Supporting Information for "Nonparametric variable importance assessment using machine learning techniques"

Brian D. Williamson*†, Peter B. Gilbert‡†, Marco Carone† and Noah Simon†

1 Proof of lemma and theorem

The following proofs rely on a study of the statistical functionals

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

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 for ϵ at $\epsilon = 0$. Every regular parametric submodel centered at P and with score h 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 $\varphi_{P,s} \in L_2(P)$ with $P\varphi_{P,s} = 0$, if we show that for any $P \in \mathcal{M}$

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

we will have established that $\Theta_s(P)$ is pathwise differentiable with respect to $\mathcal{M}$ at P with efficient influence function $\varphi_{P,s}$.

The evaluation of Θ_s on the distribution P_ϵ corresponding to p_ϵ equals

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

*brianw26@uw.edu
†Department of Biostatistics, University of Washington
‡Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center

$$= \iint \alpha_{P,s,\epsilon}(x)p(x,y)\nu(dx,dy) + \epsilon \iint \alpha_{P,s,\epsilon}(x)h(x,y)p(x,y)\nu(dx,dy) \ ,$$

where $\alpha_{P,s,\epsilon}(x) := \{\mu_{P_\epsilon,s}(x) - \mu_{P_\epsilon}(x)\}^2$, and so,

$$\frac{\partial}{\partial\epsilon}\Theta_s(P_\epsilon)\Big|_{\epsilon=0} = \iint \frac{\partial}{\partial\epsilon}\alpha_{P,s,\epsilon}(x)\Big|_{\epsilon=0} p(x,y)\nu(dx,dy) + \iint \alpha_{P,s}(x)h(x,y)p(x,y)\nu(dx,dy) \ ,$$

where $\alpha_{P,s} = \alpha_{P,s,\epsilon}|_{\epsilon=0}$. With a slight abuse of notation, we can write $\alpha_{P,s,\epsilon}(x)$ as

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

and so, we find that $\frac{\partial}{\partial\epsilon}\alpha_{P,s,\epsilon}(x)\big|_{\epsilon=0}$ equals

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

This allows us to write that

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

Because Ψ_s is the ratio of two parameters, namely Θ_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 $z \mapsto \tilde{\varphi}_P(z) := \{y - E_P(Y)\}^2 - var_P(Y)$. It follows then that the nonparametric efficient influence function of Ψ_s at P equals

$$z \mapsto \varphi^*_{P,s}(z) \ = \ \frac{\varphi_{P,s}(z)var_P(Y) - \tilde{\varphi}_P(z)\Theta_s(P)}{var^2_P(Y)} \ ,$$

which simplifies algebraically to the form provided in the Lemma.

Proof of Theorem 1.

It is straightforward to verify that $\Theta_s(P) - \Theta_s(P_0) = -\int \varphi_{P,s}(z)dP_0(z) + R_s(P,P_0)$ with

$$\begin{aligned}R_s(P,P_0) &= P_0(\mu_P - \mu_{P,s})^2 - P_0(\mu_0 - \mu_{0,s})^2 + 2P_0\{(\mu_P - \mu_{P,s})(\mu_0 - \mu_P)\} \\ &= P_0\{(\mu_{0,s} - \mu_{P,s})^2 - (\mu_0 - \mu_P)^2\} \ .\end{aligned}$$

