UW Biostatistics Working Paper Series

8-30-2017

Nonparametric variable importance assessment using machine learning techniques

Brian D. Williamson
Department of Biostatistics, University of Washington, brianw26@uw.edu

Peter B. Gilbert
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, pgilbert@scharp.org

Noah Simon
Department of Biostatistics, University of Washington, nrsimon@uw.edu

Marco Carone
Department of Biostatistics, University of Washington, mcarone@uw.edu

Suggested Citation

Williamson, Brian D.; Gilbert, Peter B.; Simon, Noah; and Carone, Marco, "Nonparametric variable importance assessment using machine learning techniques" (August 2017). *UW Biostatistics Working Paper Series.* Working Paper 422.
http://biostats.bepress.com/uwbiostat/paper422

Nonparametric variable importance assessment using machine learning techniques

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

[1]Department of Biostatistics, University of Washington
[2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center

August 30, 2017

Abstract

In a regression setting, it is often of interest to quantify the importance of various features in predicting the response. Commonly, the variable importance measure used is determined by the regression technique employed. For this reason, practitioners often only resort to one of a few regression techniques for which a variable importance measure is naturally defined. Unfortunately, these regression techniques are often sub-optimal for predicting response. Additionally, because the variable importance measures native to different regression techniques generally have a different interpretation, comparisons across techniques can be difficult. In this work, we study a novel variable importance measure that can be used with any regression technique, and whose interpretation is agnostic to the technique used. Specifically, we propose a generalization of the ANOVA variable importance measure, and discuss how it facilitates the use of possibly-complex machine learning techniques to flexibly estimate the variable importance of a single feature or group of features. Using the tools of targeted learning, we also describe how to construct an efficient estimator of this measure, as well as a valid confidence interval. Through simulations, we show that our proposal has good practical operating characteristics, and we illustrate its use with data from a study of the median house price in the Boston area, and a study of risk factors for cardiovascular disease in South Africa.

Keywords: machine learning; nonparametric R^2; statistical inference; targeted learning; variable importance.

Corresponding author: Brian D. Williamson, Department of Biostatistics, University of Washington, F-600, Health Sciences Building, Box 357232, Seattle, WA 98195-7232. Email: brianw26@uw.edu.

1 Introduction

Suppose that the observed data include independent draws $O_1, O_2, \ldots, O_n$ 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. It is often of interest to understand the association between Y and X under P_0. For this purpose, it may be useful to consider the conditional mean function μ_{P_0}, where for each $P \in \mathcal{M}$ we define

$$\mu_P(x) := E_P(Y \mid X = x) \ . \tag{1}$$

There are many tools for estimating μ_{P_0}: classical parametric techniques (e.g., linear regression), and more flexible nonparametric or semiparametric methods, including smoothing splines (Reinsch, 1967), random forests (Breiman, 2001), generalized additive models (Hastie and Tibshirani, 1990), loess smoothing (Cleveland, 1979), artificial neural networks (Barron, 1989), and kernel smoothing (Wand and Jones, 1994), 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 $s \subseteq \{1, 2, \ldots, p\}$ and distribution $P \in \mathcal{M}$, we may define the reduced conditional mean

$$\mu_{P,s}(x) := E_P\left(Y \mid X_{(-s)} = x_{(-s)}\right) \ , \tag{2}$$

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.

There has been recent work on evaluating variable importance without relying on overly strong modeling assumptions. Proposals for flexible variable importance assessment can generally be categorized as being either (a) intimately tied to a specific estimation technique for the conditional mean function or (b) agnostic to the estimation technique used. Variable importance measures falling into the former category include the native variable importance measure for random forests (Breiman, 2001), variable importance in neural networks (see, e.g., Olden et al., 2004), and ANOVA in linear models. Among these, ANOVA alone appears to allow formal statistical inference. Additionally, even restricting our attention to methods for which a variable importance measure is naturally defined, it is generally not possible to directly compare the importance assessment stemming from these different methods: they are usually measuring different quantities and thus have different interpretations. Examples of estimation technique-agnostic variable importance measures include nonparametric extensions of R^2 (Doksum and Samarov, 1995); and the risk difference, $E_{P_0}(Y \mid A = a, W = w) - E_{P_0}(Y \mid A = 0, W = w)$ for $X = (A, W)$, or expected risk difference, $E_{P_0}\{E_{P_0}(Y \mid A = a, W = w) - E_{P_0}(Y \mid A = 0, W = w)\}$ (van der Laan, 2006), with extensions studying the best linear approximation of the risk difference (Chambaz et al., 2012) and interval-censored survival outcomes (Sapp et al., 2014). These methods all allow formal inference, but they may not have a desirable interpretation in the context of many scientific problems. Despite their broad potential applicability, many of these proposals have only been studied in the context of specific estimation strategies. For example, Doksum and Samarov (1995) only consider the use of kernel-based estimators of the underlying regression function, though recent work extends their results to local polynomial regression (Huang and Chen, 2008), functional regression (Yao et al., 2005), and different test statistics for the null hypothesis of no variable importance (Fan and Li, 1996).

In our view, an ideal variable importance measure should (i) be entirely agnostic to the estimation technique, (ii) allow formal inference, and (iii) provide an interpretation that is well suited to scientific applications. In this work, we propose a variable importance measure that satisfies each of these criteria. In particular, we consider inference on the variable importance measure

$$\psi_{0,s} := \frac{\int \{\mu_{P_0}(x) - \mu_{P_0,s}(x)\}^2 \, dP_0(x)}{var_{P_0}(Y)} \,. \tag{3}$$

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) 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

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

the difference in the R^2 either obtained using the full set of covariates or the reduced set of covariates only. Thus, the parameter we focus on can be seen as a simple generalization of the classical R^2 measure to a nonparametric model. 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.

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 estimators of these regression function estimates into (3) will not yield a regular and asymptotically linear, let alone efficient, estimator of $\psi_{0,s}$. In this manuscript, 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 present some properties of our parameter of interest and give our proposed estimator in Section 2. In Section 3, we provide empirical evidence that our estimator outperforms the naive plug-in estimator in settings where the covariate vector is low- or moderate-dimensional. In Section 4, we illustrate the use of our method in the context of the benchmark Boston housing study data. In Section 5, we apply our method on data from a retrospective study of heart disease in South African men. We provide concluding remarks in Section 6. Technical details are provided in Part 1 of the Supplementary Materials.

