Received: 23 May 2020 | Accepted: 26 May 2020

DOI: 10.1111/biom.13391

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

DISCUSSION

# Discussion on "Nonparametric variable importance assessment using machine learning techniques" by Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon

**Min Lu** | **Hemant Ishwaran**

Division of Biostatistics, Department of Public Health Sciences, University of Miami, Florida (Email: m.lu6@umiami.edu)

**Correspondence**
Hemant Ishwaran, Division of Biostatistics, Department of Public Health Sciences, University of Miami, FL 33146.
Email: hishwaran@med.miami.edu

**Funding information**
National Institute of General Medical Sciences, Grant/Award Number: R01 GM125072

## 1 | INTRODUCTION: BREIMAN-CUTLER VARIABLE IMPORTANCE

Williamson *et al.* present a variable index related to $R^2$ but with the property of being free of model specification. We congratulate the authors on this very interesting paper.

Our first point is to draw an important connection between this work and existing work in machine learning. In their introduction, the authors briefly mention the variable importance measure used in Breiman (2001). The authors state that this and related measures used for random forests are intimately tied to the specific estimation technique used, in contrast to the agnostic procedure they propose. However, we will argue there is a much deeper connection than might be apparent.

Denoting PE for prediction error, the general principle underlying Breiman (2001) was to define importance using a prediction error approach. Breiman (2001) defined the variable importance $\hat{I}_{n,s}$ for a set of variables $s$ as the difference in prediction error for the full model compared to the model without $s$,

$$\hat{I}_{n,s} = \text{PE(model without } s) - \text{PE(full model)}. \quad (1)$$

The rationale being that if $s$ contains informative variables, then removing $s$ will increase prediction error and $\hat{I}_{n,s} > 0$. The larger the value, the more evidence of $s$'s importance. On the other hand, if $s$ contains only noisy variables, then removing $s$ may reduce prediction error relative to the full model, or at the very least it will not increase, and thus $\hat{I}_{n,s} \leq 0$.

How one actually calculates prediction error for the $s$-restricted model is crucial and one of the clever aspects of Breiman (2001). This idea was actually developed by Leo Breiman in collaboration with Adele Cutler and is therefore often called Breiman-Cutler variable importance. We will hereafter simply abbreviate this as BC-VIMP (where VIMP stands for variable importance, Ishwaran, 2007; Ishwaran *et al.*, 2008). Let $\hat{\mu}_T(X, \Theta_\upsilon)$ be the $\upsilon = 1, \ldots, V$ tree estimator for the unknown target function $\mu_0(X)$ estimated with respect to a loss function $\ell(Y, \mu)$. Here $\{\Theta_\upsilon\}_{\upsilon=1}^V$ are independent and identically distributed random instructions used to grow each tree. This includes, for example, instructions for splitting the learning data into training data used to grow the tree and out-of-sample data used for testing. The latter indices are denoted by $\mathcal{O}_\upsilon$ for tree $\upsilon$.

BC-VIMP is obtained by averaging over tree VIMP: $\hat{I}_{n,s} = V^{-1} \sum_{\upsilon=1}^V \hat{I}^\upsilon_{n,s}$, where $\hat{I}^\upsilon_{n,s}$ is the vimp for tree $\upsilon$ defined by

$$\hat{I}^\upsilon_{n,s} = \frac{1}{|\mathcal{O}_\upsilon|} \sum_{i \in \mathcal{O}_\upsilon} \ell\left(Y_i, \hat{\mu}_T\left(\tilde{X}_i^{(s)}, \Theta_\upsilon\right)\right) - \frac{1}{|\mathcal{O}_\upsilon|} \sum_{i \in \mathcal{O}_\upsilon} \ell(Y_i, \hat{\mu}_T(X_i, \Theta_\upsilon)). \quad (2)$$

  

The value $\tilde{X}^{(s)}$ represents $X$ when the coordinates $X_s$ have been randomly permuted. In (2), only out-of-sample cases $\mathcal{O}_v$ have their $s$ coordinates randomly permuted. These values $(\tilde{X}_i^{(s)})_{i\in\mathcal{O}_v}$ are run through the tree to estimate PE(model without $s$).

There are two key points in the above calculation that we highlight:

(P1) *Calculating the s-restricted model estimator:* How one calculates the $s$-restricted model estimator is very flexible. What BC-VIMP does is to "noise" up the $s$ coordinates, $X_s$, by permuting them, thereby obtaining a noised up estimator with the intention to mimic a model with $s$ removed. The main advantage of this is that it is computationally fast, therefore it can be used for high-dimensional and big data problems.

