85 \\ 0.85 & 1 & 0.85 \\ 0.85 & 0.85 & 1 \end{bmatrix}$$

and each of Σ_{12}, Σ_{13} and Σ_{23} are three-by-three zero matrices. The random error ϵ and the outcome Y are then generated as in setting A. In this setting, we considered the same sample sizes and groups of features to study as in setting A. The true value of the variable importance measure corresponding to each of the considered groups is also given in Tables 3.1 and A.1. As in setting A, results for the analysis of additional groupings are provided in the main manuscript.

For each of these situations, we estimate the conditional means $E_{P_0}(Y \mid X)$ and $E_{P_0}(Y \mid X_{-s})$ using gradient boosted trees, fit using the `GradientBoostingRegressor` function in the `sklearn` module in Python. We use five-fold cross-validation to select the optimal number of trees with one node, as well as the optimal learning rate for the algorithm. We computed the naive and proposed estimates and respective confidence intervals for each of 500 replications. Because of the unavailability of a simple asymptotic distribution for the naive estimator, a percentile bootstrap approach with 1,000 bootstrap samples was used to attempt to obtain approximate confidence intervals based on $\hat{\psi}_{\text{naive},s}$. For each estimator, we then computed the empirical bias scaled by $n^{1/2}$ and the empirical variance scaled by n. Finally, we computed the empirical coverage of the nominal 95% confidence intervals constructed.

The results from setting A are presented in Figures A.3–A.5. We see that when the features are uncorrelated, on these two groups, the performance of the various estimators considered is similar to the performance showcased in the main manuscript – as n grows the scaled bias of the proposed estimator tends to zero while the scaled bias of the naive estimator tends away from zero, and coverage of confidence intervals based on the proposed estimator tends to the nominal level while coverage of confidence intervals based on the naive estimator remains low. In all settings, we see that variance of the proposed estimator

Table A.1: Approximate values of ψ_0 for each simulation setting and group considered for effect size.

Group	Setting A	B
X_{11}	0.242	0.035
$(X_1, X_2, X_3, X_6, X_7)$	0.535	0.461

is similar to the variance of the naive estimator (Figure A.5).

The results from setting B are a bit different (Figures A.6–A.8). For both groups, we see some residual bias in the proposed estimator, though the magnitude of this bias is smaller than the magnitude of the scaled bias in the naive estimator. We also see some odd behavior in terms of coverage – coverage of confidence intervals based on the proposed estimator is not nearly as good when $s = 11$ under setting B as it was under setting A. However, it is encouraging that the coverage of confidence intervals based on the naive estimator approaches zero as n increases. Finally, we see that the variance of the proposed estimator is still similar to the variance of the naive estimator.

These experiments may be reproduced using code on the first author's GitHub page.

A.5 Results from the Boston housing study data

We consider data on the median house value sampled from 506 neighborhoods in the suburbs of the Boston, Massachusetts metropolitan area. These data come from Harrison and Rubinfeld [1978], and are freely available as part of the R package `MASS`. In addition to the median house value, measurements on four groups of variables are available. The first consists of accessibility features: the weighted distance to five employment centers in the Boston region; and an index of accessibility to radial highways. The second group consists of neighborhood features: the proportion of black residents in the population; the proportion of the population of lower socio-economic status, referring to adults without any high school education or male workers classified as laborers; the crime rate; the proportion of a town's

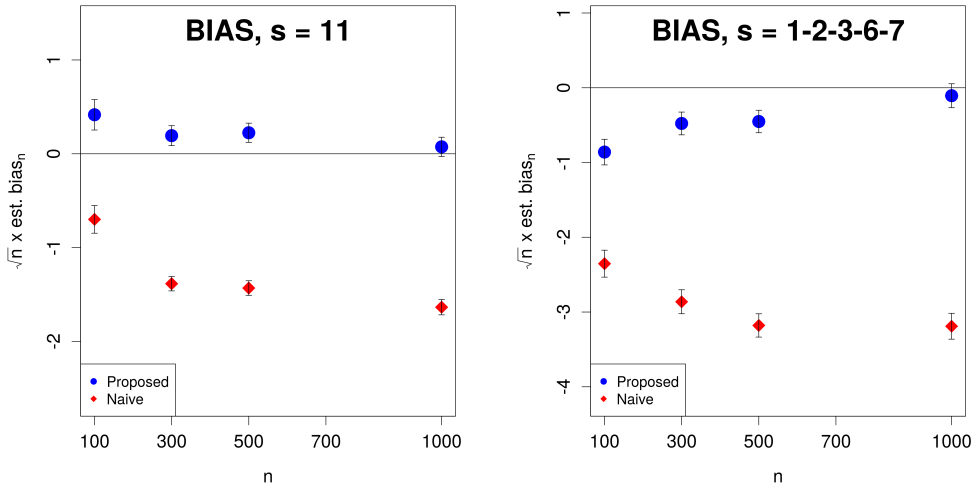

Figure A.3: Empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Tables 3.1 and A.1. Diamonds denote the naive estimator, and circles denote the proposed estimator. Monte Carlo error bars are displayed vertically.

residential land zoned for lots greater than 25,000 square feet; the proportion of non-retail business acres per town; the full value property tax rate; the pupil-teacher ratio by school district; and an indicator of whether the tract of land borders the Charles River. The third group consists of structural features: the average number of rooms in owner units; and the proportion of owner units built prior to 1940. The final group consists of one variable alone: the nitrogen oxide concentration, a measure of air pollution. In our analysis, we considered the variable importance for each individual feature, as well as the natural groups defined above, when predicting the median house value.

We estimate the conditional means using the sequential regression estimating procedure outlined in Section 3.2 of the main manuscript and using the Super Learner [van der Laan et al., 2007] via the `SuperLearner` R package. Our library of candidate learners consists of boosted trees implemented in the `gbm` R package, generalized additive models implemented in the `gam` R package, elastic net implemented in the `glmnet` R package, and random forests implemented in the `randomForest` R package, each with varying tuning parameters. We used

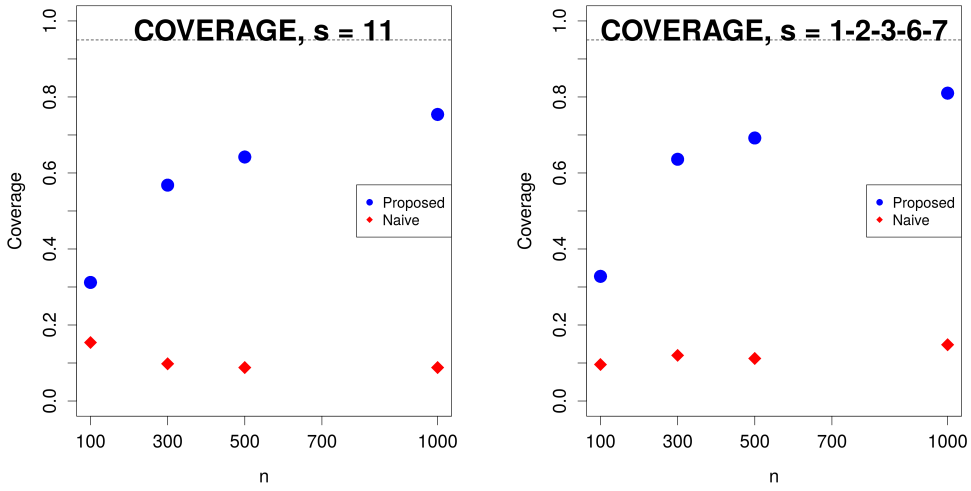

Figure A.4: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table A.1. Diamonds denote the naive estimator, and circles denote the proposed estimator.

ten-fold cross-validation to determine the optimal combination of these learners. This process allowed the Super Learner to determine the optimal tuning parameters for the individual algorithms as part of its optimal combination.

The results are presented in Figure A.9. The group of neighborhood variables appears to be the most important in predicting the median house value; this seems to be driven largely by the proportion of the population of lower socio-economic status. The group of structural variables appears to be the second most important group, and seems to be mostly driven by the average number of rooms in the house, which is also the most important individual feature. Contrary to a naive *a priori* expectation, the crime rate appears to be the least important individual feature in predicting median house value. Finally, we estimate that including all of the covariates in the model explains 97.6% of the variability in median house value, with a 95% confidence interval of (95.7%, 99.6%).

The Boston housing dataset is a popular choice as a benchmark for testing new prediction methods. Hence, there are many estimates of variable importance produced on these data,

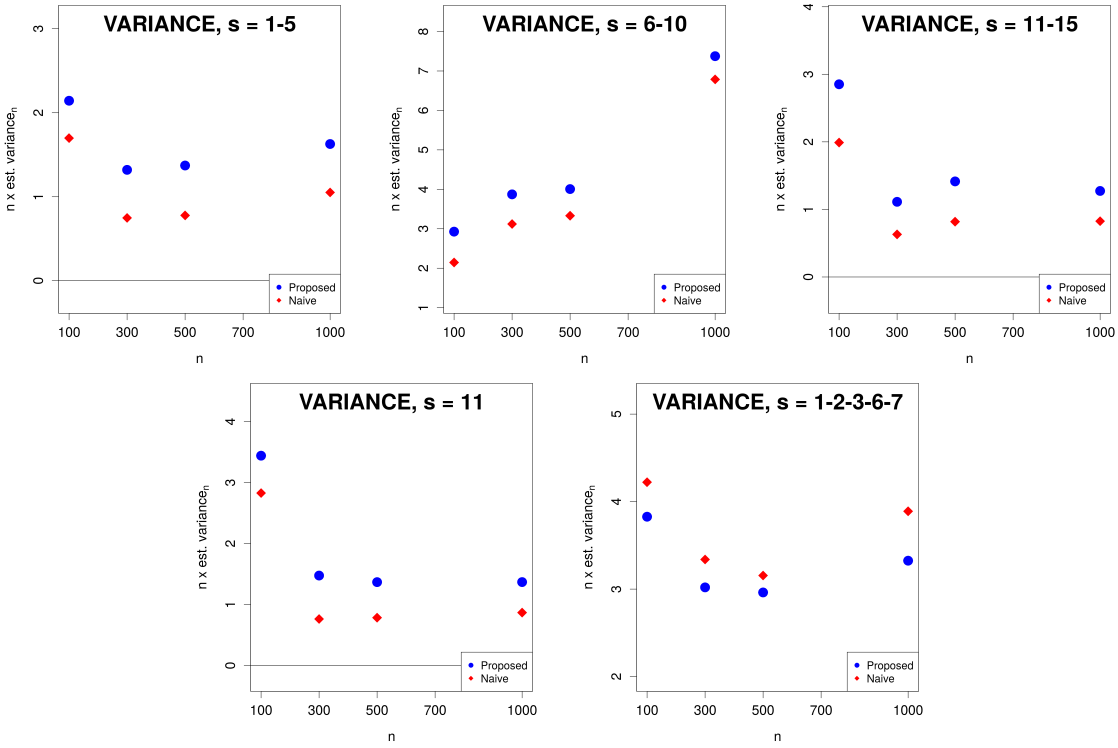

Figure A.5: Empirical variance for the proposed and naive estimators scaled by n vs n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Tables 3.1 and A.1 and the main manuscript. Diamonds denote the naive estimator, and circles denote the proposed estimator.

all of which are specific to the particular method under consideration. Comparing our results to those obtained by two other groups of investigators – Doksum and Samarov [1995] and Bi et al. [2003] – we find that our results are similar for the two most important single features, the average number of rooms and the proportion of the population designated as being of lower socioeconomic status. We estimate average number of rooms to be most important, in line with both groups of investigators. Our findings are consistent with those of Bi et al. [2003] in that distance is found to be third most important, but beyond that, our rankings differ. This is not concerning, since the other variables tend to be estimated at low importance by many methods. Importantly, we also obtain variable importance for the natural groups of variables described by Harrison and Rubinfeld [1978], in contrast to the

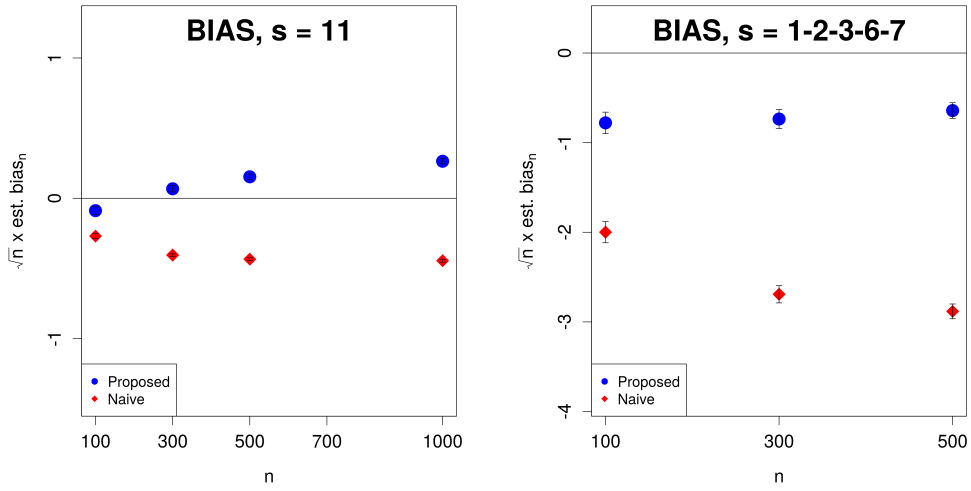

Figure A.6: Empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table A.1. Diamonds denote the naive estimator, and circles denote the proposed estimator.

method of Bi et al. [2003]. Our parameter provides a more natural interpretation than that of Doksum and Samarov [1995] – their measure provides the squared correlation between the difference $\mu_{P_0}(X) - \mu_{P_0,-s}(X)$ in means and the residual $Y - \mu_{P_0,-s}(X)$. Finally, we obtain asymptotically valid confidence intervals in addition to point estimates, which have the advantage of interpretability and generalizability to any prediction algorithm or ensemble of algorithms.

These results may be reproduced using code on the first author's GitHub page.

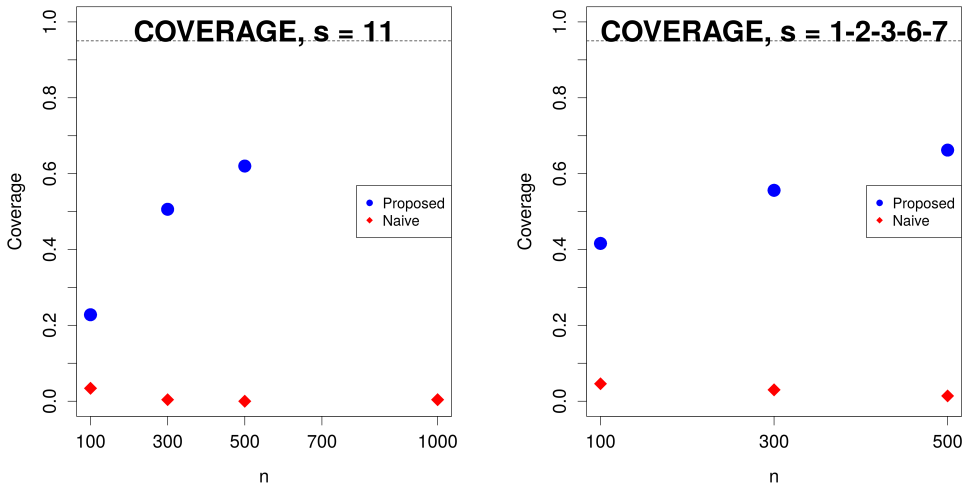

Figure A.7: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table A.1. Diamonds denote the naive estimator, and circles denote the proposed estimator.