2 Variable importance in a nonparametric model

2.1 Parameter of interest

We work in a fully unrestricted, and hence nonparametric, model $\mathcal{M}$. For given $s \subseteq \{1, 2, \ldots, p\}$ and $P \in \mathcal{M}$, we define the statistical functional

$$\Psi_s(P) := \frac{\int \left\{\mu_P(x) - \mu_{P,s}(x)\right\}^2 dP(x)}{var_P(Y)} \tag{4}$$

using the conditional means defined in (1) and (2); this is the nonparametric measure of variable importance we focus on. Using observations $O_1, O_2, \ldots, O_n$ independently drawn from $P_0 \in \mathcal{M}$, our objective is to make efficient inference about the true value $\psi_{0,s} := \Psi_s(P_0)$ of the variable importance measure corresponding to the components of X with index in s, as implied by the data-generating mechanism P_0. If we are interested in a parsimonious description of the interplay between outcome Y and covariate vector X, determining which features of X are most important may be of interest, and inference on $\psi_{0,s}$ for various choices of s may help in determining precisely this.

We note that this parameter involves two parts. *First*, the numerator of $\psi_{0,s}$ consists of the squared difference in conditional means, averaged over the marginal distribution of the features. On one hand, if the two conditional means are quite different, this expected difference is large. Hence, information is lost by excluding the components X_s when using the conditional mean to explain the outcome Y. On the other hand, if the difference is small, not much information may be lost, and perhaps using the covariates in $X_{(-s)}$ may suffice. The numerator can then be interpreted as the amount of variability in the outcome Y explained by including X_s in the conditional mean. *Second*, the denominator of $\psi_{0,s}$ is the total variability of Y. It follows then that $\psi_{0,s}$ is a proportion between 0 and 1. In particular, it is on the same scale for each s, which allows us to easily compare values for different covariates or groups of covariates. As such, $\psi_{0,s}$ is indeed a generalization of ANOVA-derived variable importance, where we consider the ratio of the amount of variability explained by including X_s in the regression to the total variability of the outcome. Because efficient estimation of $var_{P_0}(Y)$ requires no work at all – in an unconstrained model, the empirical variance estimator is optimal – we will only need smoothing techniques and flexible estimation methods to estimate the numerator of $\psi_{0,s}$.

We now discuss some properties of Ψ_s that are relevant to building an efficient estimator of $\psi_{0,s}$. Specifically, we require that the functional (4) be appropriately differentiable, and that a functional Taylor expansion holds with negligible higher-order terms.

The functional (4) is pathwise differentiable (see, e.g., Bickel et al., 1998), a result that we prove in Part 1 of the Supplementary Material. 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 given explicitly 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*

$$o \mapsto D_{P,s}(o) := 2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,s}(x)\} + \{\mu_P(x) - \mu_{P,s}(x)\}^2 - \Phi_s(P) \ , \tag{5}$$

$$o \mapsto D^*_{P,s}(o) := \frac{2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,s}(x)\} + \{\mu_P(x) - \mu_{P,s}(x)\}^2}{var_P(Y)} - \Phi_s(P)\left\{\frac{y - E_P(Y)}{var_P(Y)}\right\}^2 . \tag{6}$$

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

$$\Phi_s(P) = \Phi_s(P_0) + \int D_{P,s}(o)d(P - P_0)(o) + R_s(P, P_0) \ , \tag{7}$$

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 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, using elementary algebraic manipulations,

$$\begin{aligned}\Phi_s(\widehat{P}_n) - \Phi_s(P_0) \ &= \int D_{\widehat{P}_n,s}(o)d(\widehat{P}_n - P_0)(o) + R_s(\widehat{P}_n, P_0) \\ &= \int D_{\widehat{P}_n,s}(o)d(\mathbb{P}_n - P_0)(o) + R_s(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n} D_{\widehat{P}_n,s}(O_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} D_{P_0,s}(O_i) + \int \left\{D_{\widehat{P}_n,s}(o) - D_{P_0,s}(o)\right\} d(\mathbb{P}_n - P_0)(o) + R_s(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n} D_{\widehat{P}_n,s}(O_i) \ ,\end{aligned} \tag{8}$$

where $\mathbb{P}_n$ is the empirical distribution based on $O_1, O_2, \ldots, O_n$, 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 $\Phi_s(\widehat{P}_n)$. Its four distinct summands can be studied separately. The first summand is an empirical average of mean-zero transformations of $O_1, O_2, \ldots, O_n$ – this term will determine the asymptotic behavior of our eventual estimator, as discussed in Section 2.2. The second summand is an empirical process term that we can show is asymptotically negligible under certain conditions on $\widehat{P}_n$. The third term is a second-order remainder term that we can similarly show is asymptotically negligible. The fourth term can be thought of as the bias incurred from flexibly estimating the conditional means (1) and (2), and in general, it will tend to zero slowly. This bias term motivates our choice of estimator for $\psi_{0,s}$ in Section 2.2. Specifically, we will choose one particular method of correcting for this bias term, and the large sample properties of our proposed estimator will then be determined by the first summand in (8).

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

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

where $\bar{Y}_n$ is the empirical mean of $Y_1, Y_2, \ldots, Y_n$. In turn, this suggests using

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

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 $\mu_{P_0,s}$, tuning parameters must generally be chosen to ensure an adequate bias-variance tradeoff. Construction of $\hat{\mu}$ 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. However, we view estimation of the regression functions as a nuisance, since we are ultimately interested in estimating $\psi_{0,s}$. Hence, we need to tailor the estimation procedure to make the appropriate bias-variance tradeoff for estimating $\psi_{0,s}$ rather than each of μ_{P_0} and $\mu_{P_0,s}$. Without such tailoring, the estimator $\hat{\psi}_{\text{naive},s}$ is generally overly biased and thus neither efficient nor regular and asymptotically linear. 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 (8), will be asymptotically efficient under certain regularity conditions. This estimator, which is often referred to as the one-step estimator, is obtained by correcting for the excessive bias of the naive plug-in estimator $\hat{\phi}_{\text{naive},s}$. Upon close examination, 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 but rather to construct 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

$$\hat{\psi}_{n,s} \;=\; \frac{\hat{\phi}_{n,s}}{var_{\mathbb{P}_n}(Y)} \;=\; \hat{\psi}_{\text{naive},s} + \frac{\sum_{i=1}^n 2\{Y_i - \hat{\mu}(X_i)\}\{\hat{\mu}(X_i) - \hat{\mu}_s(X_i)\}}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}\ . \tag{9}$$

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 $\psi_{0,s}$.