This directly implies that $R_s(\widehat{P}_n, P_0) = o_P(n^{-1/2})$ provided $\hat{\mu} - \mu_0$ and $\hat{\mu}_s - \mu_{0,s}$ both $o_P(n^{-1/4})$ in $L_2(P_0)$ norm, that is, under condition (A1). Additionally, under conditions (A1)–(A3), we have that $H_{s,n}(\widehat{P}, P_0) = o_P(n^{-1/2})$ in view of Lemma 19.24 of van der Vaart (2000), since $\int \{\varphi_{\widehat{P},s}(z) - \varphi_{P_0,s}(z)\}^2 dP_0(z)$ tends to zero in probability as a consequence of condition (A1) provided condition (A3) holds. As such, since we may write that

$$\hat{\theta}_{n,s} - \theta_{0,s} = \Theta_s(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \varphi_{\widehat{P},s}(Z_i) - \Theta_s(P_0) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{P_0,s}(Z_i) + R_s(\widehat{P}, P_0) + H_{s,n}(\widehat{P}, P_0)$$

in view of linearization (7) in the main manuscript, it follows that $\hat{\theta}_{n,s} - \theta_{0,s} = \frac{1}{n}\sum_{i=1}^{n} \varphi_{P_0,s}(Z_i) + o_P(n^{-1/2})$. Additionally, it is easy to verify that $var_{\mathbb{P}_n}(Y) - var_{P_0}(Y) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}_{P_0}(Z_i) + o_P(n^{-1/2})$, where $\tilde{\varphi}_{P_0} : z \mapsto \{y - E_{P_0}(Y)\}^2 - var_{P_0}(Y)$. By the delta method, it follows then that

$$\begin{aligned}
\hat{\psi}_{n,s} - \psi_{0,s} &= \frac{\hat{\theta}_{n,s}}{var_{\mathbb{P}_n}(Y)} - \frac{\theta_{0,s}}{var_{P_0}(Y)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{var_{P_0}(Y)\varphi_{P_0,s}(Z_i) - \theta_{0,s}\tilde{\varphi}_{P_0}(Z_i)}{var_{P_0}(Y)^2} \right\} + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \varphi^*_{P_0,s}(Z_i) + o_P(n^{-1/2}) \ .
\end{aligned}$$

In other words, the proposed estimator $\hat{\psi}_{n,s}$ is an asymptotically linear estimator of $\psi_{0,s}$ with influence function $\varphi^*_{P_0,s}$. 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 $var_{P_0}\{\varphi^*_{P_0,s}(Z)\} = \int \{\varphi^*_{P_0,s}(z)\}^2 dP_0(z)$.

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 2. *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, particularly for large 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, the point estimator will then also enjoy finite-sample invariance.

Proof of Theorem 2

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 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 define the transformed covariate vector $X^* := g_X(X) = (g_1(X_1), g_2(X_2), \ldots, g_p(X_p))$. 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

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

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 $\Psi_s(P_{0,g_X}) = \Psi_s(P_0)$.

3 Cross-fitted estimation procedure

In Section 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 (A2). Theoretical details underlying the behavior of this cross-fitted procedure are provided in Williamson et al. (2020). We now present simulations results for the cross-fitted version of the one-step debiasing procedure involved in our proposed estimator.

3.1 Numerical experiments on a low-dimensional feature vector

We performed an experiment to provide empirical evidence that cross-fitting 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-fitted estimate of the empirical risk based on the squared error