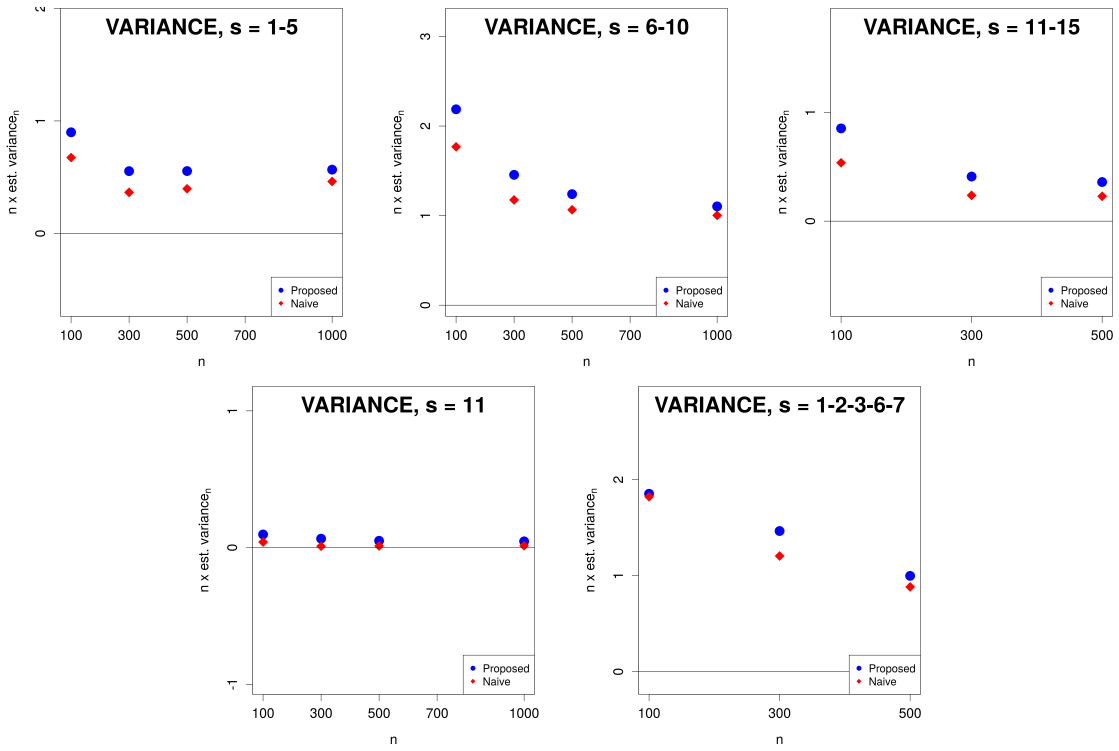

Figure A.8: Empirical variance for the proposed and naive estimators scaled by n vs n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Tables 3.1 and A.1 and the main manuscript. Diamonds denote the naive estimator, and circles denote the proposed estimator.

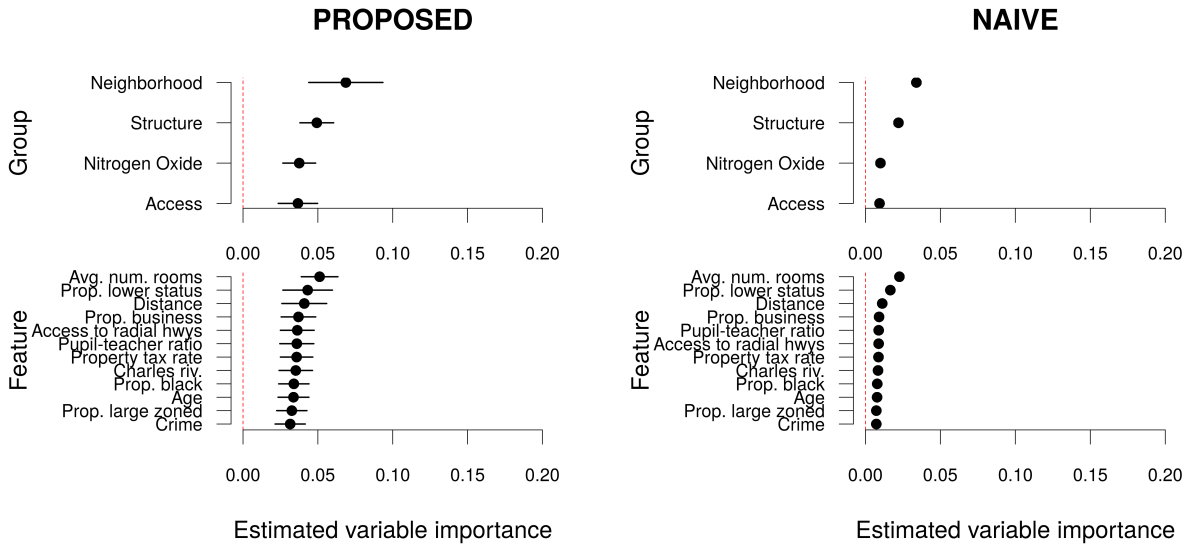

Figure A.9: Estimates from the Boston housing project study, for the proposed and naive estimators of the standardized variable importance parameter, on left and right respectively. We estimate (3.1) and (3.2) using the Super Learner with the elastic net, generalized additive models, gradient boosted trees, and random forests in its library.

Appendix B

SUPPORTING INFORMATION FOR CHAPTER 4

B.1 Proofs of theorems

As in the main manuscript, to reduce the notational complexity we introduce the following simplified notation for the predictiveness-optimizing functions:

$$f_0 \in \underset{f\in\mathcal{F}}{\operatorname{argmax}}\, V(f, P_0)$$

$$f_{0,s} \in \underset{f\in\mathcal{F}_s}{\operatorname{argmax}}\, V(f, P_0).$$

We use estimators f_n and $f_{n,s}$ of f_0 and $f_{0,s}$, respectively. For convenience, we use the notation $Pf = \int f(o)dP(o)$.

B.1.1 Proof of Theorem 2

Writing $r_n := \{V(f_n, P_n) - V(f_n, P_0)\} - \{V(f_0, P_n) - V(f_0, P_0)\}$, we first decompose

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

In view of condition (A2), the functional delta method is applicable and yields that

$$\begin{aligned} V(f_0, P_n) - V(f_0, P_0) &= \dot{V}(f_0, P_0; P_n - P_0) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + o_P(n^{-1/2}) \ , \end{aligned}$$

where $\dot{V}(f_0, P_0; h)$ is the Gâteaux derivative of the mapping $P \mapsto V(f_0, P)$ at P_0 in the direction h and δ_z is the degenerate distribution on z. Under condition (A1), we have that

$|V(f_n, P_0) - V(f_0, P_0)| = C\|f_n - f_0\|_{\mathcal{F}}^2 = o_P(n^{-1/2})$ under condition (A3). It remains to show that $r_n = o_P(n^{-1/2})$ as well. For any given $\epsilon > 0$, $h \in \mathcal{Q}$ and $f \in \mathcal{F}$, we define

$$R_0(f, \epsilon, h) := \left| \frac{V(f, P_0 + \epsilon h) - V(f, P_0)}{\epsilon} - \dot{V}(f, P_0; h) \right|.$$

Setting $\epsilon_n := n^{-1/2}$ and $h_n := n^{1/2}(P_n - P_0)$, we have that

$$\begin{aligned} n^{1/2} r_n &= \frac{[\{V(f_n, P_n) - V(f_n, P_0)\} - \{V(f_0, P_n) - V(f_0, P_0)\}]}{\epsilon_n} \\ &= \{\dot{V}(f_n, P_0; h_n) + R_0(f_n, \epsilon_n, h_n)\} - \{\dot{V}(f_0, P_0; h_n) + R_0(f_0, P_0; h_n))\} = A_n + B_n \ , \end{aligned}$$

where $A_n := \dot{V}(f_n, P_0; h_n) - \dot{V}(f_0, P_0; h_n)$ and $B_n := R_0(f_n, \epsilon_n, h_n) - R_0(f_0, \epsilon_n, h_n)$, and so, we can write that $P_0\left(n^{1/2}|r_n| > \epsilon\right) \leq P_0\left(|A_n| > \epsilon/2\right) + P_0\left(|B_n| > \epsilon/2\right)$. On one hand, since we can rewrite $A_n = \dot{V}(f_n, P_0; h_n) - \dot{V}(f_0, P_0; h_n) = n^{1/2} \int g_n(z) d(P_n - P_0)(z)$, under conditions (A4) and (A5), an application of Lemma 19.24 of van der Vaart (2000) yields that $A_n = o_P(1)$ under P_0, and so, $P_0(|A_n| > \epsilon/2) \longrightarrow 0$. On the other hand, we can write

$$\begin{aligned} P_0\left(|B_n| > \epsilon/2\right) &= P_0\left(|B_n| > \epsilon/2, \|f_n - f_0\| < \delta\right) + P_0\left(|B_n| > \epsilon/2, \|f_n - f_0\| \geq \delta\right) \\ &\leq P_0\left(\sup_{f \in \mathcal{F}: \|f - f_0\| < \delta} R_0(f, \epsilon_n, h_n) > \epsilon/4, \|f_n - f_0\| < \delta\right) + P_0\left(\|f_n - f_0\| \geq \delta\right) \\ &\leq P_0\left(\sup_{f \in \mathcal{F}: \|f - f_0\| < \delta} R_0(f, \epsilon_n, h_n) > \epsilon/4\right) + P_0\left(\|f_n - f_0\| \geq \delta\right). \end{aligned}$$

Since the first and second summands tend to zero by conditions (A2) and (A3), respectively, it follows that $P_0\left(|B_n| > \epsilon/2\right) \longrightarrow 0$. In summary, under conditions (A1)–(A5), we find that

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

under sampling from P_0, as claimed.

Using identical logic as above, if conditions (A1)–(A5) additionally hold for $f_{0,s}$, then $V(f_{n,s}, P_n)$ is an asymptotically linear estimator of $V(f_{0,s}, P_0)$ with influence function $x \mapsto$

$\dot{V}(f_{0,s}, P_0; \delta_x - P_0)$. By a simple application of the functional delta method, under conditions (A1)–(A5) we have that

$$\sqrt{n}(\hat{\psi}_{n,s} - \psi_{0,s}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \dot{V}(f_0, P_0; \delta_{X_i} - P_0) - \dot{V}(f_{0,s}, P_0; \delta_{X_i} - P_0) \right\} + o_P(1).$$

Thus $\hat{\psi}_{n,s}$ is an asymptotically linear estimator of $\psi_{0,s}$ with influence function $\phi_0 : x \mapsto \dot{V}(f_0, P_0; \delta_x - P_0) - \dot{V}(f_{0,s}, P_0; \delta_x - P)$. Since we operate in a nonparametric model space $\mathcal{M}$, the functions are the efficient influence functions of $V(f_0, P_0)$ and $\Psi_s(P_0)$ at P_0 relative to $\mathcal{M}$.

B.1.2 Proof of Theorem 3

As in the main manuscript, let $B_n := \{1, \ldots, K\}^n$ be a random vector generated independently of the data; let f_n^k and $f_{n,s}^k$ denote the estimators of f_0 and $f_{0,s}$, respectively, based on the training data $\mathcal{T}_k = \{O_i : B_{n,i} \neq k\}$; and let $P_{n,k}$ denote the empirical distribution based on the validation data with index in $\{i : B_{n,i} = k\}$.

Then we have

$$\frac{1}{K} \sum_{k=1}^{K} V(f_n^k, P_{n,k}) - V(f_0, P_0) = \frac{1}{K} \sum_{k=1}^{K} \left[\{V_4(P_{n,k}) - V_4(P_0)\} + \{V_3(f_n^k) - V_3(f_0)\} + R_n^k \right],$$

where each term on the right-hand side of the preceeding display is defined similarly as in the proof of Theorem 2, but using the cross-validated objects.

Again, the first two terms are easily controlled under condition (A1). Set $\epsilon_n = n^{-1/2}$, $h_n^k = \sqrt{n}(P_{n,k} - P_0)$, and $g_n^k : x \mapsto \dot{V}(f_n^k, P_0; h_n^k) - \dot{V}(f_0, P_0; h_n^k)$. The third term is, for each

k,

$$R_n^k \leq \left| \left\{ \frac{V(f_n^k, P_0 + \epsilon_n h_n^k) - V(f_n^k, P_0)}{\epsilon_n} - \dot{V}(f_n^k, P_0; h_n^k) \right\} - \left\{ \frac{V(f_0, P_0 + \epsilon_n h_n^k) - V(f_0, P_0)}{\epsilon_n} - \dot{V}(f_0, P_0; h_n^k) \right\} \right| + \left| \int g_n^k(x) d(P_{n,k} - P_0)(x) \right|.$$

The first two terms in this expression are second-order under condition (A2). To study the final term, we use empirical process theory and results from cross-validated targeted minimum loss-based estimation [Zheng and van der Laan, 2010]. For a class of functions $\mathcal{G}$, we define the *covering number* of $\mathcal{G}$ with respect to the $L_2(Q)$-norm as $N(\epsilon \|F\|_{Q,2}, \mathcal{G}, L_2(Q))$, where F is the envelope of $\mathcal{G}$ (i.e., $|g| \leq F$ for all $g \in \mathcal{G}$).

Suppose that $\|f_n^k - f_0\|_{\mathcal{F}} \to_P 0$. For each sample split B_n, we condition on $\mathcal{T}_k$ and consider the function class $\mathcal{G}(\mathcal{T}_k) := \{f_n^k \mapsto g(f_n^k) - g(f_0) \text{ for } g : \mathcal{F} \to \mathbb{R}\}$ of measurable functions of O. Now consider the subclass $\mathcal{G}_{\delta_n}(\mathcal{T}_k) := \{g \in \mathcal{G}(\mathcal{T}_k) : \|f_n^k - f_0\|_{\mathcal{F}} < \delta_n\}$, where δ_n is a deterministic sequence with $\delta_n \to 0$. Let $F(\delta_n, \mathcal{T}_k)$ denote the envelope of $\mathcal{G}_{\delta_n}(\mathcal{T}_k)$. If

$$E_{P_0}\left\{ \int_0^\infty \sqrt{\log \sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{G}, L_2(Q))} d\epsilon \sqrt{P_0 F(\delta_n, \mathcal{T}_k)^2} \right\} \to 0$$

as $\delta_n \to 0$, then

$$\begin{aligned} \left| \int g_n^k(x) d(P_{n,k} - P_0)(x) \right| &\leq \left| \int \sup_{g \in \mathcal{G}_{\delta_n}(\mathcal{T}_k)} \{\dot{V}(f_n^k, P_0; \delta_x - P_0) \right. \\ &\qquad \left. - \dot{V}(f_0, P_0; \delta_x - P_0)\} d(P_{n,k} - P_0)(x \mid \mathcal{T}_k) \right| \\ &= o_P(n^{-1/2}), \end{aligned}$$

by a straightforward application of Lemma 2.14.1 in van der Vaart and Wellner [1996], where we have made use of the fact that the cross-validation folds were generated independently of the data, and have used conditions (A1)–(A4) to provide the necessary convergence.