(P2) *The same estimator is used to calculate both terms in (2):* The same estimator $\hat{\mu}_T(X,\Theta_v)$ is used to calculate prediction error for both the $s$-restricted and full model. Using the same tree harness eliminates Monte Carlo variability that would occur otherwise.

## 2 | VIMP FOR REGRESSION AND A GENERAL FRAMEWORK

The framework (2) and its over-arching principle (1) are very general and applicable to many settings. For example, extensions include random survival forests (Ishwaran *et al.*, 2008) and competing risk forests (Ishwaran *et al.*, 2014). Approaches based on (1) are also used by many machine learning methods, such as boosting. In fact, there is nothing special about a tree or a random forest that is required to use (1). Thus we can describe a more general framework by replacing the estimator $\hat{\mu}_T$ with any other estimator; let us call this $\hat{\mu}^*$. As before the estimator is constructed from training data and prediction error calculated using out-of-sample data, denoted by $\mathcal{O}_v$, where $v = 1,\ldots,V$. Write $\hat{\mu}_v^*$ for the training data estimator. We now have the following general VIMP framework,

$$\hat{I}_{n,s}^{v} = \frac{1}{|\mathcal{O}_v|}\sum_{i\in\mathcal{O}_v}\ell(Y_i,\hat{\mu}_{s,v}^*(X_i)) - \frac{1}{|\mathcal{O}_v|}\sum_{i\in\mathcal{O}_v}\ell(Y_i,\hat{\mu}_v^*(X_i)). \tag{3}$$

In (3), $\hat{\mu}_{s,v}^*$ denotes the $s$-restricted estimator. The only constraint is it satisfies (P2) requiring it uses the full model estimator $\hat{\mu}_v^*$. One example we have discussed is permutation importance, $\hat{\mu}_{s,v}^* = \hat{\mu}_v^*(\tilde{X}_i^{(s)})$. However, this is not the only method, nor is necessarily the best method that can be used. For example, Lu and Ishwaran (2018) described a general technique for restricted model estimation in parametric models. Rather than permuting $X_s$, they replace the estimated coefficients for $X_s$ in the full model with values of zero.

Now we show how this is related to the authors' work. Let us begin by looking at Equation (10) of their paper. For the comparison, we rescale their estimator by the sample variance and denote the estimator using a tilde. The authors' rescaled equation (10) is

$$\tilde{\psi}_{n,s} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{\mu}_s(X_i)\}^2 - \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{\mu}(X_i)\}^2.$$

Thus $\tilde{\psi}_{n,s}$ is the difference in sum-of-squared values for the restricted versus full model. As this is a measure calculated from the full learning data, it is not directly comparable to (3). However, Williamson et al. also describe a cross-validated version of their estimator in their Algorithm 2. Their rescaled estimator (line 6, Algorithm 2) is

$$\tilde{\psi}_{n,s}^{v} = \frac{1}{|D_v|}\sum_{i\in D_v}\{Y_i - \hat{\mu}_{s,v}(X_i)\}^2 - \frac{1}{|D_v|}\sum_{i\in D_v}\{Y_i - \hat{\mu}_v(X_i)\}^2, \tag{4}$$

where $D_v$ is a test-set fold, $v = 1,\ldots,V$, using $V$-fold estimation. Here $\hat{\mu}_v$ denotes the full model estimator calculated using the $v$th-fold training data and $\hat{\mu}_{s,v}$ is an $s$-restricted model estimator calculated using the same $v$-fold training data.

When specialized to $L_2$-loss, $\ell(Y,\mu) = (Y-\mu)^2$, upon comparing (3) to (4) we see an obvious similarity. It is also clear now that (4) is an example of the general VIMP principle (1). Indeed, the key issue comes down to (P1) in how one defines the restricted model estimator. The approach used by the authors to calculate $\hat{\mu}_{s,v}$ (line 5, Algorithm 2) is to use the full model estimator $\hat{\mu}_v$ and regress this on $X_{-s}$. This is very interesting because what the authors are proposing is essentially a new technique for restricted model estimation for regression. This can be added to the growing list of such techniques (Table 1) and in our opinion is a valuable hidden contribution of the paper. Note that the rationale for using the full model estimator $\hat{\mu}_v$ in calculating both prediction error terms as in (P2) is explained by the authors on page 9 of their paper. They make special mention of this stating they did this because the more obvious method of regressing $Y$ on $X_{-s}$ was found to generally lead to incompatible results.