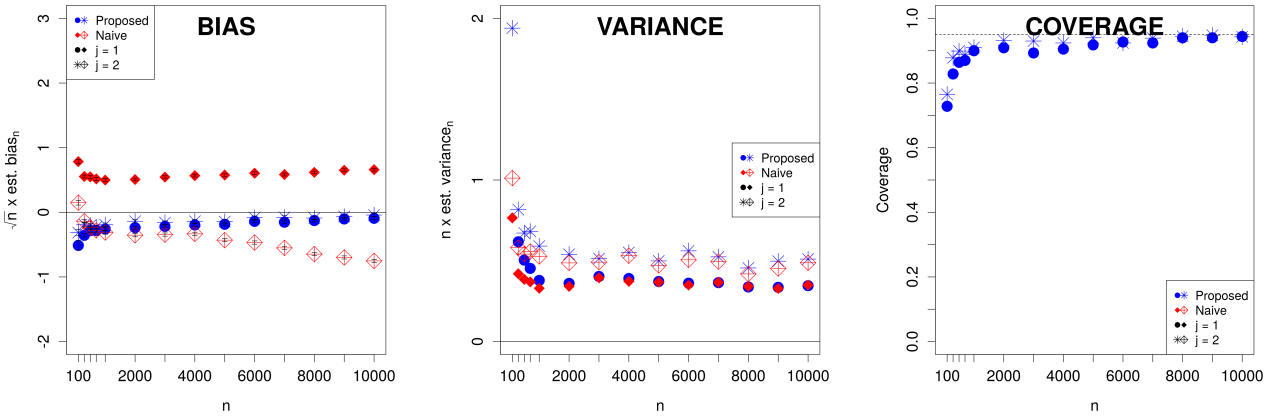

Figure 1: Empirical bias scaled by $n^{1/2}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive cross-fitted estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to 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.

loss function. To obtain the naive and corrected cross-fitted estimators, we performed an outer layer of five-fold cross validation according to the specification above. Confidence intervals based on the corrected cross-fitted estimator were computed using the cross-fitted 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 of the naive estimator should yield poor coverage of such intervals.

Figure 1 displays the results under this alternative hypothesis. In this case, we see a smaller bias of both the naive and the corrected cross-fitted estimators than we saw in the simulations in the main manuscript — this is likely due to the fact that we used locally-linear loess smoothing rather than locally-constant loess smoothing. However, we still see that the naive cross-fitted estimator does not have bias going to zero faster than $n^{-1/2}$, highlighting that the debiasing step is still necessary, even if sample splitting is used in estimation. The variance of the proposed corrected cross-fitted estimator is similar to that of the naive cross-fitted 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-fitted estimator quickly approach the nominal level of 95%.

To study the behavior of our procedure when one feature has null importance, we generated data according to the follow specification:

$$X_1, X_2 \overset{iid}{\sim} \text{Uniform}(-1, 1) \quad \text{and} \quad \epsilon \sim N(0, 1) \text{ independent of } (X_1, X_2)$$

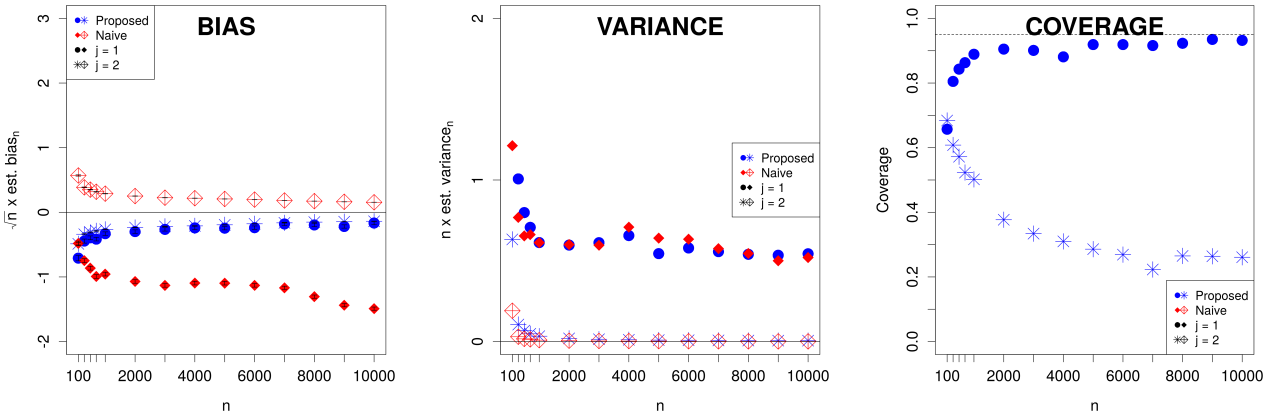

Figure 2: Empirical bias scaled by $n^{1/2}$, empirical variance scaled by n with Monte Carlo error bars, and empirical coverage of nominal 95% confidence intervals for the proposed and naive cross-fitted estimators, for $j = 1$ and 2, using loess smoothing with spans selected by cross-validation to 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.