Heuristically, this says that if the class $\mathcal{G}_{\delta_n}(\mathcal{T}_k)$ of fixed functions of the validation data has a square-integrable envelope, then the remainder term is controlled. Thus, under conditions (A1)–(A4), we may write that

$$\begin{aligned}\sqrt{n}\left\{K^{-1}\sum_{k=1}^{K}V(f_n^k,P_{n,k})-V(f_0,P_0)\right\} &= K^{-1}\sum_{k=1}^{K}\big[\sqrt{n}\{V_4(P_n)-V_4(P_0)\}\\ &\quad +\sqrt{n}\{V_3(f_n)-V_3(f_0)\}+\sqrt{n}R_n\big]\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{V}(f_0,P_0;\delta_{X_i}-P_0)+o_P(1).\end{aligned}$$

Thus $K^{-1}\sum_{k=1}^{K}V(f_n^k,P_{n,k})$ is an asymptotically linear estimator of $V(f_0,P_0)$ with influence function $\phi_{V,P}:x\mapsto\dot{V}(f,P;\delta_x-P)$.

B.1.3 Proof of Theorem 4

A sufficient condition for pathwise differentiability of $P\mapsto V(f_P,P)$ is that the two *slice parameters* $P\mapsto V(f_0,P)$ and $P\mapsto V(f_P,P_0)$ are each pathwise differentiable with derivatives $\dot{V}(f_0,P_0;h)$ and $\dot{V}_{2,f}(P_0;h)$, respectively. The definition of pathwise differentiability requires an additional concept, the tangent space $T_{\mathcal{M}}(P_0)$ of the model $\mathcal{M}$ and P_0. If the two slice parameters given above are pathwise differentiable, then we may write for all $\epsilon_1,\ldots\to 0$ and $h_0,h_1,\ldots\in T_{\mathcal{M}}(P_0)$ such that $\|h_j-h_0\|_\infty\to 0$ that

$$\begin{aligned}\left\|\frac{V(f_{P_0+\epsilon_j h_j})-V(f_0,P_0)}{\epsilon_j}-\dot{V}_{2,f_0}(P_0;h_j)\right\|_\infty &\to 0 \text{ and}\\ \left\|\frac{V(f_0,P_0+\epsilon_j h_j)-V(f_0,P_0)}{\epsilon_j}-\dot{V}(f_0,P_0;h_j)\right\|_\infty &\to 0.\end{aligned}$$

Then we can write

$$\begin{aligned}
&\left\| \frac{V(f_{P_0+\epsilon_j h_j}, P_0+\epsilon_j h_j) - V(f_0, P_0)}{\epsilon_j} - \{\dot{V}(f_0, P_0; h_j) + \dot{V}_{2,f_0}(P_0; h_j)\} \right. \\
&\leq \left\| \frac{V(f_{P_0+\epsilon_j h_j}) - V(f_0, P_0)}{\epsilon_j} - \dot{V}_{2,f_0}(P_0; h_j) \right\|_\infty \\
&+ \left\| \frac{V(f_0, P_0+\epsilon_j h_j) - V(f_0, P_0)}{\epsilon_j} - \dot{V}(f_0, P_0; h_j) \right\|_\infty + \\
&\left\| \frac{r(\epsilon_j, h_j, P_0)}{\epsilon_j} \right\|_\infty,
\end{aligned}$$

where

$$r(\epsilon, h, P_0) := \{V(f_{P_0+\epsilon h}, P_0+\epsilon h) - V(f_{P_0+\epsilon h}, P_0)\} - \{V(f_0, P_0+\epsilon h) - V(f_0, P_0)\}.$$

Set

$$R_0(f, \epsilon, h) := \left| \frac{V(f, P_0+\epsilon h) - V(f, P_0)}{\epsilon} - \dot{V}_{2,f}(P_0; h) \right|.$$

Then

$$r(\epsilon_j, h_j, P_0) = \{\dot{V}_{2,f_{P_0+\epsilon_j h_j}}(P_0; h_j) - \dot{V}_{2,f_0}(P_0; h_j)\} - \{R_0(f_0, \epsilon_j, h_j) - R_0(f_{P_0+\epsilon_j h_j}, \epsilon_j, h_j)\},$$

which is a second-order difference-in-differences term. Thus, $\left\| \frac{r(\epsilon_j, h_j, P_0)}{\epsilon_j} \right\|_\infty \to 0$, and the parameter $P \mapsto V(f_P, P)$ is pathwise differentiable at each P in $\mathcal{M}$ relative to $T_{\mathcal{M}}(P)$.

B.2 Explicit description of estimation procedure for Examples 1–4

In this section, we provide the explicit form of our proposed estimator for Examples 1–4. For each example, we describe both the simple plug-in estimator and the cross-fit estimator. When we discuss cross-fitting, recall that we generate a random partition assignment vector $B_n \in \{1, \ldots, K\}^n$ by sampling uniformly from $\{1, \ldots, K\}$ with replacement, and denote

by D_k the subset of observations with index in $\{i : B_{n,i} = k\}$ for $k = 1, \ldots, K$. For each $k = 1, \ldots, K$, we denote by f_n^k and $f_{n,s}^k$ estimators of f_0 and $f_{0,s}$, respectively, constructed on the data in $\bigcup_{j \neq k} D_j$, and we denote by P_n^k the empirical distribution estimator of P_0 based on the data in D_k.

Example 1: R^2

The difference in R^2 VIM estimator is

$$\psi_{n,s} = \left[1 - \frac{\sum_{i=1}^n \{Y_i - f_n(X_i)\}^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}\right] - \left[1 - \frac{\sum_{i=1}^n \{Y_i - f_{n,s}(X_i)\}^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}\right],$$

where $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ is the marginal empirical mean of Y. In this example, $f_n = \mu_n$ and $f_{n,s} = \mu_{n,s}$, where μ_n and $\mu_{n,s}$ are estimators of μ_0 and $\mu_{0,s}$, respectively. For each $k = 1, \ldots, K$, the fold-specific difference in R^2 VIM estimator is

$$\psi_{n,s}^k = \left[1 - \frac{\frac{1}{n_k}\sum_{i \in D_k} \{Y_i - f_n^k(X_i)\}^2}{\frac{1}{n_k}\sum_{i \in D_k} (Y_i - \overline{Y}_n^k)^2}\right] - \left[1 - \frac{\frac{1}{n_k}\sum_{i \in D_k} \{Y_i - f_{n,s}^k(X_i)\}^2}{\frac{1}{n_k}\sum_{i \in D_k} (Y_i - \overline{Y}_n^k)^2}\right],$$

where $n_k := \sum_{i=1}^n I(i \in D_k)$ is the number of observations in fold k, and $\overline{Y}_n^k = \frac{1}{n_k}\sum_{i \in D_k} Y_i$ is the marginal empirical mean of Y in fold k. The cross-fit estimator is then $\psi_{n,s}^* = \frac{1}{K}\sum_{k=1}^K \psi_{n,s}^k$.

Example 2: deviance

The difference in deviance VIM estimator is

$$\begin{aligned}\psi_{n,s} = &\left[1 - \frac{\frac{1}{n}\sum_{i=1}^n \{Y_i \log f_n(X_i) + (1 - Y_i)\log(1 - f_n(X_i))\}}{\pi_n \log(\pi_n) + (1 - \pi_n)\log(1 - \pi_n)}\right] \\ &- \left[1 - \frac{\frac{1}{n}\sum_{i=1}^n \{Y_i \log f_{n,s}(X_i) + (1 - Y_i)\log(1 - f_{n,s}(X_i))\}}{\pi_n \log(\pi_n) + (1 - \pi_n)\log(1 - \pi_n)}\right],\end{aligned}$$

where $\pi_n = \frac{1}{n}\sum_{i=1}^n Y_i$ is the empirical estimator of the marginal probability $P_0(Y = 1)$. Again, in this example, $f_n = \mu_n$ and $f_{n,s} = \mu_{n,s}$. For each $k = 1, \ldots, K$, the fold-specific

difference in deviance VIM estimator is

$$\psi_{n,s}^k = \left[1 - \frac{\frac{1}{n_k}\sum_{i\in D_k}\{Y_i \log f_n^k(X_i) + (1-Y_i)\log(1-f_n^k(X_i))\}}{\pi_n^k \log(\pi_n^k) + (1-\pi_n^k)\log(1-\pi_n^k)}\right] - \left[1 - \frac{\frac{1}{n_k}\sum_{i\in D_k}\{Y_i \log f_{n,s}^k(X_i) + (1-Y_i)\log(1-f_{n,s}^k(X_i))\}}{\pi_n^k \log(\pi_n^k) + (1-\pi_n^k)\log(1-\pi_n^k)}\right],$$

where $\pi_n^k = \frac{1}{n_k}\sum_{i\in D_k} Y_i$ is the marginal estimator of $P_0(Y=1)$ in fold k. The cross-fit estimator is then $\psi_{n,s}^* = \frac{1}{K}\sum_{k=1}^K \psi_{n,s}^k$.

Example 3: classification accuracy

The difference in classification accuracy VIM estimator is $\psi_{n,s} = \frac{1}{n}\sum_{i=1}^n I\{Y_i = f_n(X_i)\} - \frac{1}{n}\sum_{i=1}^n I\{Y_i = f_{n,s}(X_i)\}$. Sensible estimators of f_0 and $f_{0,s}$ are given by

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

The fold-specific difference in classification accuracy VIM estimator is

$$\psi_{n,s}^k = \frac{1}{n_k}\sum_{i\in D_k} I\{Y_i = f_n^k(X_i)\} - \frac{1}{n_k}\sum_{i\in D_k} I\{Y_i = f_{n,s}^k(X_i)\}.$$

The cross-fit estimator is then $\psi_{n,s}^* = \frac{1}{K}\sum_{k=1}^K \psi_{n,s}^k$.

Example 4: area under the ROC curve

The difference in AUC VIM estimator is

$$\psi_{n,s} = \frac{1}{n_0 n_1}\sum_{i=1}^n\sum_{j=1}^n I\{f_n(X_i) < f_n(X_j)\}(1-Y_i)Y_j - \frac{1}{n_0 n_1}\sum_{i=1}^n\sum_{j=1}^n I\{f_{n,s}(X_i) < f_{n,s}(X_j)\}(1-Y_i)Y_j,$$

where $n_1 = \sum_{i=1}^n Y_i$ is the number of observations with corresponding $Y=1$ and $n_0 = n - n_1$. As above, in this example, we can take $f_n = \mu_n$ and $f_{n,s} = \mu_{n,s}$. The fold-specific difference

in AUC VIM estimator is

$$\psi_{n,s}^k = \frac{1}{n_{0,k}n_{1,k}} \sum_{i \in D_k} \sum_{j \in D_k} I\{f_n^k(X_i) < f_n^k(X_j)\}(1 - Y_i)Y_j \\ - \frac{1}{n_{0,k}n_{1,k}} \sum_{i \in D_k} \sum_{j \in D_k} I\{f_{n,s}^k(X_i) < f_{n,s}^k(X_j)\}(1 - Y_i)Y_j,$$

where $n_{1,k} = \sum_{i \in D_k} I(Y_i = 1)$ is the number of observations with corresponding $Y = 1$ in fold k and $n_{0,j} = n_k - n_{1,k}$. The cross-fit estimator is then $\psi_{n,s}^* = \frac{1}{K} \sum_{k=1}^K \psi_{n,s}^k$.

B.3 Additional technical details

B.3.1 Classification accuracy is maximized by the Bayes classifier

Suppose that $Y \in \{0, 1\}$ is a binary random variable. Define the Bayes classifier $b_0 : x \mapsto I\{\mu_0(x) > 1/2\}$ with $\mu_0(x) = E_{P_0}(Y \mid X = x)$. For any fixed $x \in \mathcal{X}$, we have that

$$\begin{aligned} P_0\{f(X) = Y \mid X = x\} &= P_0\{Y = 1, f(X) = 1 \mid X = x\} + P_0\{Y = 0, f(X) = 0 \mid X = x\} \\ &= f(x)P_0(Y = 1 \mid X = x) + \{1 - f(x)\}P_0(Y = 0 \mid X = x) \\ &= f(x)\mu_0(x) + \{1 - f(x)\}\{1 - \mu_0(x)\} \ , \end{aligned}$$

which allows us to write that

$$\begin{aligned} &P_0\{f(X) = Y \mid X = x\} - P_0\{b_0(X) \neq Y \mid X = x\} \\ &\quad = \mu_0(x)\{f(x) - b_0(x)\} + \{1 - \mu_0(x)\}[\{1 - f(x)\} - \{1 - b_0(x)\}] \\ &\quad = \{2\mu_0(x) - 1\}\{f(x) - b_0(x)\} \ \leq \ 0 \end{aligned}$$

by definition of b_0. It follows then that

$$\begin{aligned} P_0\{f(X) = Y\} - P_0\{b_0(X) = Y\} &= E_0[P_0\{f(X) = Y \mid X\}] - E_0[P_0\{b_0(X) = Y \mid X\}] \\ &= E_0[P_0\{f(X) = Y \mid X\} - P_0\{b_0(X) = Y \mid X\}] \leq 0 , \end{aligned}$$

so that b_0 is the maximizer of the classification accuracy $P_0\{Y = f(X)\}$.

B.3.2 AUC is maximized by the conditional mean

Suppose that $Y \in \{0, 1\}$ is a binary random variable. For a given function $f \in \mathcal{F}$ and a cutoff point c, we define the sensitivity and specificity of f with respect to P_0 as

$$\begin{aligned} 1 - F_1(P_0, f)(c) &:= P_0\{f(X) > c \mid Y = 1\} \text{ and} \\ F_0(P_0, f)(c) &:= P_0\{f(X) < c \mid Y = 0\} . \end{aligned}$$

With that in mind, consider the change of variables $w = 1 - F_0(P_0, f)(c)$; this implies that $c = F_0^{-1}(P_0, f)(1 - w)$

The AUC $AUC(P_0, f)$ can be written as

$$\begin{aligned} P_0\{f(X_1) < f(X_2) \mid Y_1 = 0, Y_2 = 1\} &= \int_0^\infty \{1 - F_1(P_0, f)(c)\} dF_0(P_0, f)(c) \\ &= -\int_1^0 \left[1 - F_1(P_0, f)\left\{F_0^{-1}(P_0, f)(1 - w)\right\}\right] dw \\ &= \int_0^1 \left[1 - F_1(P_0, f)\left\{F_0^{-1}(P_0, f)(1 - w)\right\}\right] dw. \end{aligned}$$

A further change of variables, setting $u = 1 - w$, yields that

$$AUC(P_0, f) = \int_0^1 \left[1 - F_1(P_0, f)\left\{F_0^{-1}(P_0, f)(u)\right\}\right] du.$$

But u is equivalent to 1 - the specificity; hence, we have achieved our first objective by

defining AUC as an integral with respect to specificity.