**TABLE 1** Different methods and their relationship to the general VIMP framework

| Method | Loss, $\ell(Y,\mu)$ | Learner, $\mu^*$ | $\hat{\mu}^*_{s,v}$, $s$-restricted estimator |
|---|---|---|---|
| $\hat{\psi}^{v}_{n,s}$ [a] | $(Y-\mu)^2$ | Any | Regress $\hat{\mu}_v$ on $X_{-s}$ |
| BC-VIMP [b] | Any | Tree | Permute $X_s$ |
| LI-VIMP[c] | Any | Parametric model | Set coefficients for $X_s$ to zero |
| LI-VIMP[c] | Any | Nonparametric model | Permute $X_s$ |
| $\hat{I}^{v}_{n,s}$ [d] | Any | Any | Any |

[a]Williamson, B. et al.'s estimator from Algorithm 2.
[b]Breiman-Cutler VIMP.
[c]Refers to method of Lu and Ishwaran (2018).
[d]Generalized VIMP; see (3).

## 3 | DIMENSIONALITY: SOME EXTENSIONS AND THE BENEFIT OF PREDICTION ERROR

Another point we wish to mention relates to dimensionality and noise. We show that the magnitude of the author's estimator $\hat{\psi}_{n,s}$ can depend on the size of $s$, hence values of $\hat{\psi}_{n,s}$ may not be comparable across models of different dimension. As an attempt to account for this phenomenon, we can define an adjusted $\hat{\psi}_{n,s}$ that is similar to adjusted $R^2$,

$$\hat{\psi}^{\text{adj}}_{n,s} = \left[ \frac{1}{n-(p-|s|)-1} \sum_{i=1}^{n} \{Y_i - \hat{\mu}_s(X_i)\}^2 - \frac{1}{n-p-1} \sum_{i=1}^{n} \{Y_i - \hat{\mu}(X_i)\}^2 \right] \Big/ \text{VAR}_{\text{tot}},$$

where $\text{VAR}_{\text{tot}} = \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2/(n-1)$. However, this adjustment could be too weak because the value of $n-(p-|s|)-1$ could be dominated by $n$. Another option is to add a ratio to the first term, $k := k(p,s)$, and to modify the estimator as follows:

$$\hat{\psi}^{k}_{n,s} = \left[ \frac{k(p,s)}{n-(p-|s|)-1} \sum_{i=1}^{n} \{Y_i - \hat{\mu}_s(X_i)\}^2 - \frac{1}{n-p-1} \sum_{i=1}^{n} \{Y_i - \hat{\mu}(X_i)\}^2 \right] \Big/ \text{VAR}_{\text{tot}}.$$

One example for the ratio could be $k(p,s) = \ln(p/|s|)/\ln(p)$, where when $|s| = 1$, we have $k(p,s) = 1$ and $\hat{\psi}^{k}_{n,s} = \hat{\psi}^{\text{adj}}_{n,s}$.

We use the simulation of setting $A$ in Section 3.3 to display how $\hat{\psi}_{n,s}$, $\hat{\psi}^{\text{adj}}_{n,s}$, and $\hat{\psi}^{k}_{n,s}$ change with expanding feature sizes, $s = \{6\},\{6,7\},\{6,7,8\},\{6,7,8,9\},\{6,7,8,9,10\}$. We follow Algorithm 1 and estimate $\hat{\mu}$ and $\hat{\mu}_s$ using gradient boosted trees as the authors did. A total of 50 independent datasets of size $n = 300$ and $n = 500$ were simulated. Results are displayed in Figure 1. Since $X_6$ and $X_7$ are the only informative variables, we expect the variable importance to increase from $s = \{6\}$ to $s = \{6,7\}$. In other words for an estimator $\hat{\psi}^{*}_{n,s}$, we would expect $\hat{\psi}^{*}_{n,\{6,7\}} > \hat{\psi}^{*}_{n,\{6\}}$ if $\hat{\psi}^{*}_{n,\{7\}} > \hat{\psi}^{*}_{n,\{6\}}$ and $\hat{\psi}^{*}_{n,s}$ measures "average" effect size of features in $s$. Or $\hat{\psi}^{*}_{n,\{6,7\}} > \hat{\psi}^{*}_{n,\{6\}}$ if $\hat{\psi}^{*}_{n,\{7\}} > 0$ and $\hat{\psi}^{*}_{n,s}$ measures the "joint" effect size of features in $s$. However, we would not wish such increases to occur from $s = \{6,7\}$ to $s = \{6,7,8\}$, from $s = \{6,7,8\}$ to $s = \{6,7,8,9\}$, and so forth, since $X_8$, $X_9$, and $X_{10}$ are noise variables.