$$Y = \tfrac{25}{9} X_1^2 + \epsilon \ .$$

In this case, X_1 and X_2 are the non-null and null features, respectively. We used the same estimation procedure as above, and computed the same summaries. Figure 2 displays the results of this experiment. Here, we see that there is much more residual bias in the corrected cross-fitted estimator for the non-null feature. The naive cross-fitted estimator again does not have bias that goes to zero faster than $n^{-1/2}$, highlighting once more 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-fitted estimator, a phenomenon that could be of interest in future research.

Each of these experiments may be reproduced using code available on GitHub.

3.2 Additional numerical experiments for moderate-dimensional feature vector

We considered settings A and B as described in the main manuscript. We recall that in setting A we generated data according to the following specification:

$$X_1, X_2, \ldots, X_{15} \overset{iid}{\sim} N(0,4) \ \text{ and } \ \epsilon \sim N(0,1) \text{ independent of } (X_1, X_2, \ldots, X_{15})$$

$$Y = I_{(-2,+2)}(X_1) \cdot \lfloor X_1 \rfloor + I_{(-\infty,0]}(X_2) + I_{(0,+\infty)}(X_3) + \left|\tfrac{X_6}{4}\right|^3 + \left|\tfrac{X_7}{4}\right|^5 + \tfrac{7}{3}\cos\left(\tfrac{X_{11}}{2}\right) + \epsilon \ .$$

Table 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

We generated 500 random datasets of size $n \in \{100, 300, 500, 1000\}$, and consider here 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 1. In setting B, the covariate distribution was modified to include clustering. We recall that in that setting 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 block-diagonal with blocks

$$\begin{bmatrix} 1 & 0.15 & 0.15 \\ 0.15 & 1 & 0.15 \\ 0.15 & 0.15 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 0.85 & 0.85 \\ 0.85 & 1 & 0.85 \\ 0.85 & 0.85 & 1 \end{bmatrix}$$

and all other off-diagonal entries equal to zero. 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 for each considered group is also given in Table 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 μ_0 and $\mu_{0,s}$ 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.

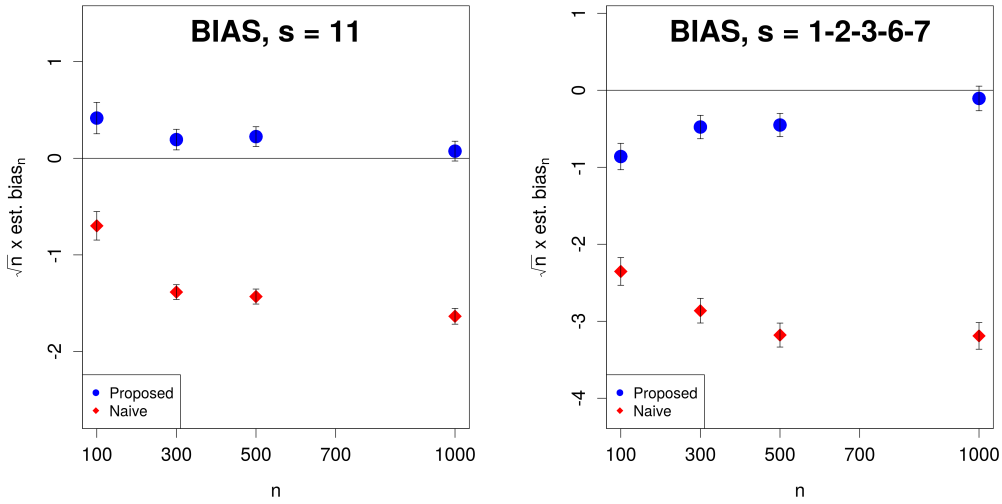

Figure 3: Empirical bias for the proposed and naive estimators scaled by $n^{1/2}$ for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 1. Diamonds denote the naive estimator, and circles denote the proposed estimator. Monte Carlo error bars are displayed vertically.