For a fixed u, the integrand $1 - F_1(P_0, f)\left\{F_0^{-1}(P_0, f)(u)\right\}$ is the sensitivity for a cutoff c and function $f \in \mathcal{F}$. But u is also equivalent to the classical type I error, and thus for a fixed u, the Neyman-Pearson Lemma states that the classification function based on the likelihood ratio $\frac{P_0(Y=1|X=x)}{P_0(Y=0|X=x)}$ is the most powerful test among all classification functions mapping $\mathcal{X}$ to $\{0, 1\}$, which utilize a function $f : \mathcal{X} \mapsto (0, 1)$. In this context power is equivalent to sensitivity, so the likelihood ratio yields a classification function that maximizes sensitivity for a fixed specificity.

However, a large likelihood ratio implies a large conditional probability; some simple algebraic manipulations yield that

$$\frac{P_0(Y = 1 \mid X = x)}{P_0(Y = 0 \mid X = x)} > c \text{ implies that}$$
$$(1 + c)P_0(Y = 1 \mid X = x) > c.$$

Therefore, basing decisions on the conditional probability is equivalent to basing decisions on the likelihood ratio. This fact, in conjunction with the Neyman-Pearson Lemma, yields that the conditional probability maximizes sensitivity for each fixed 1 - specificity and cutoff c among all functions $f \in \mathcal{F}$. Since the conditional probability maximizes the integrand of the AUC for each fixed 1 - specificity = u, then the conditional probability also maximizes AUC among all functions $f \in \mathcal{F}$.

B.3.3 Verification of conditions (A1) and (A2) for Examples 1–4

Example 1: R^2

We have that $|V(f, P_0) - V(f_0, P_0)| = E_0\{f(X) - f_0(X)\}^2/\sigma^2(P_0)$ so that $|V(f, P_0) - V(f_0, P_0)| = O(\|f - f_0\|_{\mathcal{F}}^2)$ and condition (A1) holds with $\alpha = 2$. We can verify that $\dot{V}(f, P_0; h) = -\int\{y - f(x)\}^2 h(dz)/\sigma^2(P_0)$. Since $P \mapsto E_P\{Y - f(X)\}^2$ is linear and thus Hadamard differentiable uniformly in f, condition (A2) can be shown to hold for any $\delta > 0$ provided the marginal distribution of Y under P_0 has bounded support.

Example 2: deviance

Using that $f_0 = \mu_0$ and setting $a_0 := -2/\{\log P_0(Y = 0) + \log P_0(Y = 1)\}$, a standard argument based on Taylor approximations allows to write that

$$\begin{aligned} |V(f, P_0) - V(f_0, P_0)| &= a_0 \left| E_0 \left[f_0(X) \log \left\{ \frac{f(x)}{f_0(x)} \right\} + \{1 - f_0(x)\} \log \left\{ \frac{1 - f(x)}{1 - f_0(x)} \right\} \right] \right| \\ &\leq \frac{a_0}{2} E_0 \left[\{f(x) - f_0(x)\}^2 \left\{ \frac{f_0(x)}{\xi_0(x)} + \frac{1 - f_0(x)}{1 - \xi_1(x)} \right\} \right] \end{aligned}$$

for some $\xi_0, \xi_1 : \mathcal{X} \to \mathcal{Y}$ lying pointwise between f and f_0. If $f(X), f_0(X) \in (\delta, 1 - \delta)$ almost surely under P_0, then we find that $|V(f, P_0) - V(f_0, P_0)| \leq a_0 \left(\frac{1-\delta}{\delta}\right) \|f - f_0\|_{\mathcal{F}}^2$. Thus, condition (A1) then holds with $\alpha = 2$. Since $P \mapsto E_P [Y \log f(X) + (1 - Y) \log\{1 - f(X)\}]$ is linear and thus Hadamard differentiable uniformly in f, condition (A2) can again be shown to hold for any $\delta > 0$.

Example 3: classification accuracy

By definition, and using that $f_0 = b_0$,

$$\begin{aligned} 0 &\leq P_0(Y = f_0(X)) - P_0(Y = f(X)) \\ &= [2P_0(Y = 1 \mid \mu_0(X) \geq 1/2 > \mu(X)) - 1] P_0(\mu_0(X) \geq 1/2 < \mu(X)) \\ &\quad + [2P_0(Y = 0 \mid \mu_0(X) < 1/2 \leq \mu(X)) - 1] P_0(\mu_0(X) < 1/2 \leq \mu(X)), \end{aligned}$$

where we have used the fact that any function $f : \mathcal{X} \to \{0, 1\}$ may be written as $f(x) = I\{\mu(x) > 1/2\}$ for some function $\mu : \mathcal{X} \to [0, 1]$. But

$$P_0(Y = 1 \mid \mu_0(X) \geq 1/2 > \mu(X)) - \frac{1}{2} = E_0\{\mu_0(X) - \frac{1}{2} \mid \mu_0(X) \geq 1/2 > \mu(X)\},$$

which implies that $|P_0(Y = 1 \mid \mu_0(X) \geq 1/2 > \mu(X)) - \frac{1}{2}| \leq \|\mu - \mu_0\|_\infty$. The same holds for $P_0(Y = 0 \mid \mu_0(X) < 1/2 \leq \mu(X))$. Finally, provided that the classification margin condition

$$P_0(\mu_0(X) \geq 1/2 < \mu(X)) = P_0(|\mu_0(X) - 0.5| < |\mu(X) - \mu_0(X)|) \leq \kappa \|\mu - \mu_0\|_\infty$$

holds for all $\|\mu - \mu_0\|_\infty$ small, (A1) holds. Condition (A2) can easily be shown to hold for any $\delta > 0$.

Example 4: AUC

For convenience, we consider $Y \in \{-1, +1\}$ and set $Z = (Y_1 - Y_2)/2$. We first note that in this setting, we can rewrite $AUC(f, P_0)$ as $1 - [2p(1-p)]^{-1} P_0([2I\{f(X_1) < f(X_2)\} - 1]Z < 0)$, where $p = P_0(Y = 1)$.

Then we can write

$$\begin{aligned} 0 &\leq AUC(f_0, P_0) - AUC(f, P_0) \\ &= P_0([2I\{f(X_1) < f(X_2)\} - 1]Z < 0) - P_0([2I\{f_0(X_1) < f_0(X_2)\} - 1]Z < 0) \\ &= E_0(|f_0(X_1) - f_0(X_2)| I[\{f(X_1) - f(X_2)\}\{f_0(X_1) - f_0(X_2)\} < 0]). \end{aligned}$$

This form no longer relies on $Y \in \{-1, +1\}$.

We first control the indicator function in the above display. We have that the set

$$\begin{aligned} \{[f(X_1) - f(X_2)][f_0(X_1) - f_0(X_2)] < 0\} &= \{(A + Z)Z < 0\} \\ &= \{(Z + \frac{1}{2}A)^2 - \frac{1}{4}A^2 < 0\} \\ &= \{|Z| < |A|, ZA < 0\} \\ &= \{|Z| < |A|\} \\ &= \{|f_0(X_1) - f_0(X_2)| < t(X_1) + t(X_2)\}, \end{aligned}$$

where $t(x) = |f(x) - f_0(x)|$ and where $A := [f(X_1) - f_0(X_1)] + [f_0(X_2) - f(X_2)]$ and $Z := [f_0(X_1) - f_0(X_2)]$.

Using this result, we have that

$$\begin{aligned}
&E_0(|f_0(X_1) - f_0(X_2)|I[\{f(X_1) - f(X_2)\}\{f_0(X_1) - f_0(X_2)\} < 0]) \\
&\leq E_0[|f_0(X_1) - f_0(X_2)|I\{|f_0(X_1) - f_0(X_2)| < t(X_1) + t(X_2)\}] \\
&\leq E_0[|f_0(X_1) - f_0(X_2)|I\{|f_0(X_1) - f_0(X_2)| < 2\|t\|_\infty\}] \\
&\leq 2\|t\|_\infty P_0\{|f_0(X_1) - f_0(X_2)| < 2\|t\|_\infty\} \\
&\leq 2C_0\|t\|_\infty^2,
\end{aligned}$$

provided that $P_0\{|f_0(X_1) - f_0(X_2)| < t\} \leq C_0 t$ for all t small. Thus, condition (A1) holds. Condition (A2) can easily be shown to hold for any $\delta > 0$.

B.3.4 Sufficient conditions for standardized V-measures

Recall that a standardized V-measure has the form $V(f, P) = a + V_1(f, P)/V_2(P)$ with

$$V_1(f, P) := E_P\left\{G\left((Y_1, f(X_1)), \ldots, (Y_m, f(X_m))\right)\right\}$$

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. Using the symmetry of G, the Gâteaux derivative of $P \mapsto V_1(f, P)$ at P_0 in the direction h is given by

$$\dot{V}(f, P_0; h) = m \int \cdots \int G((y_1, f(x_1)), \ldots, (y_m, f(x_m)))dh(z_1)dP_0(z_2)\ldots dP_0(z_m).$$

Consider a sequence $\epsilon_j \to 0$ and $h, h_1, \cdots \in \mathcal{R}$ such that $\|h_j - h\|_\infty \to 0$. Then we can write

$$\frac{V(f, P_0 + \epsilon_j h_j) - V(f, P_0)}{\epsilon_j} - \dot{V}(f, P_0; h_j) + \dot{V}(f, P_0; h_j) = Rem_f^{\epsilon_j}(f, P_0; h_j) + \dot{V}(f, P_0; h_j).$$

Focusing on the remainder term, we can write

$$Rem_f^{\epsilon_j}(f, P_0; h_j) = \int \cdots \int G((y_1, f(x_1)), \ldots, (y_m, f(x_m))) \left[\frac{\prod_{i=1}^m d(P_0 + \epsilon_j h_j)(z_i) - \prod_{i=1}^m dP_0(z_i)}{\epsilon_j} \right]$$
$$= \int \cdots \int G((y_1, f(x_1)), \ldots, (y_m, f(x_m)))$$
$$\times \left(\sum_{i=1}^m \left[\prod_{\ell=1}^{i-1} \{d(P_0 + \epsilon_j h_j)(z_\ell)\} \left\{ \frac{d(P_0 + \epsilon_j h_j)(z_j) - dP_0(z_j)}{\epsilon_j} \right\} \right] \prod_{k=i+1}^m dP_0(z_k) \right)$$
$$= \int \cdots \int G((y_1, f(x_1)), \ldots, (y_m, f(x_m))) \left(\sum_{i=1}^m \left[\prod_{\ell=1}^{i-1} \{d(P_0 + \epsilon_j h_j)(z_\ell)\} \{dh_j(z_\ell)\} \right] \prod_{k=i+1}^m dP_0(z_k) \right),$$

using the telescoping identity of van der Laan [1991]. Using again the symmetry of G, we may rewrite this as

$$\begin{aligned} Rem_f^{\epsilon_j}(f, P_0; h_j) = \sum_{i=1}^m \int \cdots \int & G((y_1, f(x_1)), \ldots, (y_m, f(x_m))) dh_j(z_1) d(P_0 + \epsilon_j h_j)(z_2) \ldots \\ & d(P_0 + \epsilon_j h_j)(z_i) dP_0(z_{i+1}) \ldots dP_0(z_m) \\ = \epsilon_j \binom{m}{2} \int \cdots \int & G((y_1, f(x_1)), \ldots, (y_m, f(x_m))) dh_j(z_1) dh_j(z_2) \\ & dP_0(z_3) \ldots dP_0(z_m) \\ + \sum_{k=2}^{m-1} \Bigg[\epsilon_j^k \binom{m}{k+1} \int \cdots \int & G((y_1, f(x_1)), \ldots, (y_m, f(x_m))) dh_j(z_1) \ldots dh_j(z_{k+1}) \\ & dP_0(z_{k+2}) \ldots dP_0(z_m) \Bigg]. \end{aligned}$$

Thus, since $\|h_j - h\|_\infty \to 0$ and under the assumption that G is bounded almost surely under P_0 in the sense that for some $M < \infty$, $P_0(|G((Y_1, f(X_1)), \ldots, (Y_m, f(X_m)))| \leq M) = 1$, we have that

$$\lim_{j \to \infty} Rem_f^{\epsilon_j}(f, P_0; h_j) = 0.$$

Thus, condition (A2) is satisfied if G is bounded in the sense mentioned above.

Table B.1: Approximate values of $\psi_{0,s}$ for the numerical experiments in Section 3.3. The values are different for the different VIMs.

Both features are important		Only X_1 is important		
X_1	X_2	X_1	X_2	Importance measure
0.143	0.300	0.299	0	Deviance
0.051	0.116	0.181	0	Accuracy
0.040	0.106	0.356	0	AUC

B.4 Additional numerical experiments

B.4.1 Using cross-fitting to estimate the VIM

In Section 3.3 of the main manuscript, we estimated each VIM using cross-fitting, using an inner layer of cross-validation to estimate the predictiveness maximizing functions. In this section, we describe the behavior of our proposed estimation procedure under different scenarios. We use in all cases the same data-generating mechanism described in Section 3.3 of the main manuscript. We provide the results of each experiment in the next three subsections, and provide concluding remarks in Section B.4.1. The true VIM values of Scenarios 1 and 2 (corresponding to X_2 having non-null and null importance, respectively) are provided in Table 4.1.

The results based on the deviance VIM are presented in Figures B.1 and B.2. These results are similar to the results presented in the main manuscript, so we do not discuss them further here.

Simple estimators with no cross-fitting

Here, we use only generalized linear models and the sample mean as candidate estimators and do not use cross-fitting for the VIM estimator. Note that the true conditional mean model implied by the data-generating mechanism is in this case a generalized linear model, so our estimator is correctly specified and has a parametric rate of convergence to the true conditional mean. We use the hypothesis test outlined in Algorithm 4 (main manuscript). We

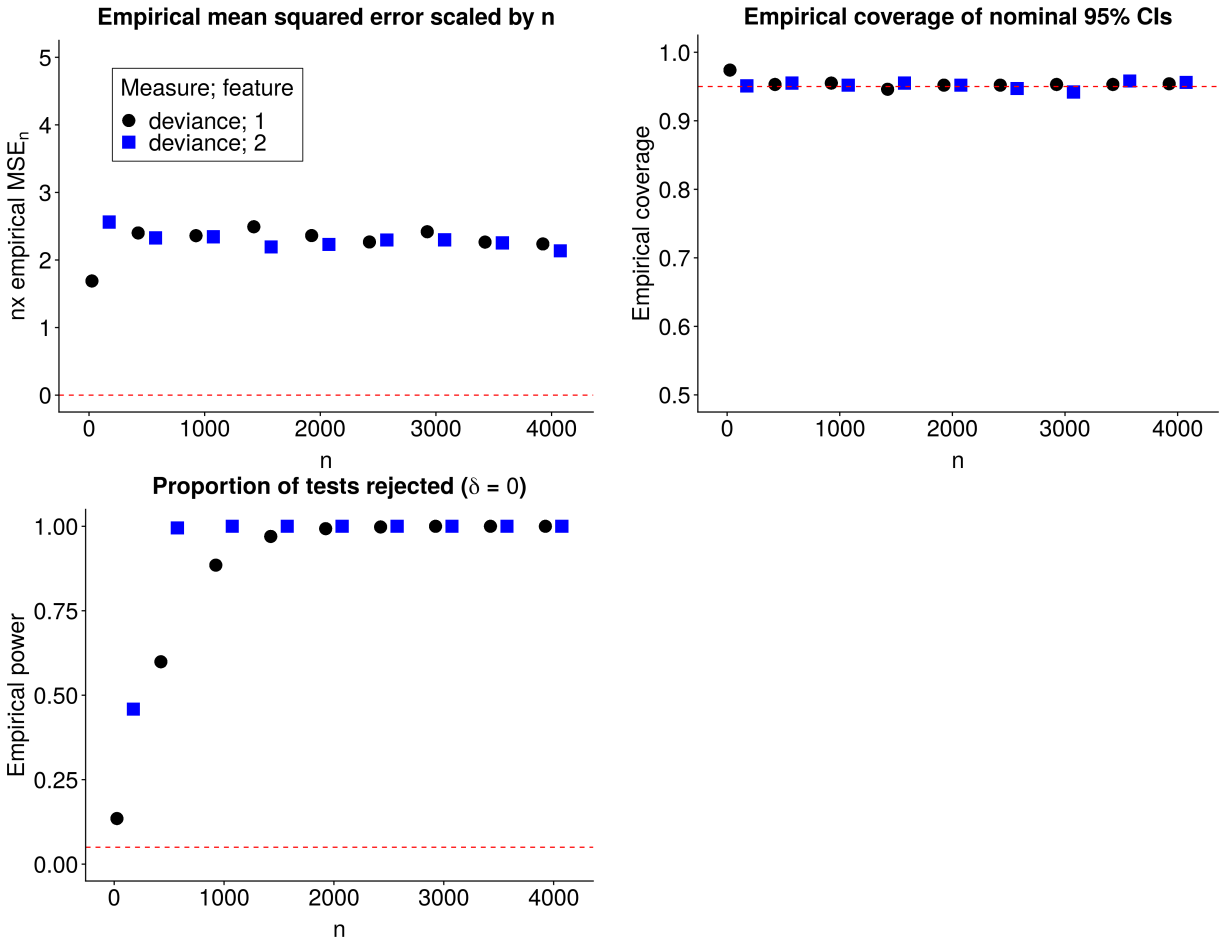

Figure B.1: Performance of plug-in estimators for estimating importance via the deviance. Clockwise from top left: empirical mean squared error for the proposed plug-in estimator scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; power of the split-sample hypothesis testing procedure vs n when $\delta = 0.05$; and power of the split-sample hypothesis testing procedure vs n when $\delta = 0$. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively.