Figure 1 shows that as noise variables are added and size of $s$ increases, $\hat{\psi}^{k}_{n,s}$ helps reduce inflated values seen for $\hat{\psi}_{n,s}$. Values for $\hat{\psi}^{\text{adj}}_{n,s}$ are similar to $\hat{\psi}_{n,s}$, confirming that its dimensionality adjustment is too weak. Of course, from such small sample sizes, the estimation of $\hat{\psi}_{n,s}$ could be biased, hence variable importance may not be able to measure the average or joint effect sizes in perfect proportion. Therefore, to test this, we have added to Figure 1 another line "BC-VIMP" that are Breiman-Cutler importance values, standardized by $\text{VAR}_{\text{tot}}$. Values were calculated using the `vimp` function from the `randomForestSRC` R-package (Ishwaran and Kogalur, 2020). We can immediately see that BC-VIMP values conform to what we had expected to see: importance values increase from $s = \{6\}$ to $s = \{6,7\}$ and then immediately flatten off. In fact there even seems to be a downward trend for BC-VIMP for the overfit models. The latter occurs as a benefit of using prediction error as prediction error will increase with addition of noise. Finally, we have also adjusted BC-VIMP by $k$ for comparison. As can be seen the adjustment further pushes importance downwards for the overfit models.

## 4 | CONCLUSIONS

In their paper, the authors have introduced not one, but actually two estimators. The first $\hat{\psi}_{n,s}$ (Algorithm 1) is constructed from the full learning dataset. Although we did