The results from setting A are presented in Figures 3–5. 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 is similar to the variance of the naive estimator (Figure 5).

The results from setting B are a bit different (Figures 6–8). 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, the coverage of confidence intervals based on the naive estimator do approach zero as n increases. Finally, we see that the variance of the proposed estimator remains similar to the variance of the naive estimator.

These experiments may be reproduced using code available on GitHub.

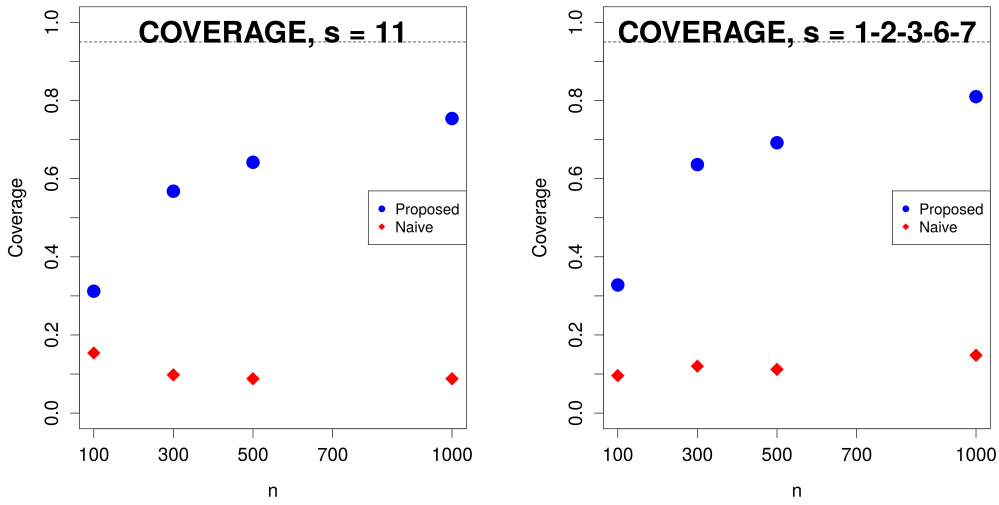

Figure 4: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 1. Diamonds denote the naive estimator, and circles denote the proposed estimator.