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. One way to estimate these two conditional mean functions is to use any specific regression technique to regress the outcome Y on the full covariate vector X and then on the reduced vector $X_{(-s)}$ of covariates. However, this strategy 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};
2. regress $\hat{\mu}(X)$ on $X_{(-s)}$ to obtain an estimate $\hat{\mu}_s$ of $\mu_{P_0,s}$.

The final estimating procedure we recommend for $\psi_{0,s}$ consists of the estimator (9), where the conditional means involved are estimated using flexible regression estimators and this sequential regression approach.

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, 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 $R_s(P, P_0)$ 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. *The linearization (7) holds with second-order remainder term given explicitly by*

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

Furthermore, $R_s(\widehat{P}_n, P_0) = o_P(n^{-1/2})$ *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 both* $o_P(n^{-1/2})$.

Each remainder term is a sum of several terms, each of which is a product of two terms that 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 (8) is asymptotically negligible.

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 (van der Vaart, 2000) 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 very flexible regression estimators, there may be reason for concern regarding the validity of the Donsker class condition. In such cases, a cross-validated version of the one-step procedure involved in our proposed estimator (see, e.g., van der Laan and Rubin, 2005) can be used to circumvent this condition altogether. While this cross-validated estimator is only marginally more complex than the estimator proposed here, we restrict attention to studying the simpler estimator.

The following theorem builds upon these two lemmas to describe the asymptotic behavior of the proposed estimator.

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 normal random variable with variance $\sigma^2_{0,s} := \int \{D^*_{P_0,s}(o)\}^2 dP_0(o)$.*

A natural plug-in estimator of the standard error of $\hat{\psi}_{n,s}$ 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.

Since we must already compute $\hat{\mu}$ and $\hat{\mu}_s$ to obtain the naive estimator, computing the proposed estimator and its standard error estimate takes minimal extra time. When flexible estimators of the involved regression are used, the naive estimator is generally not asymptotically linear: it will usually be irregular and have 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 prohibitive computational burden of such an approach, theory suggests that this strategy is likely invalid in this context.

2.4 Invariance to transformations

So far, we have defined a nonparametric measure of variable importance and proposed an efficient estimator for this parameter that allows valid inference under mild regularity conditions. Our parameter can be interpreted as the additional proportion of variability in the outcome explained by including a single covariate or group of covariates when using the conditional mean as a proxy for the outcome.

In some applications, it is common to center and standardize the features – and sometimes even the outcome – 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 2. *Suppose 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 $g_j : \mathbb{R} \to \mathbb{R}$, $j = 1, 2, \ldots, p$, and that $g_Y : \mathbb{R} \to \mathbb{R}$ is a linear function. If $P_{0,g}$ is the distribution of $(g_X(X), g_Y(Y))$ induced by P_0, then $\Psi_s(P_{0,g}) = \Psi_s(P_0)$.*

The variable importance measure we have proposed 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 individual features and the outcome.

We note here that 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 also 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 reasonable 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.

2.5 Behavior under the zero-importance null hypothesis

This work primarily focuses on developing an efficient estimator of a variable importance measure proposed 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 $D_{P_0,s}$ is identically zero under this null, 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. (2014), 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 Experiments on simulated data

We now present empirical results describing the performance of the proposed estimator compared to that of the naive estimator. We consider settings in which the total number of features is relatively low or moderately large. In both settings, we compute and display the empirical bias and variance of the estimators as well as the empirical coverage of nominal 95% confidence intervals. In all implementations, we use the sequential regression estimating procedure described in Section 2.2 to compute compatible estimates of the required regression functions, and we compute Wald-type confidence intervals as outlined in Section 2.3.

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 $n \in \{100, 300, 500, 700, 1000, 2000, \ldots, 10000\}$ 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-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. Because we obtained essentially the same results 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 estimator and respective confidence intervals for each of $B = 1,000$ replications. Because of the unavailability of a simple asymptotic distribution for the naive estimator, a percentile bootstrap approach with 80 bootstrap samples was used to 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 1 displays the results of this simulation. Values relating to the proposed estimator are depicted in blue, while they are in red for the naive estimator. Circles and stars denote $j = 1$ and $j = 2$, respectively. In the left panel, we note that the Monte Carlo error is relatively small, regardless of sample size – since B is large, this is not surprising. 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 the naive estimator has substantial bias that does not tend to zero faster than $n^{-1/2}$. This coincides with our expectations, since 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. However, we also see that the proposed estimator for $j = 2$ appears to dip slightly below zero for large n, though we expect for larger n to see the scaled bias of the proposed estimator get closer to zero. Numerical error in our computations may explain why this does not exactly happen, though there is very substantial bias reduction from using the proposed estimator regardless. These results provide empirical evidence that the one-step correction performed 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 1, we see that the variance of the proposed estimator is essentially the same as that of the naive estimator – we have thus not suffered much at all from removing excess bias in our estimation procedure. 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 the naive estimator decreases instead and very quickly becomes completely unsatisfactory. When we take into account the fact that bootstrapping a confidence interval adds computation time, the procedure based on the proposed estimator appears to substantially outperform that using the naive estimator in both computation time and coverage.

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 3,000 random datasets of size $n \in \{100, 300, 700, 1000, 2000, \ldots, 4000\}$ 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 as in the previous simulation. The empirical bias and variance for each estimator were computed and scaled as before, and the empirical coverage of confidence intervals was also evaluated.

Figure 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 ($j = 1$). However, when we remove the feature with zero importance ($j = 2$), the proposed estimator has slightly higher bias. While this is somewhat 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. The ratio of the variances is close to one for the predictive feature, but is less than one for the null feature. This indicates a somewhat larger variance when 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 ahead.

3.3 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 $\{1,2,3,4,5\}$, $\{6,\ldots,10\}$ and $\{11,\ldots,15\}$ for each sample size. Details on the analysis of additional groups of features are provided in Part 2 of the Supplementary Materials. The true value of the variable importance measure corresponding to each of the considered groups is given in Table 2.

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 Table 2. As in setting A, results for the analysis of additional groupings are provided in Part 2 of the Supplementary Materials.

For each scenario considered, we estimated the conditional mean functions using gradient boosted trees (Friedman, 2001) fit using the GradientBoostingRegressor function in the sklearn module in Python. 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 Figures 9 and 10. First, from Figure 9, we note that as n increases, the scaled empirical bias of the proposed estimator approaches zero while that of the naive estimator increases in magnitude across all three groupings s considered. From Figure 10, 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 Figures 12 and 13. From Figure 12, 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. We observe similar coverage results as in setting A from Figure 13: intervals based on the proposed estimator have uniformly higher coverage than those based on the naive estimator.