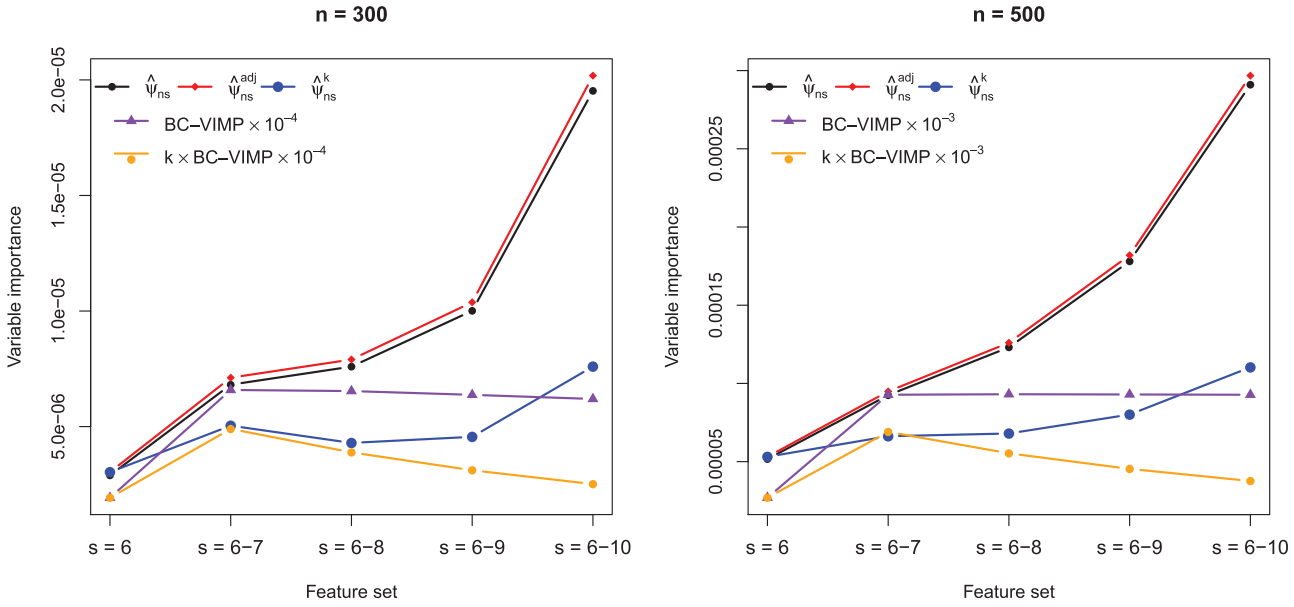


**FIGURE 1** Comparison of variable importance measures with increasing size of feature set $s$ for sample sizes $n = 300$ (left) and $n = 500$ (right). Adjusted variable importance measures $\hat{\psi}^{\text{adj}}_{n,s}$ and $\hat{\psi}^{k}_{n,s}$ are shown in red and blue, respectively, and the unadjusted measure $\hat{\psi}_{n,s}$ is marked in black color. The datasets are generated according to the simulation of setting $A$ in Section 3.3, where only $X_6$ and $X_7$ are informative variables in all the chosen feature sets. Values of $\hat{\psi}_{n,s}$ and $\hat{\psi}^{\text{adj}}_{n,s}$ are similar and both become inflated as noise variables are added. On the other hand, $\hat{\psi}^{k}_{n,s}$ performs much better due to its heavy dimensionality penalty. Also included in figure are BC-VIMP values displayed in purple. By being based on prediction error, BC-VIMP automatically adjusts for dimensionality and does not overfit with addition of noise and performs correctly. Orange lines are BC-VIMP multiplied by dimensionality parameter $k$. This further pushes VIMP values downwards for overfit models. Finally, note BC-VIMP has been scaled by $10^{-4}$ and $10^{-3}$ for left and right figures. Thus for model $s = \{6, 7\}$, the value is approximately 7% for $n = 300$ and 10% for $n = 500$. This is very interpretable and is stating that the model with the two nonzero variables explains a relatively high fraction of the variance over test data. This type of interpretation is not possible with estimators like $\hat{\psi}_{n,s}$ that are not prediction error based. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

not comment much about this version in our discussion, one major advantage of $\hat{\psi}_{n,s}$ is that it is far more amenable to theoretical analysis than a prediction error based estimator, such as their second estimator, $\hat{\psi}^{v}_{n,s}$. In fact this is what the authors have done. Using empirical processes and semiparametric theory, they have developed a comprehensive analysis for $\hat{\psi}_{n,s}$ that provides justification for the procedure and identifies its asymptotic limiting distribution for certain cases, and we commend the authors on doing so. Regarding the limiting distribution of Theorem 1, the authors use this for developing confidence intervals using a plug-in estimator. We would like to mention another technique that has been used for constructing confidence intervals for VIMP based on subsampling (Ishwaran and Lu, 2019). This method is applicable to many settings including survival, regression, and classification problems. By being based on subsampling theory, the conditions needed are different than those used here. In particular, condition (A1) requires a rate condition for the underlying estimation technique, whereas for the subsampling estimator, one only requires the existence of a limiting distribution. This could be useful as proving (A1) may be difficult for machine learning methods. From a more practical standpoint, subsampling is highly computationally efficient and therefore opens up applications to high dimensional and big data scenarios.

This brings us to the authors second estimator $\hat{\psi}^{v}_{n,s}$ (Algorithm 2), which we have discussed in more detail. This estimator unlike the full data estimator is prediction error based and as we have argued is an example of the general principle (1) described by Breiman (2001). As we commented, the authors are proposing a new technique for restricted model estimation (Table 1). To calculate the first term in (1), they regress the full estimator on $X_{-s}$. This is interesting and we wish they had provided empirical results of its effectiveness. In fact, we believe this estimator will prove superior to $\hat{\psi}_{n,s}$. We illustrated already some problems with the latter. Other issues to mention are in the results of real data applications: all values of $\hat{\psi}_{n,s}$ appear positive and confidence intervals do not cover zero, which could mislead one to believe $\hat{\psi}_{n,s}$ can only rank variables, but is unable to separate noise variables. We believe these and other issues will be remedied by a prediction error based VIMP.

In closing, we congratulate the authors for their contributions to the area of variable importance. We also thank

the editor(s) for giving us an opportunity for sharing our insights on this work. There are many interesting ideas and potentially important theoretical work that may find inspiration from this article and its discussion.

## ORCID

*Hemant Ishwaran* https://orcid.org/0000-0003-2758-9647

## REFERENCES

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

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

Ishwaran, H., Gerds, T.A., Kogalur, U.B., Moore, R.D., Gange, S.J. and Lau, B.M. (2014) Random survival forests for competing risks. *Biostatistics*, 15(4), 757–773.

Ishwaran, H. and Kogalur, U.B. (2020) Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.9.3.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008) Random survival forests. *The Annals of Applied Statistics*, 2(3), 841–860.

Ishwaran, H. and Lu, M. (2019) Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, 38(4), 558–582.

Lu, M. and Ishwaran, H. (2018) A prediction-based alternative to P values in regression models. *The Journal of Thoracic and Cardiovascular Surgery*, 155(3), 1130–1136.

**How to cite this article:** Lu M, Ishwaran H. Discussion on "Nonparametric variable importance assessment using machine learning techniques" by Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. *Biometrics*. 2021;77:23–27. https://doi.org/10.1111/biom.13391