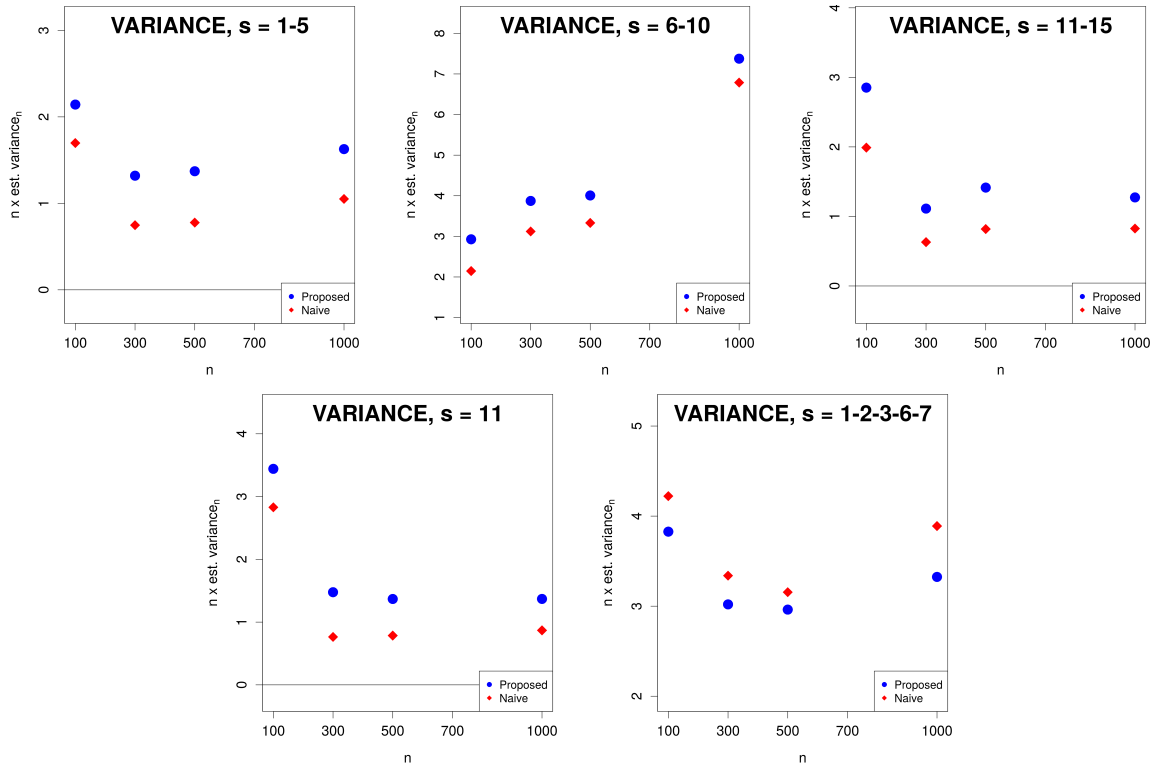

Figure 5: Empirical variance for the proposed and naive estimators scaled by n for setting A, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 1 and the main manuscript. Diamonds denote the naive estimator, and circles denote the proposed estimator.

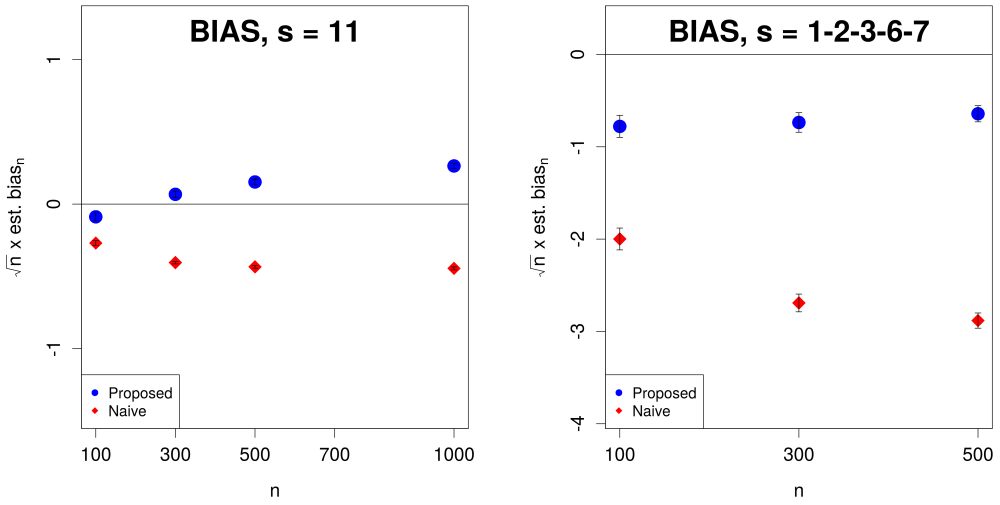

Figure 6: Empirical bias for the proposed and naive estimators scaled by $n^{1/2}$ for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 1. Diamonds denote the naive estimator, and circles denote the proposed estimator.