Regardless of whether or not the data are correlated, the proposed estimator performs significantly better than the naive estimator in these simulations. How well the proposed estimator performs appears to be tied to the degree of correlation, with higher levels of correlation 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. Indeed, this is a sensible approach for dealing with correlated features, which necessarily render variable importance assessment challenging. We note that 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 13) approaches the nominal level with increasing sample size, indicating that the one-step 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. Each group of five features has at least two null features, and some have more. For example, when evaluating the importance of the group $(X_1, X_2, \ldots, X_5)$, the group $(X_8, X_9, X_{10}, X_{12}, X_{13}, X_{14}, X_{15})$ 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.

4 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 on the UC Irvine Machine Learning Repository. 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, with housing prices expected to increase with decreased distance to employment centers; and an index of accessibility to radial highways, with housing prices expected to increase with increased highway access. 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 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 2.2 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 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 7. First, we see a difference in the ordering of features based on estimated importance using the naive and proposed estimators. 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. Since the neighborhood group is so large, its large importance is not surprising. The average number of rooms in a house also tends to increase its price, which thus contributes to most of the relative importance of the structural group. Interestingly, the crime rate appears to be the least important individual feature in predicting median house value. One might expect *a priori* that crime rate would have a large effect on 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, all of which are specific to the particular method under consideration. Comparing our results to those obtained by three other groups of investigators – Doksum and Samarov (1995), Friedman and Popescu (2008) 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 Bi et al. (2003), Doksum and Samarov (1995), and many applications of random forest alone; this is not consistent with the findings of Friedman and Popescu (2008). After these two features, the ranking tends to differ based on the prediction algorithm used. 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 every method besides that of Doksum and Samarov (1995). 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.

5 Results from the South African heart disease study data

We consider a subset of the data from the Coronary Risk Factor Study, a retrospective cross-sectional sample of 462 white males aged 15 – 64 in a region of the Western Cape, South Africa. The primary aim of this study was to establish the prevalence of ischemic heart disease risk factors in this high incidence region. These data are a subset of a larger dataset described in Rousseauw et al. (1983), and are publicly available as one of the datasets used in Hastie et al. (2009). For each participant, the presence or absence of myocardial infarction (MI) at the time of the survey is recorded. This dataset includes 160 cases and 302 controls. In addition, measurements of systolic blood pressure (mmHg), cumulative tobacco consumption (kg), LDL cholesterol (mg/dL), adiposity (similar to body mass index), family history of heart disease (binary), type A behavior (binary), obesity, current alcohol consumption, and age are available.

These features can naturally be grouped into two sets: behavioral features (tobacco consumption, alcohol consumption, and type A behavior), and biological features (systolic blood pressure, LDL cholesterol, adiposity, obesity, family history, 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 estimate the conditional means using the Super Learner, with the same library of learners as in the previous section. Then, we used the sequential regression estimating procedure to calculate both the naive and proposed estimators, and 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 8. 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 the two groups of features – biological and behavioral – are important, with biological factors more important than behavioral factors (tobacco consumption, alcohol consumption, and type A behavior). The most important individual feature is family history of heart disease – this is consistent with the fact that family history has been found to be a risk factor of MI in previous studies. The fact that both groups of features are more important than any individual feature besides family history appears scientifically sensible.

We compared these results to the logistic regression model fit to these data in Hastie et al. (2009). Based on the absolute values of z-statistics, logistic regression picks age as most important ($z = 4.184$) followed immediately by family history ($z = 4.178$). This slight difference is captured in our uncertainty estimates (Figure 8): there, we see that the point estimates for age and family history are close, and their confidence intervals almost overlap. Logistic regression picks the next two most important variables as LDL cholesterol ($z = 3.129$) and tobacco consumption ($z = 3.034$); we find the opposite ordering, but again see remarkably similar point estimates and nearly overlapping intervals. 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 estimating 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 much less likely to be driven by potential model misspecification.

6 Conclusion

We have developed a novel measure of variable importance, interpreted as the additional proportion of variability in the outcome explained by including a single feature or a group of features in the conditional mean outcome given all available features. 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 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, both when the vector of covariates is low or moderate-dimensional. 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 seem 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 influenced 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.

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 an aggressive use of super learning with a wide range of candidate learners, ranging from the very parametric to the fully nonparametric. This flexibility mitigates concerns regarding model misspecification. Additionally, we also suggest the use of sequential regressions to minimize any incompatibility between the two conditional means estimated.

Software

We implement the methods discussed above in the R package `vimp` and the Python package `vimpy`, both freely available on the author's Github page at https://github.com/bdwilliamson/vimp and https://github.com/bdwilliamson/vimpy, respectively.

Supplementary Materials

Technical details and additional results from the moderate-dimensional simulations are available in the supplementary document.

Acknowledgements

The authors wish to thank Professor Antoine Chambaz for insightful comments that improved this manuscript. This work was supported by U.S. Public Health Service Grant UM1 AI068635 [SDMC: HIV Vaccine Trials Network] by the National Institute of Allergy and Infectious Diseases (NIAID). The opinions expressed in this article are those of the authors and do not necessarily represent the official views of the NIAID or the National Institutes of Health.

References

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

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.

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

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

M. Carone, I. Díaz, and M. van der Laan. Higher-order targeted minimum loss-based estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 2014.

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

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

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*, 23:1443–1473, 1995.

Y. Fan and Q. Li. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the Econometric Society*, 64:865–890, 1996.

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

J. Friedman and B. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2:916–954, 2008.

R. Gill, M. van der Laan, and J. 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.

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

T. Hastie and R. Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

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

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

J. Olden, M. Joy, and R. Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389–397, 2004.

C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

J. Rousseauw, 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(430-436):216, 1983.

S. Sapp, M. 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. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

M. van der Laan and D. Rubin. Estimating function based cross-validation and learning. *University of California at Berkeley Division of Biostatistics Working Paper Series*, (180), 2005.

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

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

M. Wand and M. Jones. *Kernel Smoothing.* CRC Press, 1994.

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

Table 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.