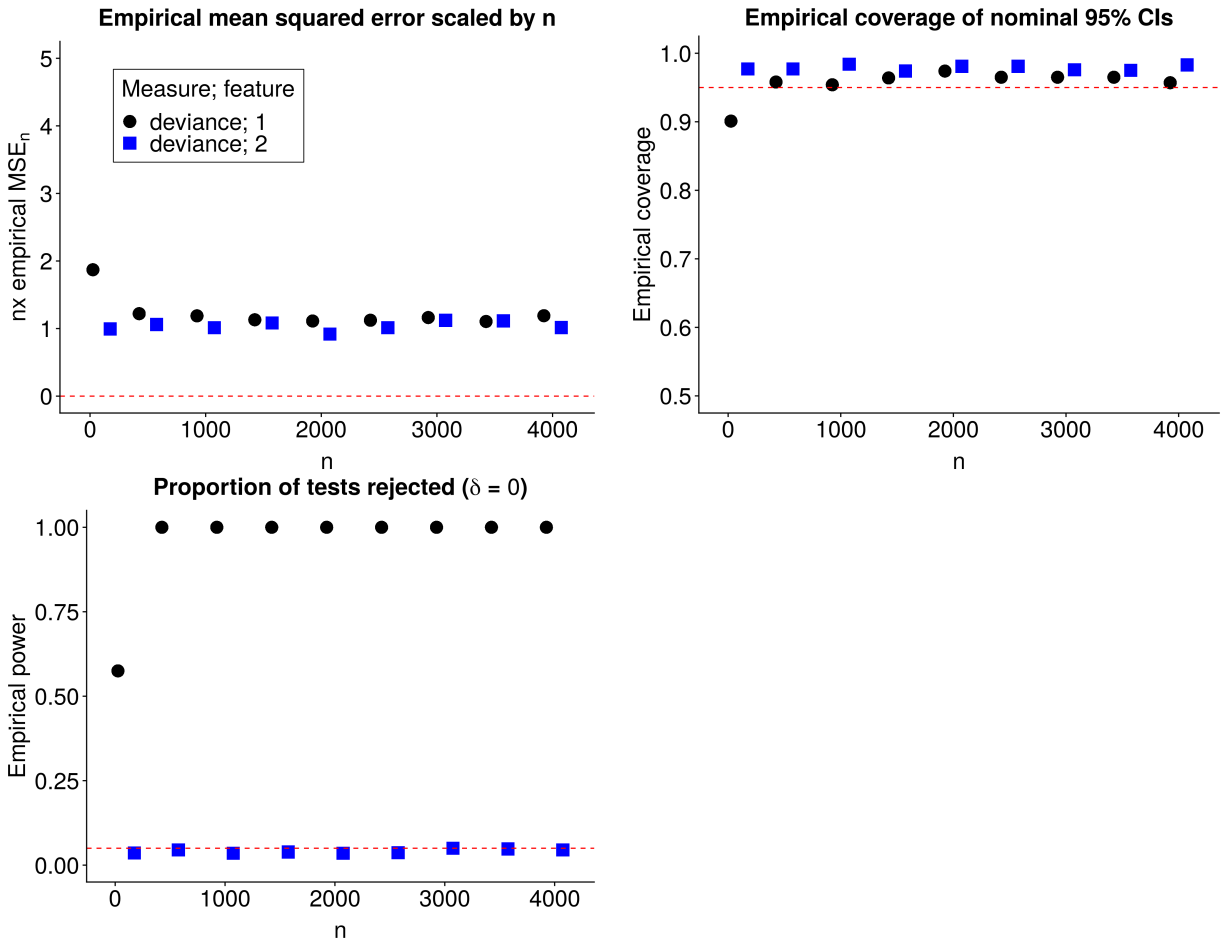

Figure B.2: Performance of plug-in estimators for estimating importance via the deviance. Clockwise from top left: empirical mean squared error for the proposed plug-in estimator scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; power of the split-sample hypothesis testing procedure vs n when $\delta = 0.05$; and power of the split-sample hypothesis testing procedure vs n when $\delta = 0$. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively. We operate here under the null hypothesis for X_2: $\psi_{0,s} = 0$.

display the results of an experiment where both features are truly important in Figure B.3. The results match with our expectations based on having a correctly specified estimator: the scaled MSE of the proposed estimator is tightly controlled; approximate coverage of nominal 95% confidence intervals quickly approaches the nominal level; and the test has high power (even for X_1, which has importance closer to the null than X_2).

In Figure B.4, we display the results of an experiment with the same estimation procedure, but where the true importance of X_2 is zero. Here, we see that there is some residual bias for the estimator of the importance of X_1, the non-null feature, with respect to the difference in AUCs. However, the scaled bias for the other estimators approaches zero; coverage approaches the nominal level in all cases; and power is high for all non-null features and type I error is controlled at the nominal level for the null features.

Simple estimators with cross-fitting

In this experiment, we use the same conditional mean estimators as in the previous section. We now use cross-fitting for the VIM estimator. We display the results of an experiment where both features are truly important in Figure B.5. The results match with our expectations based on having a correctly specified estimator: the scaled MSE of the proposed estimator is tightly controlled, and approximate coverage of nominal 95% confidence intervals quickly approaches the nominal level. The power of the test in this experiment (lower-right panel of Figure B.5) is lower than the power in the experiment of Section B.4.1. This may be explained in part by the fact that the approximate cross-validated test uses a smaller *effective sample size* than the split-sample hypothesis test. We hypothesize that if we used nested cross-validation to do the hypothesis test, that it would have higher power; we refrain from doing so because it adds computation time.

We display the results of a similar experiment where X_2 has zero importance in Figure B.6. In contrast to Figure B.4, in this scenario there is no residual bias for the non-null feature's importance relative to the difference in AUCs. The results are otherwise similar to those in Figure B.5.

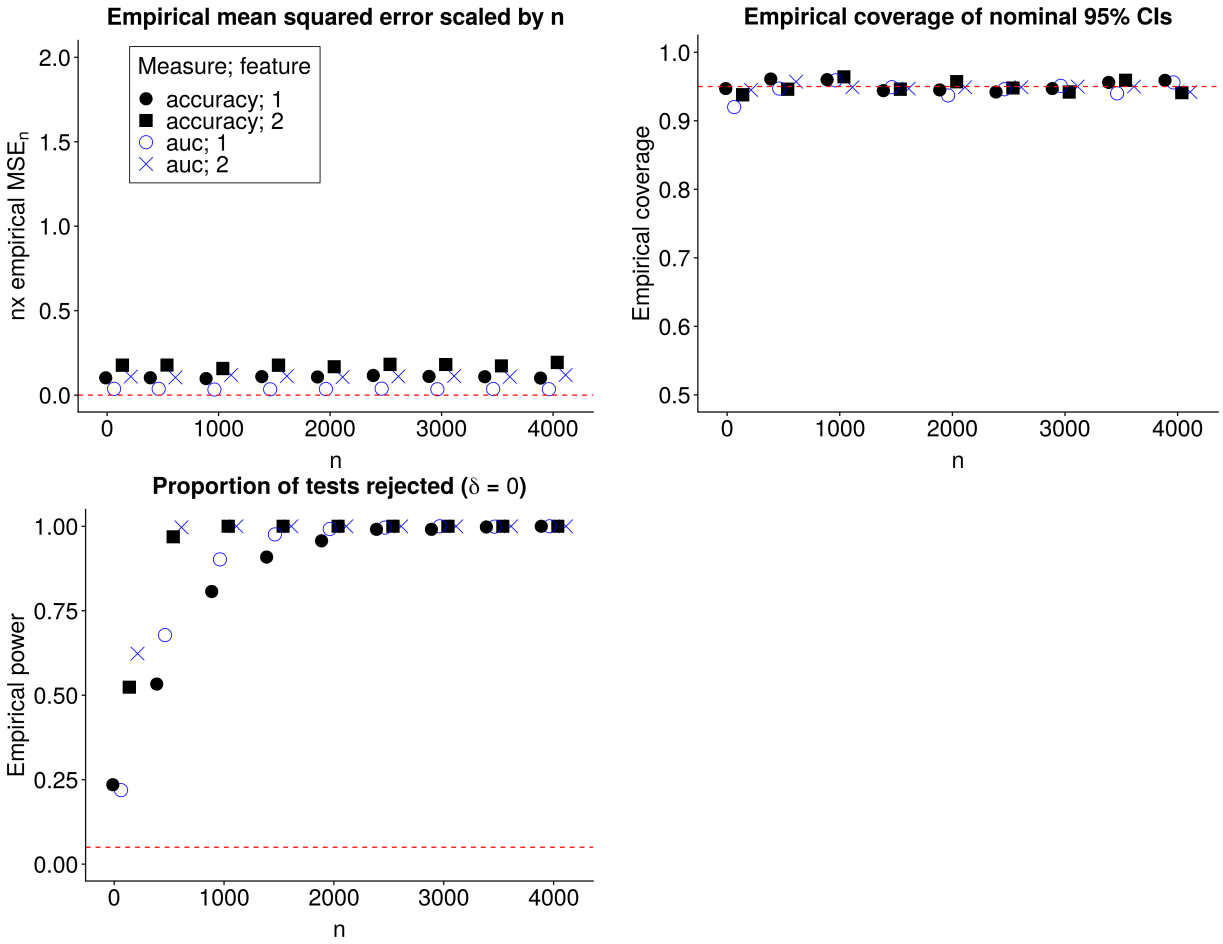

Figure B.3: Performance of plug-in estimators for estimating importance using only simple conditional mean estimators, without cross-fitting. Clockwise from top left: empirical MSE for the plug-in estimators scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; and power of the split-sample hypothesis testing procedure vs n. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively.

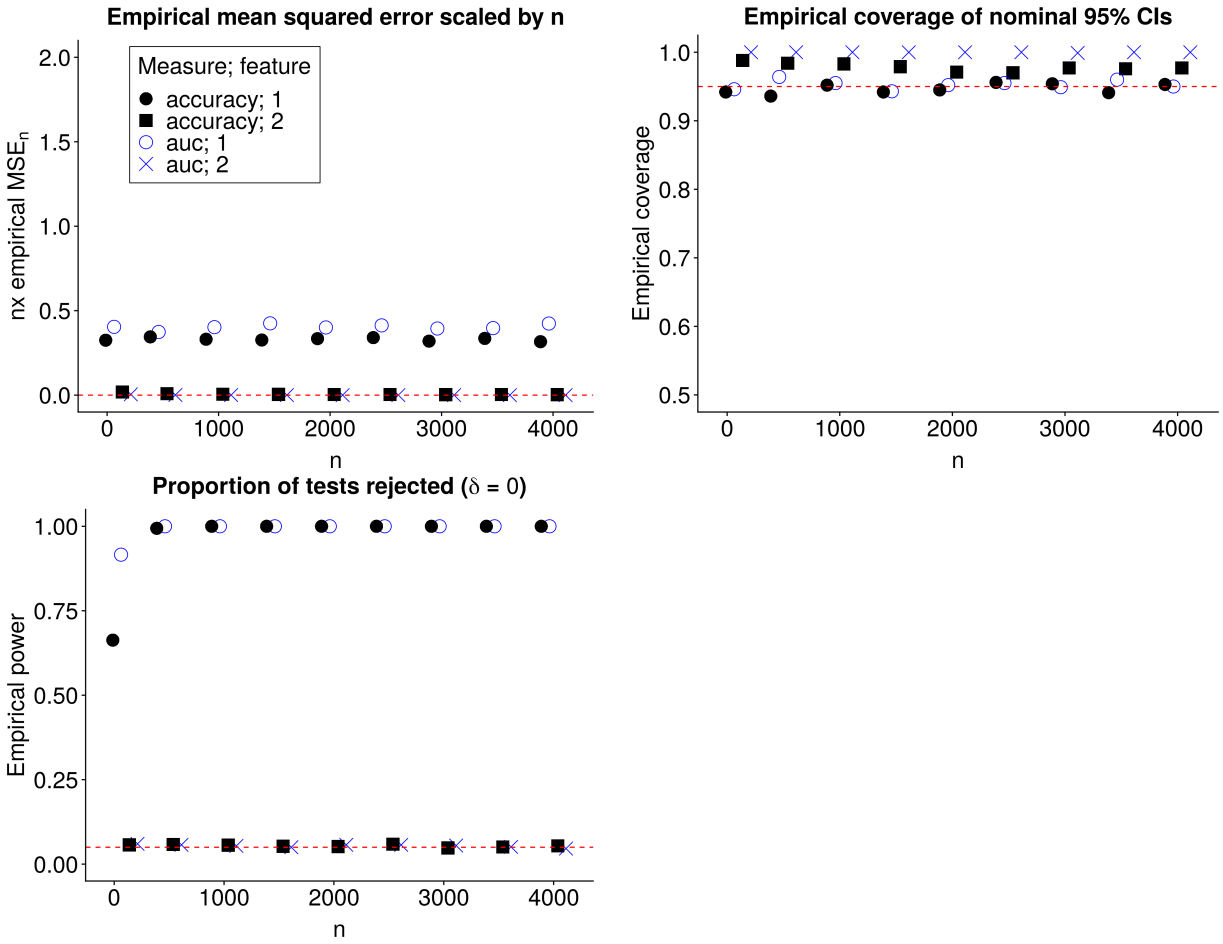

Figure B.4: Performance of plug-in estimators for estimating importance using only simple conditional mean estimators, without cross-fitting. Clockwise from top left: empirical MSE for the plug-in estimators scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; and power of the split-sample hypothesis testing procedure vs n. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively. In this experiment, the importance of X_2 is zero.