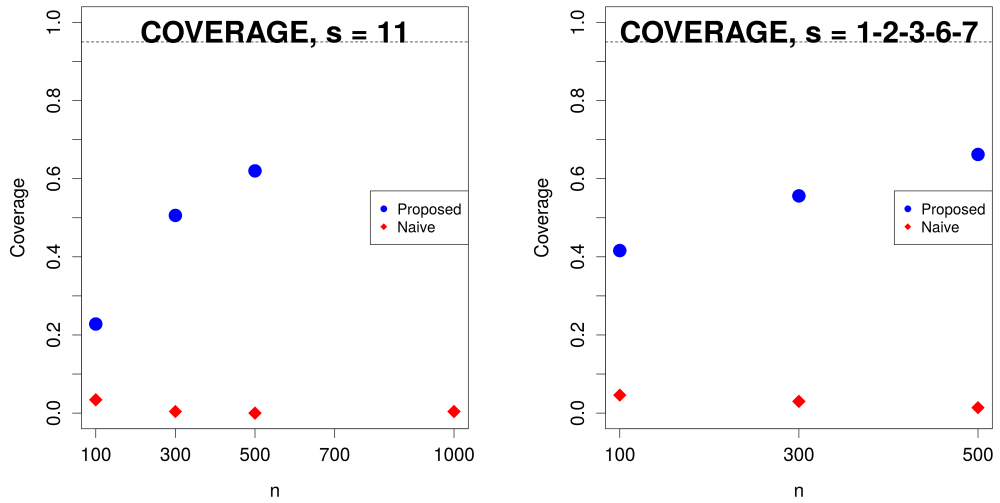

Figure 7: Empirical coverage of nominal 95% confidence intervals for the proposed and naive estimators for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 1. Diamonds denote the naive estimator, and circles denote the proposed estimator.