Group	Setting A	Setting 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

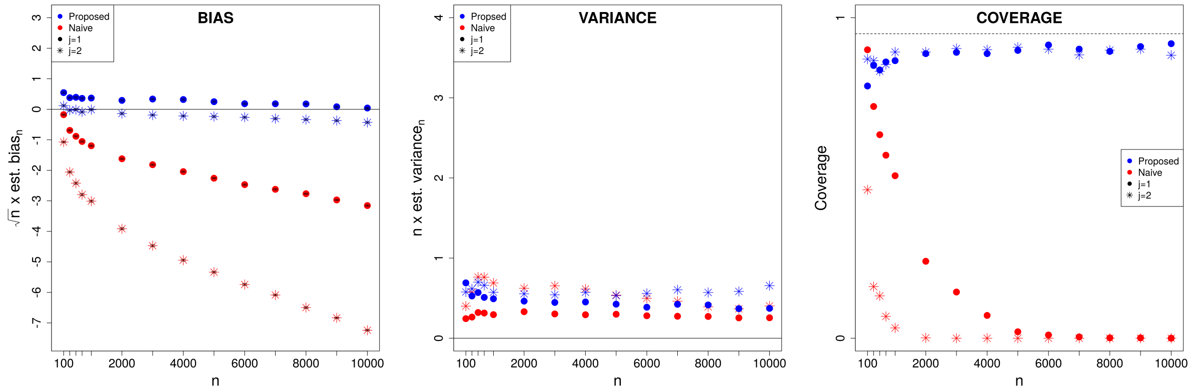

Figure 1: Empirical bias scaled by $\sqrt{n}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to estimate (1)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

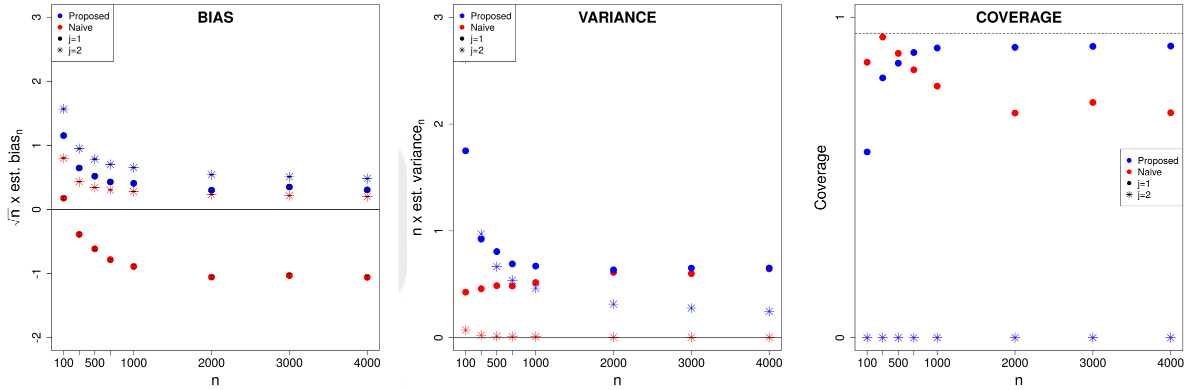

Figure 2: Empirical bias scaled by $\sqrt{n}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to estimate (1)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

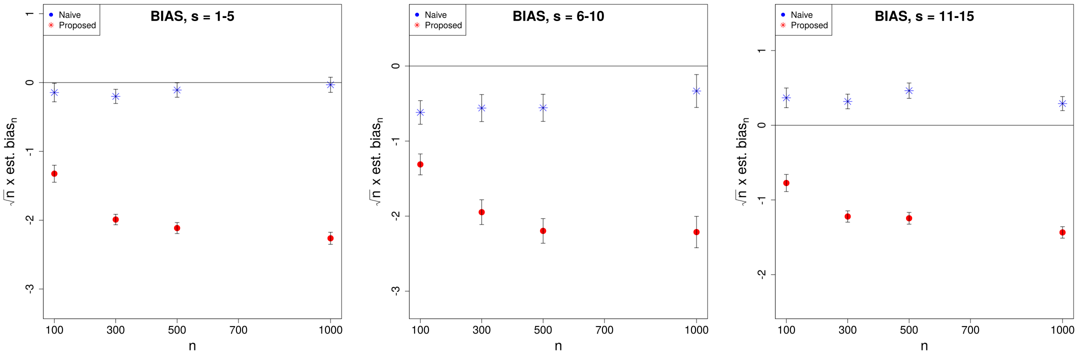

Figure 3: Empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting A, using gradient boosted trees to estimate (1) and (2). We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator. Monte Carlo error bars are displayed vertically.

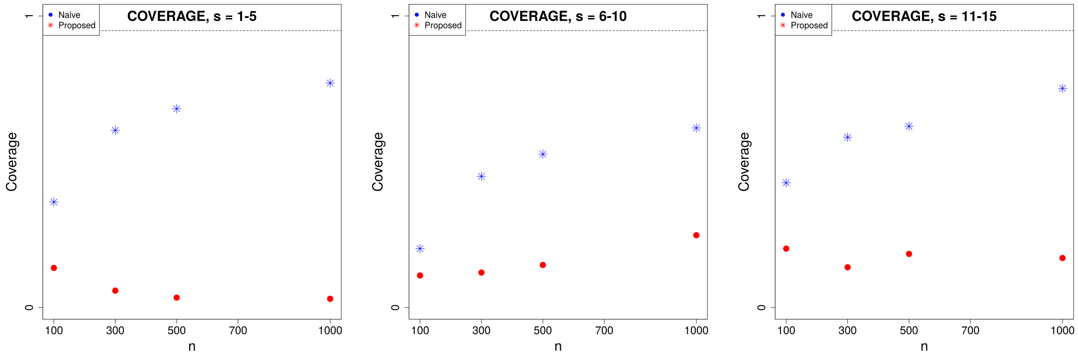

Figure 4: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting A, using gradient boosted trees to estimate (1) and (2). We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

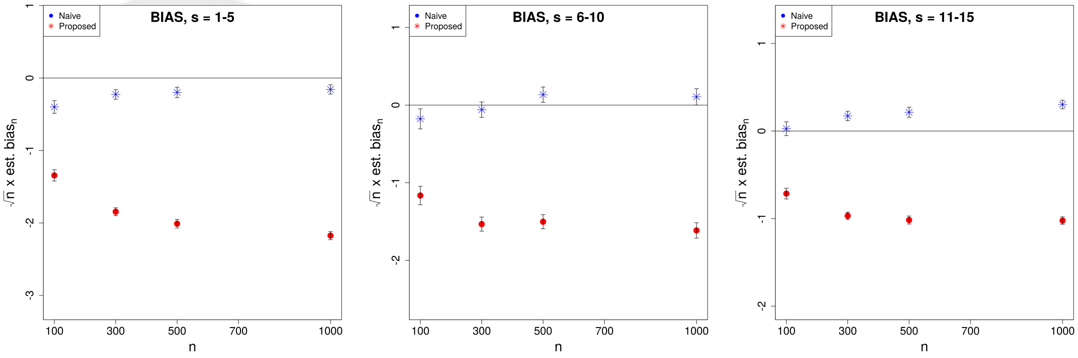

Figure 5: Empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting B, using gradient boosted trees to estimate (1) and (2). We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator. Monte Carlo error bars are displayed vertically.