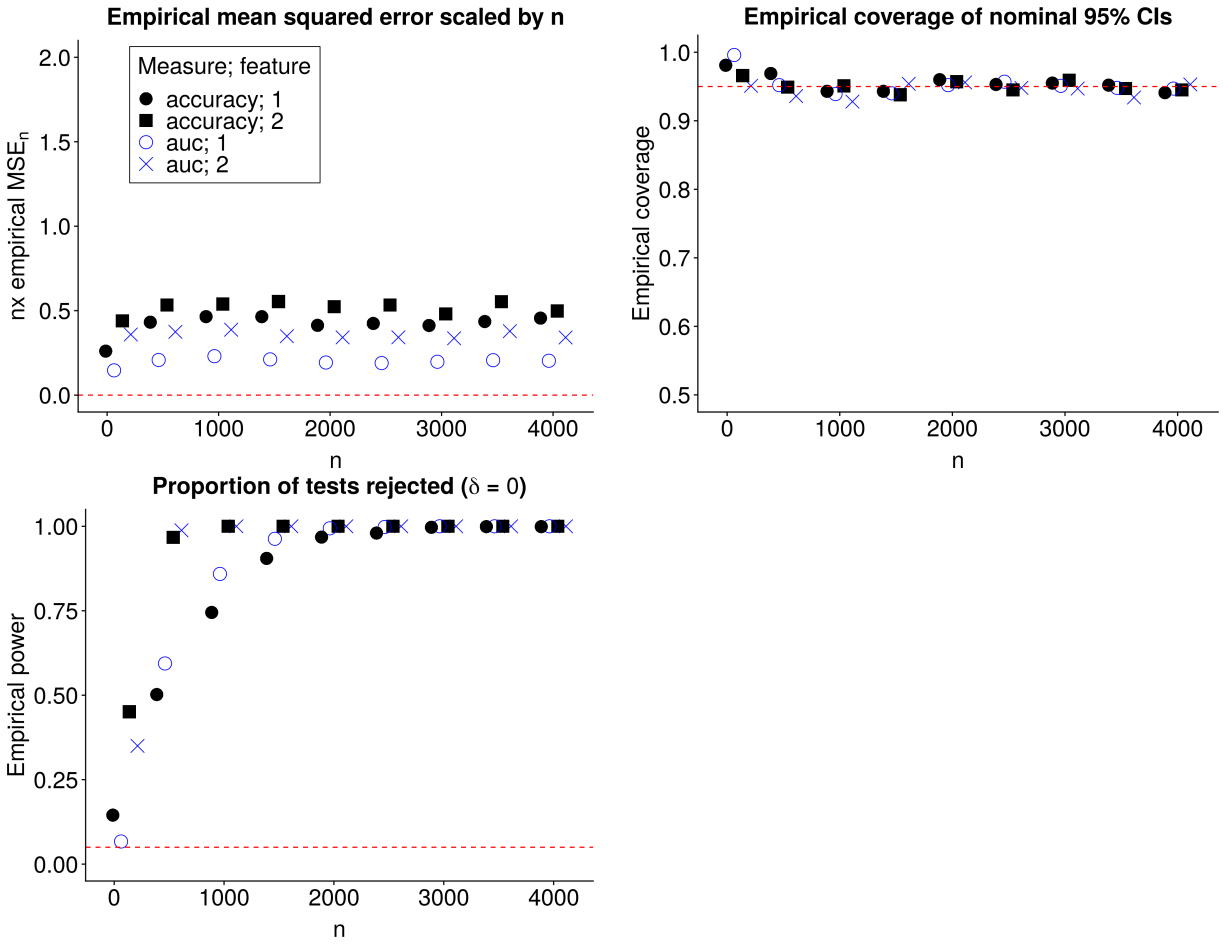

Figure B.5: Performance of plug-in estimators for estimating importance using only simple conditional mean estimators, with cross-fitting. Clockwise from top left: empirical MSE for the plug-in estimators scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; and power of the split-sample hypothesis testing procedure vs n. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively.

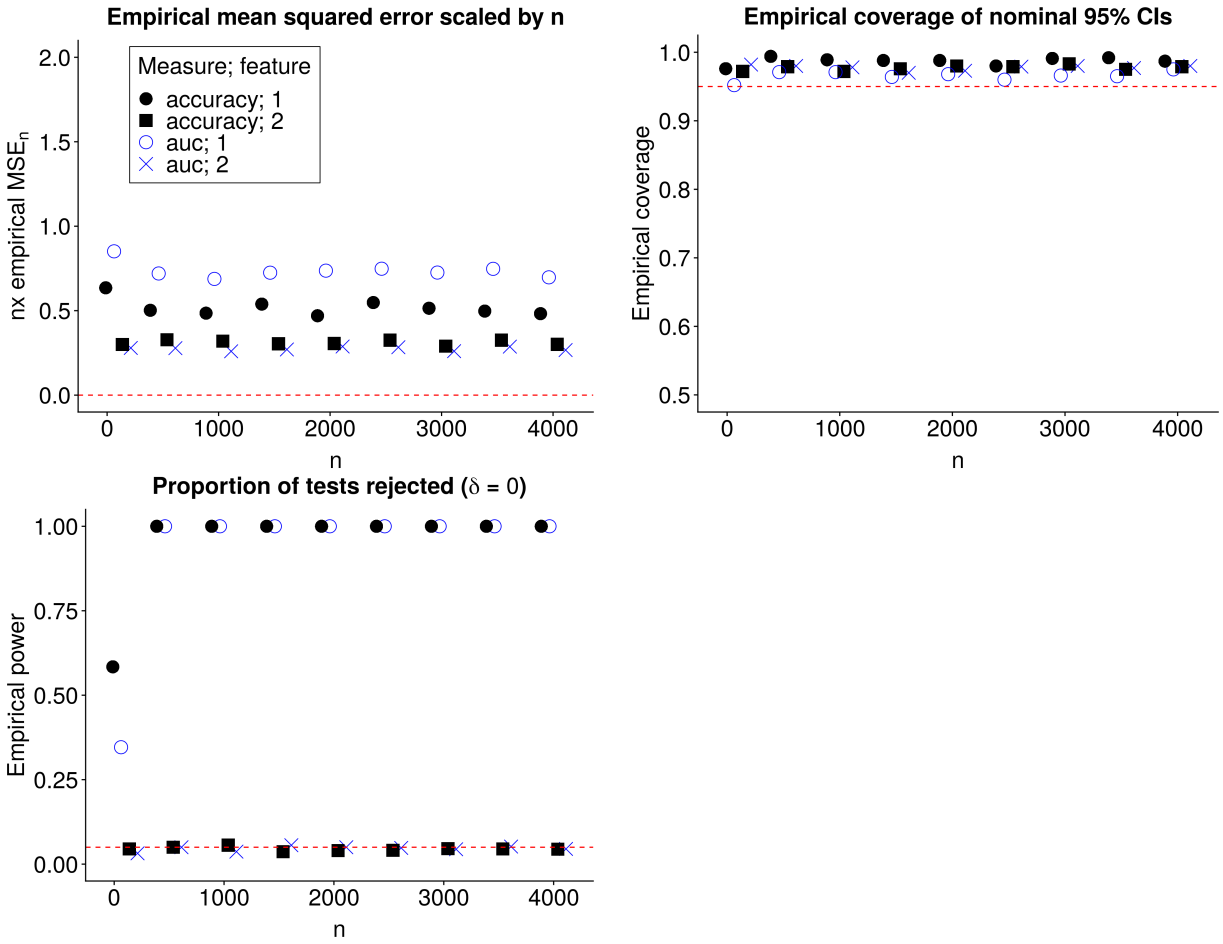

Figure B.6: Performance of plug-in estimators for estimating importance using only simple conditional mean estimators, with cross-fitting. Clockwise from top left: empirical MSE for the plug-in estimators scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; and power of the split-sample hypothesis testing procedure vs n. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively. In this experiment, the true importance of X_2 is zero.

Flexible estimators with no cross-fitting

In this experiment, we use the same conditional mean estimators as in the main manuscript, but do not use cross-fitting to estimate the VIMs. We display the results of this experiment in Figure B.7. Here, there is some residual bias for many of the estimators, and coverage is below the nominal level.

We present the results of an experiment where X_2 has zero importance in Figure B.8. Here, we see much more residual bias and correspondingly poorer coverage of confidence intervals.

Conclusing remarks on the use of cross-fitting

When using simple (and correctly specified) estimators of the conditional mean functions, using cross-fitting appears to only affect V-measures of degree two under the null. There is not much difference between Figures B.5 and B.3, while the only difference between Figures B.6 and B.4 is that the scaled bias of the AUC-VIM estimator approaches zero. This bias may come from several sources. First, V-measures of degree one are linear, while V-measures of degree two are only asymptotically linear. Second, the true AUC of an estimator with only the null feature is 0.5, which is on the boundary of the parameter space. This behavior may affect the higher-order terms in the asymptotic expansion. Further work is necessary to fully describe this expansion.

However, it is rarely the case that we know that a simple estimator is correctly specified. When using flexible estimators, it is crucial to use cross-fitting to estimate variable importance, as seen by comparing Figures B.7 and B.8 to Figures 4.1 and 4.2 in the main manuscript. The residual bias in Figures B.7 and B.8 may be due to the Donsker class conditions not being satisfied for some of our candidate learners (e.g., random forests or boosted trees). These results suggest that cross-fitting does weaken the conditions necessary for our plug-in VIM estimators to be asymptotically linear estimators of the true VIMs.

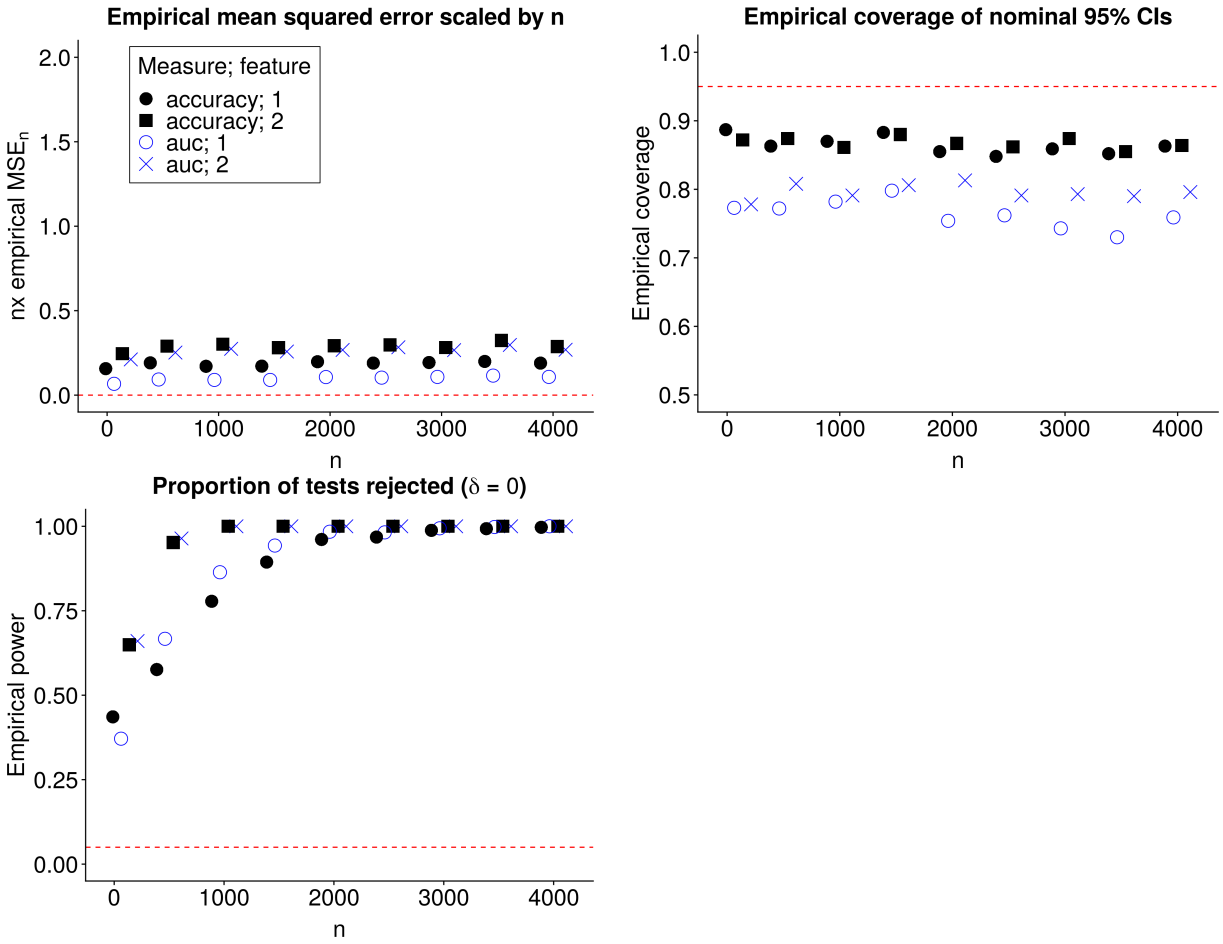

Figure B.7: Performance of plug-in estimators for estimating importance using flexible conditional mean estimators, without cross-fitting. Clockwise from top left: empirical MSE for the plug-in estimators scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; and power of the split-sample hypothesis testing procedure vs n. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively.

