Nonparametric variable importance assessment using flexible estimation procedures

Brian D. Williamson

UW Biostatistics Colloquium

8 November 2018

Motivation

The Antibody Mediated Prevention trials study prevention efficacy of VRC01, a broadly neutralizing antibody, against HIV-1 infection.

Key question: how does prevention efficacy of VRC01 vary with genotypic characteristics of the HIV-1 virus?

Motivation

The Antibody Mediated Prevention trials study prevention efficacy of VRC01, a broadly neutralizing antibody, against HIV-1 infection.

Key question: how does prevention efficacy of VRC01 vary with genotypic characteristics of the HIV-1 virus?

Potential issues:

- Many ways to define genotype based on amino acid sequence
  - Low statistical power after adjusting for multiple comparisons
  - Typically pre-specify small set of features

Motivation

The Antibody Mediated Prevention trials study prevention efficacy of VRC01, a broadly neutralizing antibody, against HIV-1 infection.

Key question: how does prevention efficacy of VRC01 vary with genotypic characteristics of the HIV-1 virus?

Potential issues:

- Many ways to define genotype based on amino acid sequence
  - Low statistical power after adjusting for multiple comparisons
  - Typically pre-specify small set of features
- Using machine learning-based methods in prediction
  - What information do we gain about the population of interest?
  - Formal statistical inference often difficult

Motivation

Variable importance may help to address these issues:

- Pre-existing data: identify important features and groups
  - maintain statistical power, while
  - making fuller use of the data at hand

Motivation

Variable importance may help to address these issues:

- Pre-existing data: identify important features and groups
  - maintain statistical power, while
  - making fuller use of the data at hand
- May obtain valid statistical inference on the importance
  - necessary for decision making
  - understand the population-level interplay between variables

Motivation

What is the importance of different amino acid sequence features for predicting the neutralization sensitivity of HIV-1 to VRC01?

X_1 = CD4 binding site

X_2 = VRC01 binding footprint

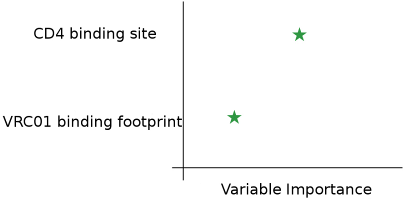
Y = Neutralization sensitivity

Motivation

What is the importance of different amino acid sequence features for predicting the neutralization sensitivity of HIV-1 to VRC01?

We need:

- A broadly-relevant definition of variable importance

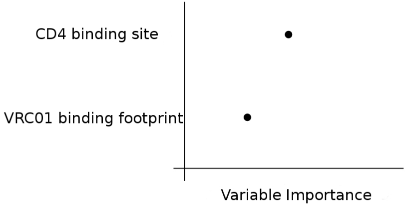


Motivation

What is the importance of different amino acid sequence features for predicting the neutralization sensitivity of HIV-1 to VRC01?

We need:

- A broadly-relevant definition of variable importance
- A method that:
    - Estimates variable importance

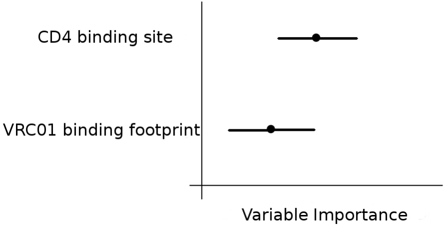


Motivation

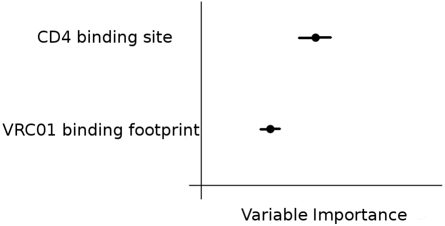
What is the importance of different amino acid sequence features for predicting the neutralization sensitivity of HIV-1 to VRC01?

We need:

- A broadly-relevant definition of variable importance
- A method that:
  - Estimates variable importance
  - Provides valid uncertainty assessment for our estimates



vs.



Motivation

What is the importance of different amino acid sequence features for predicting the neutralization sensitivity of HIV-1 to VRC01?

We need:

- A broadly-relevant definition of variable importance
- A method that:
  - Estimates variable importance
  - Provides valid uncertainty assessment for our estimates
  - May be used with flexible estimation procedures

Variable importance: the data

Consider data $O_1, \ldots, O_n$ drawn from an unknown distribution P_0:

Variable importance: the data

Consider data $O_1, \ldots, O_n$ drawn from an unknown distribution P_0:

- $O_i := (X_i, Y_i)$;

Variable importance: the data

Consider data $O_1, \ldots, O_n$ drawn from an unknown distribution P_0:

- $O_i := (X_i, Y_i)$;
- $X_i \in \mathbb{R}^p$ is a vector of covariates, and

Variable importance: the data

Consider data $O_1, \ldots, O_n$ drawn from an unknown distribution P_0:

- $O_i := (X_i, Y_i)$;
- $X_i \in \mathbb{R}^p$ is a vector of covariates, and
- $Y_i \in \mathbb{R}$ is the outcome of interest.

Variable importance: the data

Consider data $O_1, \ldots, O_n$ drawn from an unknown distribution P_0:

- $O_i := (X_i, Y_i)$;
- $X_i \in \mathbb{R}^p$ is a vector of covariates, and
- $Y_i \in \mathbb{R}$ is the outcome of interest.