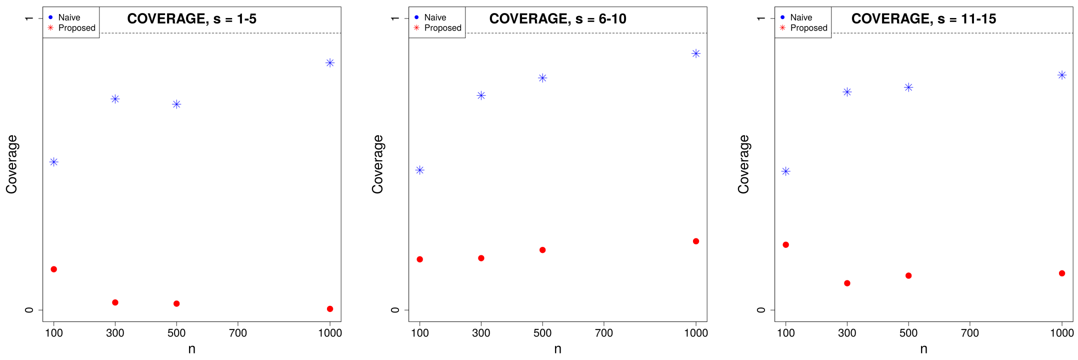

Figure 6: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting B, using gradient boosted trees to estimate (1) and (2). We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

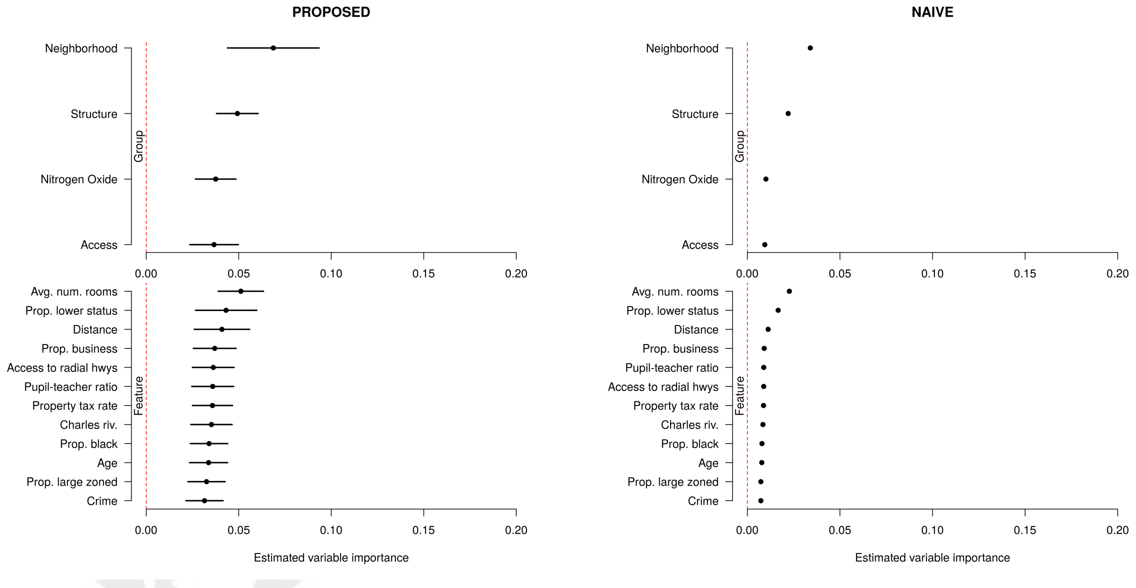

Figure 7: Estimates from the Boston housing project study, for the proposed and naive estimators of the standardized variable importance parameter, on left and right respectively. We estimate (1) and (2) using the Super Learner with the elastic net, generalized additive models, gradient boosted trees, and random forests in its library.

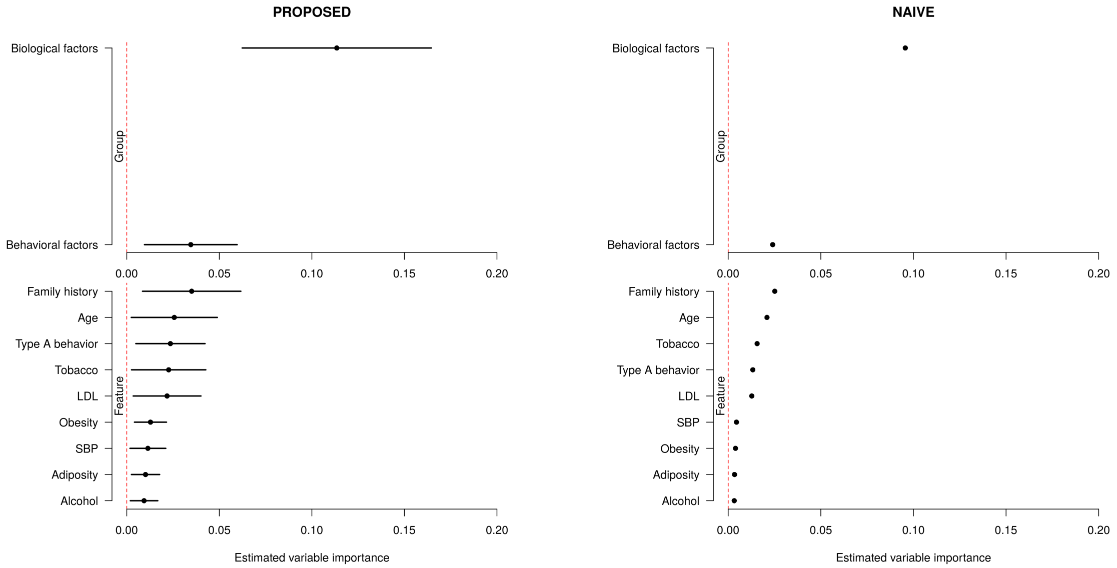

Figure 8: Estimates from the South African heart disease study, for the proposed and naive estimators of the standardized variable importance parameter, on left and right respectively. We estimate (1) and (2) using the Super Learner with the elastic net, generalized additive models, gradient boosted trees, and random forests in its library.