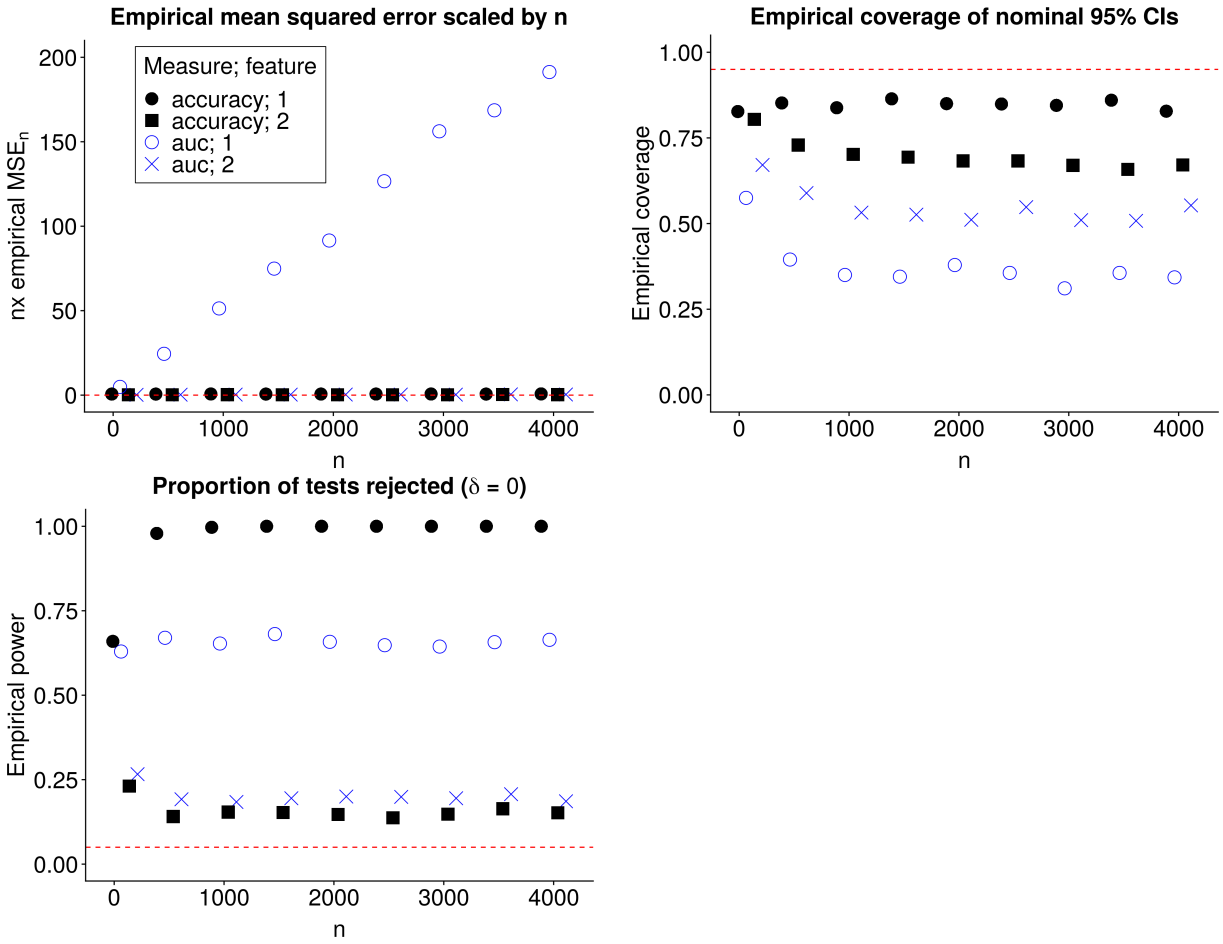

Figure B.8: Performance of plug-in estimators for estimating importance using flexible conditional mean estimators, without cross-fitting. Clockwise from top left: empirical MSE for the plug-in estimators scaled by n vs n for $j = 1$ and 2; approximate coverage of nominal 95% confidence intervals for the true importance vs n; and power of the split-sample hypothesis testing procedure vs n. Circles and squares denote that we are estimating the importance via the accuracy and AUC, respectively. In this experiment, X_2 has zero importance.

Table B.2: Library of canididate learners for the Super Learner estimator in the HIV-1 antibody resistance data with descriptions.

Function name	Description
SL.mean	intercept only regression
SL.xgboost1	boosted regression trees with maximum depth of 1
SL.xgboost2	boosted regression trees with maximum depth of 2
SL.xgboost4	boosted regression trees with maximum depth of 4
SL.xgboost6	boosted regression trees with maximum depth of 6
SL.xgboost8	boosted regression trees with maximum depth of 8
SL.ranger.small	random forest with mtry equal to one-half times square root of number of predictors
SL.ranger.reg	random forest with mtry equal to square root of number of predictors
SL.ranger.large	random forest with mtry equal to two times square root of number of predictors
SL.glmnet.mycv	GLMNET with lambda selected by CV and alpha equal to 0
SL.glmnet.25	GLMNET with lambda selected by 5-fold CV and alpha equal to 0.25
SL.glmnet.50	GLMNET with lambda selected by 5-fold CV and alpha equal to 0.5
SL.glmnet.75	GLMNET with lambda selected by 5-fold CV and alpha equal to 0.75
SL.glmnet.1	GLMNET with lambda selected by 5-fold CV and alpha equal to 1

B.5 Additional details for the study of an antibody against HIV-1

B.5.1 Library of candidate learning algorithms

In this section, we describe the library of candidate learning algorithms for our analysis replicating the results of Magaret et al. [2019]. We used a wide arrange of flexible machine learning-based algorithms in the hopes that this large library of algorithms would yield a cross-validated algorithm with good predictive performance. The machine learning techniques were: the lasso with logit link function (implemented in the `glmnet` R package), random forests (implemented in the `ranger` R package), and gradient boosted decision trees (implemented in the `xgboost` R package), each with varying tuning parameters. In Table B.2, we provide a description of each candidate learning algorithm in our library. Our final estimator is the convex combination of these algorithms chosen to minimize the ten-fold cross-validated negative log likelihood. In all cases, we adjusted for geographic region as a potential confounding variable.

B.5.2 Super Learner performance

We now describe the empirical performance of the Super Learner. In Table B.3, we show the coefficients in the final Super Learner of each of the candidate learners described in Table

Table B.3: Table of Super Learner weights for each candidate learner and cross-validation fold. We remove "SL" from each learner name and abbreviate "xgboost" with "xgb". The abbreviations "S" and "L" denote a small and large number of candidate variables tried at each split in the random forest, respectively. The abbreviation "mn" denotes the mean.

mn	xgb1	xgb2	xgb4	xgb6	xgb8	ranger.S	ranger.reg	ranger.L	glmnet.0	glmnet.25	glmnet.50	glmnet.75	glmnet.1
0	0	0.00	0.00	0	0	0	0	0.54	0.12	0.34	0.00	0.00	0.00
0	0	0.00	0.00	0	0	0	0	0.38	0.00	0.49	0.13	0.00	0.00
0	0	0.00	0.00	0	0	0	0	0.75	0.00	0.00	0.25	0.00	0.00
0	0	0.00	0.07	0	0	0	0	0.09	0.00	0.60	0.00	0.00	0.24
0	0	0.00	0.00	0	0	0	0	0.63	0.00	0.00	0.00	0.32	0.04
0	0	0.00	0.00	0	0	0	0	0.98	0.00	0.00	0.00	0.02	0.00
0	0	0.00	0.00	0	0	0	0	0.59	0.00	0.41	0.00	0.00	0.00
0	0	0.00	0.00	0	0	0	0	0.74	0.00	0.26	0.00	0.00	0.00
0	0	0.07	0.00	0	0	0	0	0.37	0.56	0.00	0.00	0.00	0.00
0	0	0.00	0.00	0	0	0	0	0.52	0.17	0.31	0.00	0.00	0.00

B.2. The rows of this table are each of the ten cross-validation folds, while the columns are the individual learners. Here, we see that boosted trees with a maximum depth of four or six were infrequently chosen as part of the Super Learner; random forests with a large number of variables considered at each split were often chosen as part of the Super Learner; and glmnet with varying values of α were often chosen by the Super Learner.

In Figure B.9, we display the cross-validated AUC and 95% confidence intervals for each of the candidate learning algorithms in the Super Learner, along with the Super Learner algorithm and the classical cross-validated selector (the "discrete Super Learner"). Similar to Magaret et al. [2019], we see that of the individual algorithms, random forests have the best performance in these data, followed by the lasso and boosted trees. Additionally, we estimate the cross-validated AUC of the overall Super Learner to be 0.90, with a 95% confidence interval of [0.84, 0.98]. Magaret et al. [2019] performed the analysis separately on two independent splits of these data, and obtained cross-validated AUCs of 0.86 [0.81, 0.92] and 0.87 [0.81, 0.93] on the two datasets.

In Figure B.10, we display cross-validated ROC curves for the Super Learner, discrete Super Learner, and the top-performing individual algorithm. These ROC curves are similar to those presented in Magaret et al. [2019] — in both analyses, we see a large cross-validated true positive rate for each chosen cross-validated false positive rate. Thus, we can conclude that our predictor allows reasonably accurate classification between right-censored IC_{50} viruses

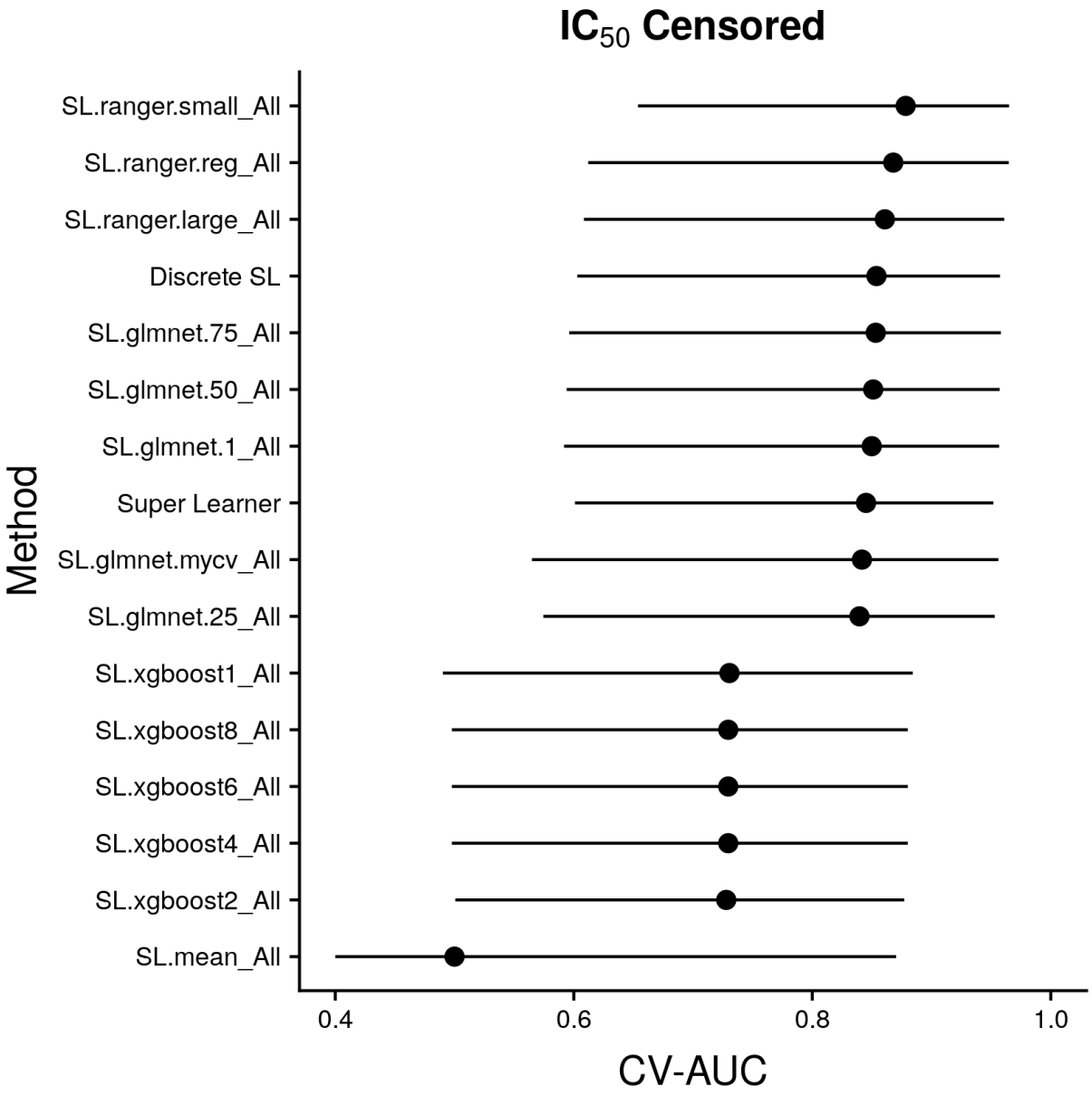

Figure B.9: Point estimates of cross-validated AUC in predicting HIV-1 antibody resistance, with 95% confidence intervals, for each candidate learning algorithm in the Super Learner.

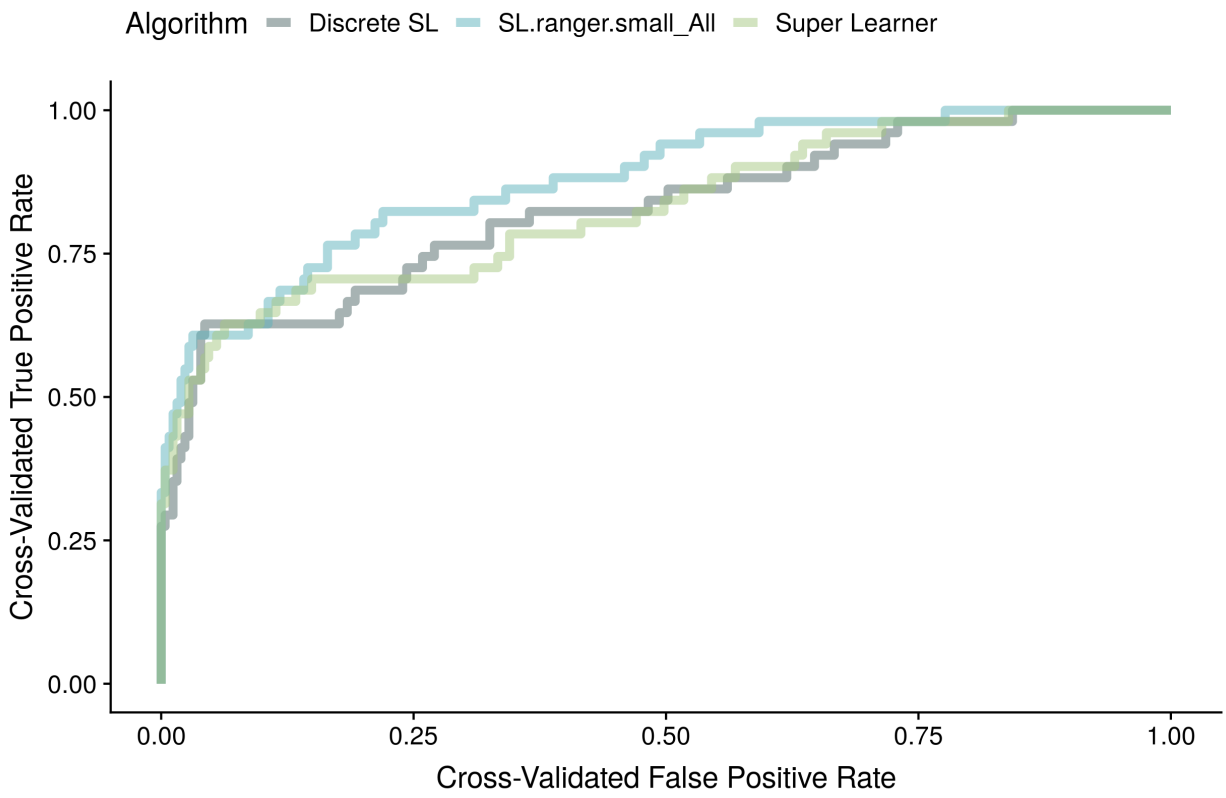

Figure B.10: Cross-validated ROC curves for the Super Learner (light green), discrete Super Learner (gray), and top-performing individual algorithm (random forests) on the HIV-1 antibody resistance data.

and non-censored IC_{50} viruses. The former group of viruses are thought to be resistant to VRC01, while the latter are thought to be sensitive to VRC01.