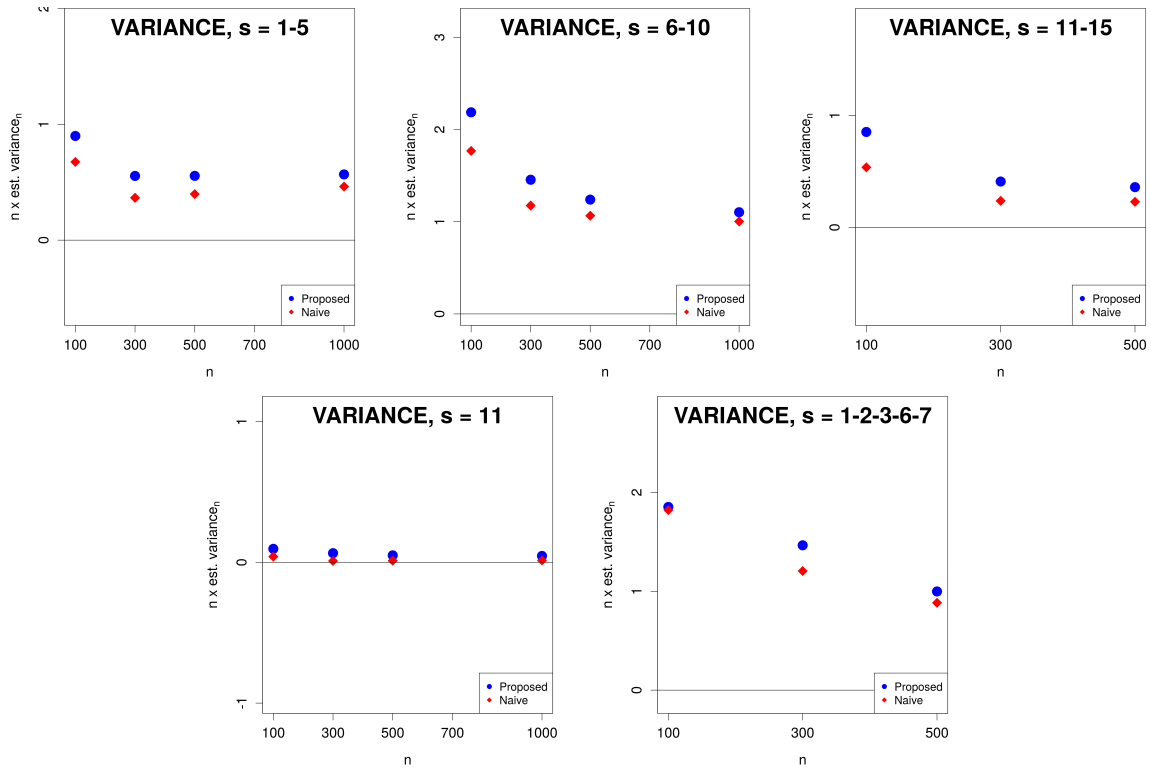

Figure 8: Empirical variance for the proposed and naive estimators scaled by n for setting B, using gradient boosted trees to estimate the conditional means. We consider all s combinations from Table 1 and the main manuscript. Diamonds denote the naive estimator, and circles denote the proposed estimator.

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 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 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 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 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 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.

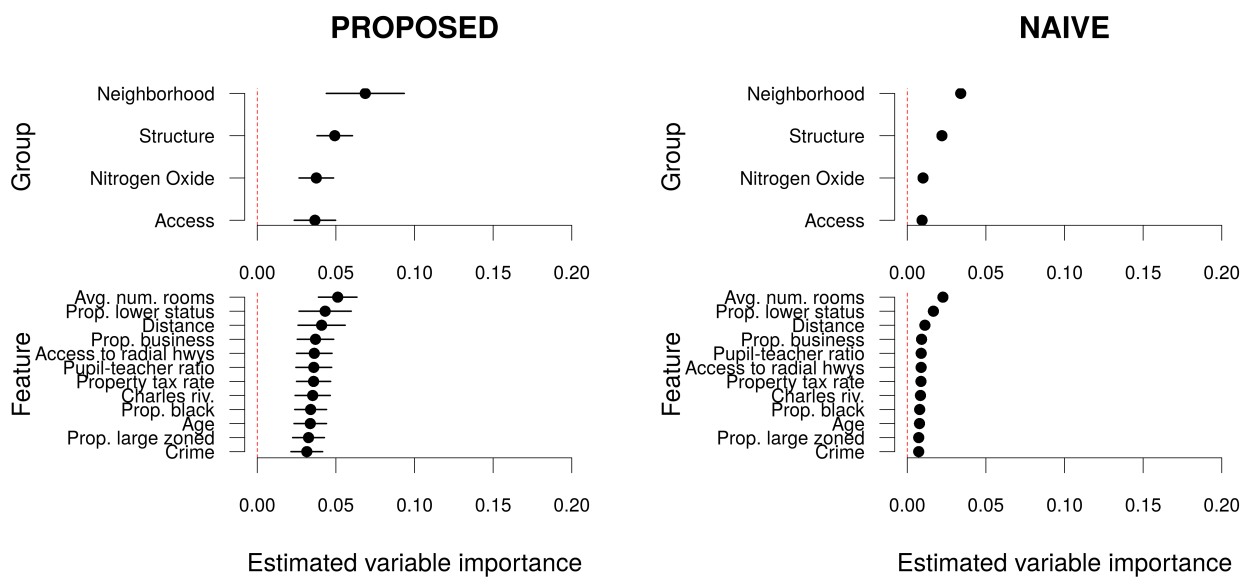

Figure 9: Estimates from the Boston housing project study, for the proposed and naive estimators of the standardized variable importance parameter, on left and right respectively. We estimate μ_0 and $\mu_{0,s}$ using the Super Learner with the elastic net, generalized additive models, gradient boosted trees, and random forests in its library.

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 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 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 $\mu_0(X) - \mu_{0,s}(X)$ in means and

the residual $Y - \mu_{0,s}(X)$. 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 available on GitHub.

References

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

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

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

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

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

Williamson, B., Gilbert, P., Simon, N., and Carone, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. *arXiv* .