Supplementary material

6.1 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 2.1. 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

$$\frac{\partial}{\partial \epsilon} \Phi_s(P_\epsilon)\bigg|_{\epsilon=0} = \iint \frac{\partial}{\partial \epsilon}\theta_{s,\epsilon}(x)\bigg|_{\epsilon=0} p(x,y)\nu(dx,dy) + \iint \theta_s(x)h(x,y)p(x,y)\nu(dx,dy)\ , \tag{10}$$

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 (10), this allows us to write that

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

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 2.2. 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:

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

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

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 2.3. This is a direct application of Lemma 19.24 of van der Vaart ((2000)). □

Proof of Theorem 2.4. 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 normal variate with variance $var_{P_0}\{D^*_{P_0,s}(O)\} = P_0 D^{*2}_{P_0,s}$.

□

Proof of Theorem 2.5. 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 $\Psi_s(P_{0,g_X}) = \Psi_s(P_0)$. □

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:

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

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 2.

In the second setting (setting B), the covariate distribution was modified to include clustering. Specifically, we generated $(X_1, X_2, \dots, X_{15}) \sim MVN_{15}(\mu, \Sigma)$, 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 Table 2. 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 80 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 9–11. 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 is similar to the variance of the naive estimator (Figure 11).

Table 2: Approximate values of ψ_0 for each simulation setting and group considered for effect size.

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

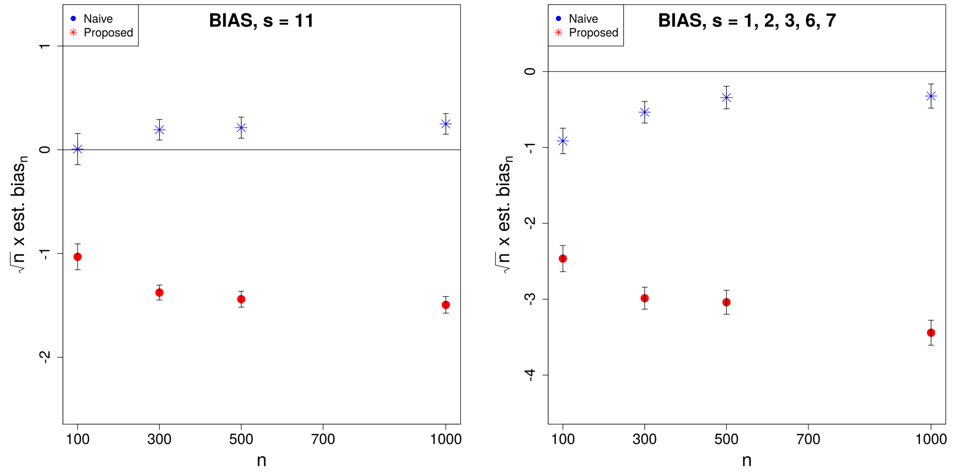

Figure 9: Empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

The results from setting B are a bit different (Figures 12–14). 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.

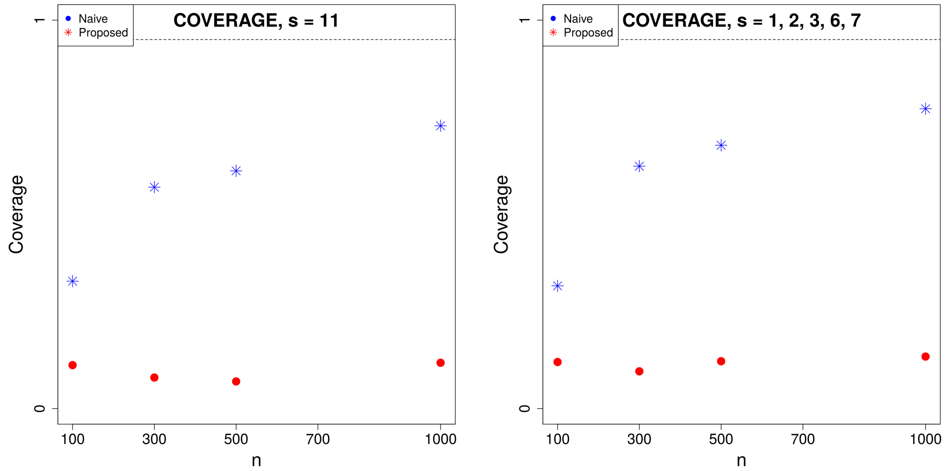

Figure 10: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

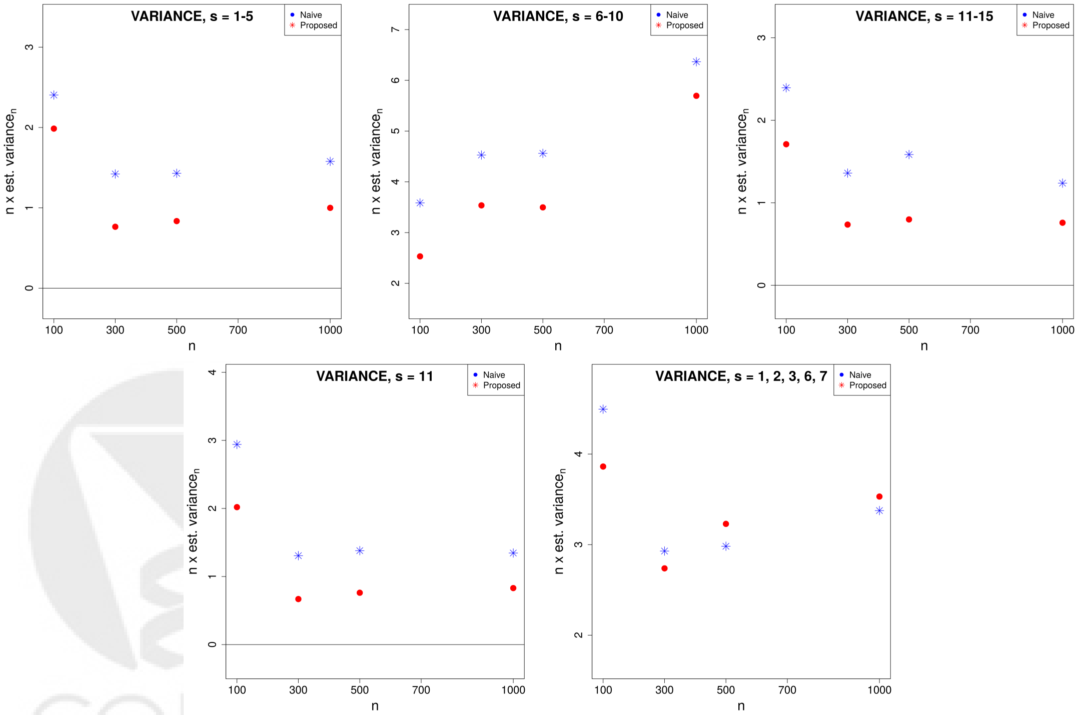

Figure 11: Empirical variance for the proposed and naive estimators scaled by n vs n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 2 and the main manuscript. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