Appendix C

SUPPORTING INFORMATION FOR CHAPTER 5

C.1 Proofs of Theorem 1

Recall that

$$\psi_{m,n} := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{ E_{Q_m}[w(S)(z(S)\psi - v_{n,S})^2] + \lambda^\top (G\psi - c_n) \}, \tag{C.1}$$

and we have decomposed

$$\psi_{m,n} - \psi_0 = (\psi_{0,n} - \psi_0) + (\psi_{m,0} - \psi_0) + r_{m,n}, \tag{C.2}$$

where $\psi_{m,0}$ is obtained by replacing $v_{n,S}$ with $v_{0,S}$ in (C.1) and $r_{m,n} := (\psi_{m,n} - \psi_{m,0}) - (\psi_{0,n} - \psi_0)$.

We first control the first term in (C.2). Recall that

$$\psi_{0,n} := \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{ \| \sqrt{W}(Z\psi - v_n) \|_2^2 + \lambda^\top (G\psi - c_n) \}. \tag{C.3}$$

Thus, we can write $\psi_{0,n}$ in closed form as

$$\psi_{0,n} = A^{-1}[Z^\top W - G^\top \{ GA^{-1}G^\top \}^{-1} \{ GA^{-1}Z^\top W - e_{\emptyset,2^p} \}] v_n \tag{C.4}$$

$$= H(Q_0) v_n, \tag{C.5}$$

where $A := (Z^\top W Z)$, $e_{\emptyset,2^p} = [e_1^\top, e_{2^p}^\top]^\top \in \mathbb{R}^{2 \times 2^p}$, and e_j denotes the unit vector for position

j. Similarly, we can write

$$\psi_0 = H(Q_0)v_0. \tag{C.6}$$

Under the collection of conditions implied by (A1)–(A4) for each subset $s \in \mathcal{S}$, a straightforward application of the functional delta method and Theorem 2 of Chapter 4 yields that $\psi_{0,n}$ is an asymptotically linear estimator of ψ_0 with influence function given by

$$\phi_{0,1} : o \mapsto H(Q_0)\dot{V}_0(o), \tag{C.7}$$

where $\dot{V}_0$ is defined as in the main manuscript.

We now control the second term in (C.2). Note that $\psi_{m,0}$ is an M-estimator, maximizing the weighted least-squares Lagrangian $L(\psi, v_0, s, \lambda) = -\frac{1}{2}(v_{0,s} - z(s)\psi)^2 + \lambda^\top(G\psi - c_0)$. A straightforward application of Theorem 5.23 in van der Vaart [2000] yields that

$$\begin{aligned}\sqrt{m}(\psi_{m,0} - \psi_0) = & -\frac{1}{\sqrt{m}}\sum_{j=1}^{m}(-1)A^{-1}[z(S_j)^\top\{v_{0,S_j} - z(S_j)\psi_0\} + G^\top\lambda^*] + o_P(1) \\ = & \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\phi_{0,2}(S_j) + o_P(1),\end{aligned} \tag{C.8}$$

since

$$\begin{aligned}&\frac{\partial}{\partial\psi}L(\psi, v_0, s, \lambda) = z(s)^\top(v_{0,s} - z(s)\psi) + G^\top\lambda^* \text{ and} \\ &\frac{\partial^2}{\partial\psi_j\partial\psi_{j'}}\ell(\psi, v, s) = -z(s)_j z(s)_{j'}.\end{aligned}$$

Finally, we control the remainder term $r_{m,n}$. We can write

$$\psi_{m,n} = H(Q_m)v_n, \tag{C.9}$$

where

$$H(Q_m) := A_m^{-1}[Z_m^\top W - G^\top\{GA_m^{-1}G^\top\}^{-1}\{GA_m^{-1}Z_m^\top W - e_{\emptyset,2^p}\}],$$

where $A_m := (Z_m^\top W Z_m)^{-1}$ and Z_m is the matrix Z where the rows corresponding to elements of $\{s \in \mathcal{S} : s \notin \{S_1, \ldots, S_m\}\}$ are set to zero. Since the empirical distribution Q_m converges weakly to Q_0, then $H(Q_m) \to_p H(Q_0)$, since $m = n\gamma_n$ under (A5). Therefore, under the collection of conditions implied by (A1)–(A4) for each subset $s \in \mathcal{S}$ and (A5), an application of Slutsky's theorem yields that

$$\begin{aligned}(\psi_{m,n} - \psi_{m,0}) - (\psi_{0,n} - \psi_0) &= \{H(Q_m) - H(Q_0)\}(v_n - v_0) \\ &= o_P(n^{-1/2}).\end{aligned} \tag{C.10}$$

In view of (C.7), (C.8), and (C.10), we can write

$$\begin{aligned}\sqrt{n}(\psi_{m,n} - \psi_0) &= \sqrt{n}(\psi_{0,n} - \psi_0) + \sqrt{n}(\psi_{m,0} - \psi_0) + \sqrt{n}r_{m,n} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi_{0,1}(O_i) + \gamma_n^{-1/2}\phi_{0,2}(S_i)\} + o_P(1),\end{aligned} \tag{C.11}$$

where we have used the fact that O and S are sampled independently. Thus $\psi_{m,n}$ is an asymptotically linear estimator of ψ_0 with influence function $\phi_0 : (o, s) \mapsto \phi_{0,1}(o) + \phi_{0,2}(s)$. Since we operate in a nonparametric model space $\mathcal{M}$, this is the efficient influence function. Finally, note that as $\gamma_n \to \infty$, implying that all subsets are sampled, then the second term goes to zero.

C.2 Additional technical details

C.2.1 Shapley values as the solution of a weighted least squares problem

For convenience, we set $N := \{1, \ldots, p\}$ and $\mathcal{S} := \{S : S \subseteq N\}$. Then $|\mathcal{S}| = 2^p$. Let v_0 be the vector of all measures of predictiveness, with $v_{0,\emptyset}$ well-defined. Let $v_{0,S}$ be the predictiveness

of $\{X_j\}_{j\in S}$. Let $z(S)$ be a binary vector with a 1 in position one and a 1 in position $j+1$ if $j \in S$. We define the weights $w(S) = \binom{p-2}{|S|-1}^{-1}$ for $S \in \mathcal{N} \setminus \{\emptyset, N\}$ and $w(\emptyset), w(N)$ are irrelevant due to the equality constraints (so we set them equal to one). We additionally define $W \in \mathbb{R}^{2^p \times 2^p}$ as a diagonal matrix of weights. Finally, $Z \in \mathbb{R}^{2^p \times (p+1)}$ consists of the stacked $z(S)$ vectors.

Recall the classical Shapley formula

$$\psi_{0,j} = \frac{1}{p} \sum_{s \in \mathcal{S}} \binom{p-1}{|s|}^{-1} \{V(f_{0,s\cup j}, P_0) - V(f_{0,s}, P_0)\}.$$

Then we can write, for any p, $\psi_{0,j} = [B(p)]_{j+1} v_{0,j}$, where $B(p)$ is a matrix encoding the weights for the classical Shapley values. Notice that the first row of $B(p)$, denoted $[B(p)]_1$, is given by $[B(p)]_1 = z(\emptyset)$. For each $j = 2, \ldots, p+1$, and for $i = 1, \ldots, 2^p$,

$$[B(p)]_{ji} := \frac{1}{p}(-1)^{I\{(j-1)\in s_i\}} \binom{p-1}{|s_i| - I\{(j-1) \in s_i\}}^{-1},$$

where the s_i have the same order as the rows of Z. This is easy to verify in the case $p = 2$; there,

$$B(2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

We now state and prove three results that we will use to prove that the Shapley values are the solution of a weighted least squares problem.

The first result provides a valid way of writing the classical Shapley values as a vector.

Lemma 5. *$B(p)Z = I_{(p+1)\times(p+1)}$ for any p.*

Proof. Set $Z = [T_1, T_2, \ldots, T_{p+1}]$, where $T_1 = 1_{2^p}$ and $T_{ji} = I\{(j-1) \in s_i\}$ for $j = 2, \ldots, p+1$

and $i = 1, \ldots, 2^p$. Then by definition, $B(p)T_1 = z(\emptyset)$. Additionally, for $j = 2, \ldots, p+1$,

$$\begin{aligned}
[B(p)]_j T_j &= \sum_{i=1}^{2^p} \frac{1}{p}(-1)^{I\{(j-1)\in s_i\}} \binom{p-1}{|s_i| - I\{(j-1)\in s_i\}}^{-1} I\{(j-1)\in s_i\} \\
&= \frac{1}{p} \sum_{S\subseteq N:(j-1)\in S} \binom{p-1}{|S|-1}^{-1} \\
&= \frac{1}{p} \sum_{k=1}^{p} \sum_{S\subseteq N:(j-1)\in S, |S|=k} \binom{p-1}{k-1}^{-1} \\
&= \frac{1}{p} \sum_{k=1}^{p} \binom{p-1}{k-1}\binom{p-1}{k-1}^{-1} = 1.
\end{aligned}$$

For $j = 2, \ldots, p+1$ and $k = 2, \ldots, p+1$ with $j \neq k$,

$$\begin{aligned}
[B(p)]_j T_k &= \sum_{i=1}^{2^p} \frac{1}{p}(-1)^{I\{(j-1)\in s_i\}} \binom{p-1}{|s_i| - I\{(j-1)\in s_i\}}^{-1} I\{(k-1)\in s_i\} \\
&= \frac{1}{p}\left[\sum_{S\subseteq N:(j-1),(k-1)\in S} \binom{p-1}{|S|-1}^{-1} - \sum_{S\subseteq N:(j-1)\notin S,(k-1)\in S} \binom{p-1}{|S|}^{-1} \right] \\
&= \frac{1}{p}\left[\sum_{k=2}^{p} \binom{p-2}{k-2}\binom{p-1}{k-1}^{-1} - \sum_{k=1}^{p-1} \binom{p-1}{k-1}\binom{p-1}{|S|}^{-1} \right] = 0.
\end{aligned}$$

Thus, $[B(p)]_j Z$ has a zero in the first position and a one in position j, for $j = 2, \ldots, p+1$, so $B(p)Z = I$. □

Now define the $(p+1)$-dimensional vector $\psi_0 := [v_{0,\emptyset}, \psi_{0,1}, \ldots, \psi_{0,p}]$. Then clearly $\psi_0 = B(p)v_0$. The next lemma states that ψ_0 solves an unconstrained minimization problem.

Lemma 6. *The Shapley value vector ψ_0 solves the minimization problem*

$$\underset{\psi\in\mathbb{R}^{p+1}}{\text{minimize}} \sum_{S\in\mathcal{N}} w(S)[z(S)\psi - v_{0,S}]^2.$$

Proof. Recall that $\psi_0 = B(p)v_0$. Rewriting the minimization in Lagrangian form yields

$$\begin{aligned} L(\psi) &= \sum_{S \in \mathcal{N}} w(S)[z(S)\psi - v_{0,S}]^2 \\ &= \|\sqrt{W}(Z\psi - v_0)\|_2^2. \end{aligned}$$

Thus, the Karush-Kuhn-Tucker (KKT) conditions of this Lagrangian are

$$2(Z^\top W Z)\psi^* = 2Z^\top W v_0.$$

Plugging in ψ_0 to the left-hand side of the above equality and using the result of Lemma 5 yields

$$\begin{aligned} 2(Z^\top W Z)\psi_0 &= 2(Z^\top W Z)B(p)v_0 \\ &= 2Z^\top W v_0. \end{aligned}$$

Thus ψ_0 satisfies the KKT conditions for the unconstrained minimization problem, and

$$\psi_0 \in \underset{\psi \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|\sqrt{W}(Z\psi - v_0)\|_2^2.$$

□

Our final lemma states that ψ_0 additionally satisfies the KKT conditions of a constrained optimization problem.

Lemma 7. *The Shapley value vector ψ_0 satisfies the constraints that $\psi_{0,0} = v_{0,\emptyset}$ and $\sum_{j=0}^{p} \psi_j = v_{0,N}$.*

Proof. Define $G = \begin{bmatrix} z(\emptyset) \\ z(N) \end{bmatrix}$ and $c_0 = [v_{0,\emptyset}, v_{0,N}]^\top$. By definition, we have that

$$G\psi_0 = GB(p)v_0 = c_0,$$

thus proving the claim. □

Now consider the minimization problem

$$\underset{\psi \in \mathbb{R}^{p+1}}{\text{minimize}} \sum_{S \in \mathcal{N}} w(S)[z(S)\psi - v_{0,S}]^2 \tag{C.12}$$

$$\text{subject to } \psi_0 = v_{0,\emptyset} \text{ and } \sum_{j=0}^{p} \psi_j = v_{0,N}. \tag{C.13}$$

Using Lemmas 5–7, we have that the Shapley value vector ψ_0 satisfies the KKT conditions of the Lagrangian corresponding to this constrained minimization problem, and thus that

$$\psi_0 \in \underset{\psi \in \mathbb{R}^{p+1}}{\text{argmin}} \sum_{S \in \mathcal{N}} w(S)[z(S)\psi - v_{0,S}]^2$$
$$\text{subject to } \psi_0 = v_{0,\emptyset} \text{ and } \sum_{j=0}^{p} \psi_j = v_{0,N}.$$

C.3 Additional details for predicting mortality of patients in the intensive care unit

In this section, we describe our analysis of data on patients' stays in the intensive care unit (ICU) [Silva et al., 2012] in more detail.

First, we computed the same summary features as in Feng et al. [2018]. These features are: 27 computed features, including the mean, minimum and maximum, and the last measurement, based on 18 of the 37 original variables; and the minimum, maximum, and mean (from fitting linear regression) from the time series of the 18 original variables if they were not already included. In addition, we added five variables that are used in the simplified

acute physiology (SAPS) I and SAPS II scores but were not in the set of 18 original variables used to compute features above. Finally, we included all general descriptors measured at admission. This procedure resulted in a total of 55 features that we used to predict mortality. We present all of these features in Table C.1 (this is a reproduction of Table C.2 in the Supplement of Feng et al. [2018]). The features are also naturally grouped in to *medical test groups* (summary features for variables measured by the same medical test)

We estimate the importance of 25 variable groups which fall into two categories: "medical test groups" contain summary features for variables measured by the same medical test and "individual variable groups" contain summary features from the same variable. Here, we discuss our results for the individual variable groups; medical test groups are discussed in the main manuscript.

Variable group	Variable name	Summary (computed or original)
GCS	GCS	last, weighted mean, max, min, slope
Metabolic panel	HCO3	min, max, last, weighted mean
	BUN	min, max, last, weighted mean
	Na	min, max, weighted mean
	K	min, max, weighted mean
	Glucose	min, max, weighted mean
Systolic blood pressure	SysABP	min, max, last, weighted mean
CBC	WBC	min, max, last, weighted mean
	HCT	min, max, weighted mean
Temp	Temp	min, max, last, weighted mean
Lactate	Lactate	min, max, last, weighted mean
HR	HR	min, max, weighted mean
Respiration	RespRate	min, max, weighted mean
	MechVent	max
	FiO2, PaO2	ratio of means
Urine	Urine	sum (based on SAPS II urine item)
General Desc.	Gender	measured at admission
	Height	measured at admission
	Weight	measured at admission
	Age	measured at admission
	ICU admission type	measured at admission

Table C.1: Features included for analysis of the ICU data. CBC: complete blood count test. Weighted mean: estimated response at mean measurement time from a linear regression of response on time. Slope: estimated slope from a linear regression of response on time. Last: last measurement. Impossible values (e.g., ≤ 0 for many variables) were dropped.