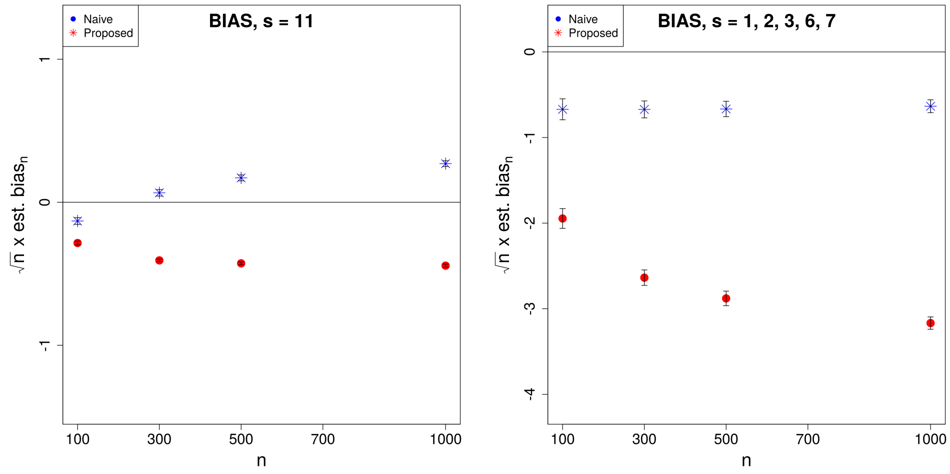

Figure 12: Empirical bias for the proposed and naive estimators scaled by $\sqrt{n}$ vs n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

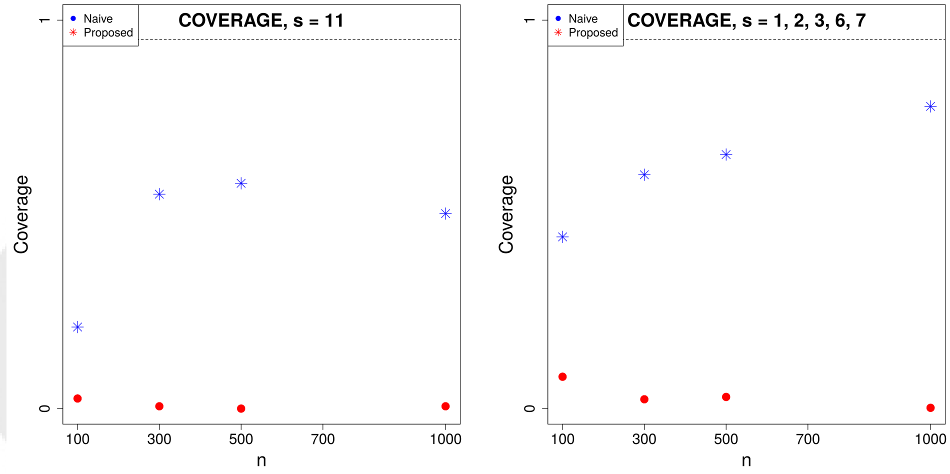

Figure 13: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators vs n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 2. Red circles denote the naive estimator, and blue stars denote the proposed estimator.

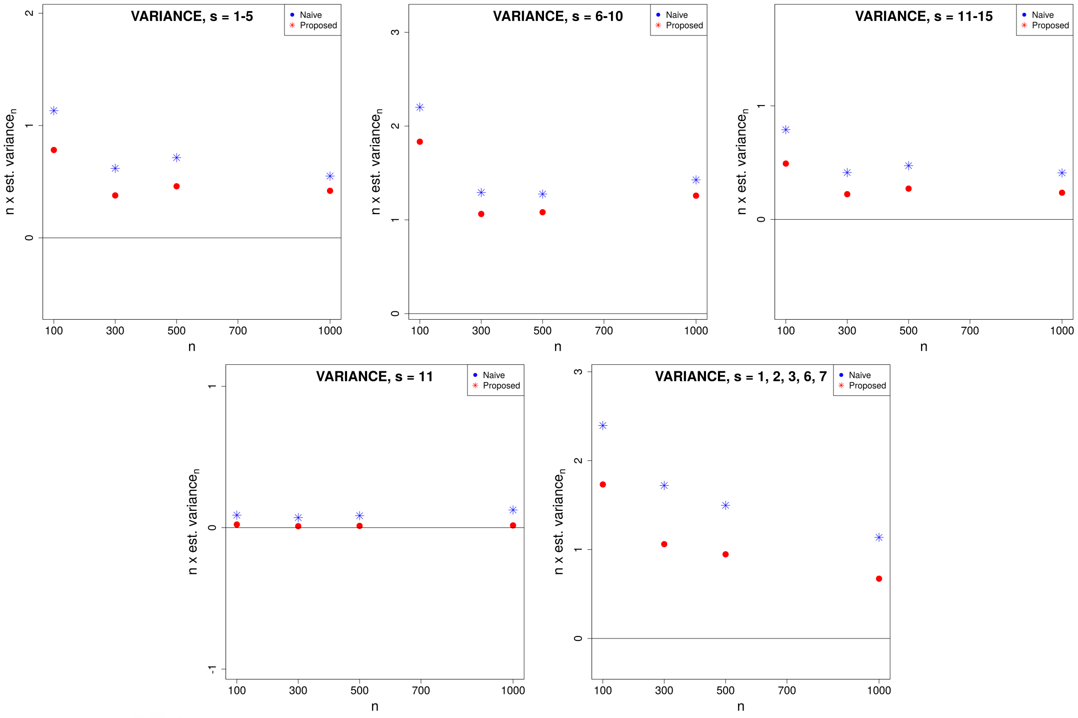

Figure 14: Empirical variance for the proposed and naive estimators scaled by n vs n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 2 and the main manuscript. Red circles denote the naive estimator, and blue stars denote the proposed estimator.