Our goal: to describe the importance of some subset of the covariates for predicting the outcome in the population.

Key object: the conditional mean, $E_{P_0}(Y \mid X = x)$.

Variable importance: linear regression

Objective: estimate the importance of X_s, $s \subseteq \{1, \ldots, p\}$.

How is variable importance typically measured in linear regression?

Variable importance: linear regression

Objective: estimate the importance of X_s, $s \subseteq \{1, \dots, p\}$.

How is variable importance typically measured in linear regression?

1. Fit a linear regression of Y on $X \to \hat{\mu}(X)$

Variable importance: linear regression

Objective: estimate the importance of X_s, $s \subseteq \{1, \ldots, p\}$.

How is variable importance typically measured in linear regression?

1. Fit a linear regression of Y on $X \to \hat{\mu}(X)$
2. Fit a linear regression of Y on $X_{(-s)} \to \hat{\mu}_{-s}(X)$

Variable importance: linear regression

Objective: estimate the importance of X_s, $s \subseteq \{1, \ldots, p\}$.

How is variable importance typically measured in linear regression?

1. Fit a linear regression of Y on $X \to \hat{\mu}(X)$
2. Fit a linear regression of Y on $X_{(-s)} \to \hat{\mu}_{-s}(X)$
3. Compare the fitted values $[\hat{\mu}(X_i), \hat{\mu}_{-s}(X_i)]$ of each regression

Variable importance: linear regression

Objective: estimate the importance of X_s, $s \subseteq \{1, \ldots, p\}$.

How is variable importance typically measured in linear regression?

1. Fit a linear regression of Y on $X \rightarrow \hat{\mu}(X)$
2. Fit a linear regression of Y on $X_{(-s)} \rightarrow \hat{\mu}_{-s}(X)$
3. Compare the fitted values $[\hat{\mu}(X_i), \hat{\mu}_{-s}(X_i)]$ of each regression

Both sets of fitted values estimate a conditional mean!

Variable importance: linear regression

Objective: estimate the importance of X_s, $s \subseteq \{1, \ldots, p\}$.

How is variable importance typically measured in linear regression?

1. Fit a linear regression of Y on $X \to \hat{\mu}(X)$
2. Fit a linear regression of Y on $X_{(-s)} \to \hat{\mu}_{-s}(X)$
3. Compare the fitted values $[\hat{\mu}(X_i), \hat{\mu}_{-s}(X_i)]$ of each regression

Both sets of fitted values estimate a conditional mean!

Many ways to compare fitted values, including:

- Difference in R^2
- ANOVA decomposition

Variable importance: linear regression

The mean squared error (MSE) of a linear regression function f:

$$MSE(f) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(X_i)\}^2$$

Variable importance: linear regression

The mean squared error (MSE) of a linear regression function f:

$$MSE(f) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(X_i)\}^2$$

Difference in R^2:

$$\left[1 - \frac{MSE(\hat{\mu})}{MSE(\overline{Y}_n)}\right] - \left[1 - \frac{MSE(\hat{\mu}_{-s})}{MSE(\overline{Y}_n)}\right]$$

Variable importance: linear regression

The mean squared error (MSE) of a linear regression function f:

$$MSE(f) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(X_i)\}^2$$

Difference in R^2:

$$\left[1 - \frac{MSE(\hat{\mu})}{MSE(\overline{Y}_n)}\right] - \left[1 - \frac{MSE(\hat{\mu}_{-s})}{MSE(\overline{Y}_n)}\right]$$

ANOVA decomposition:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(X_i) - \hat{\mu}_{-s}(X_i)\}^2}{MSE(\overline{Y}_n)}$$

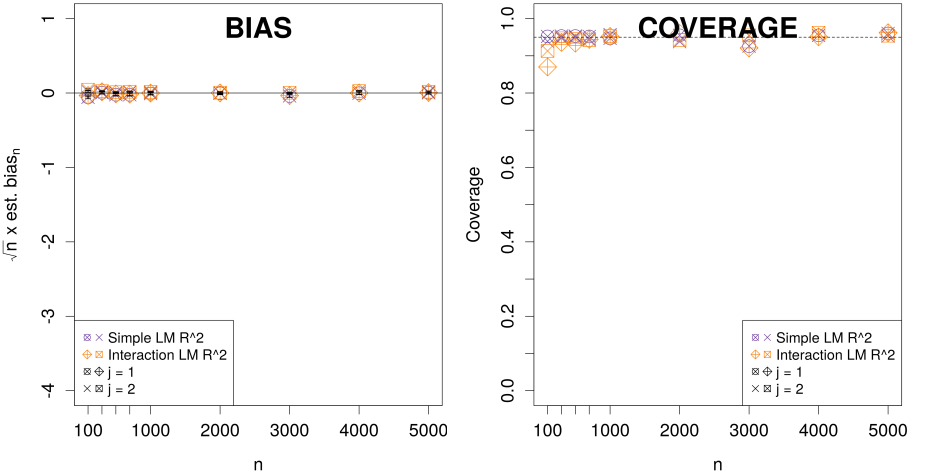
Experiment in a linear model

$X = (X_1, X_2)$ independent, and $Y \mid X = x \sim N(3x_1 + x_2, 1)$.

Estimation procedure:

1. $\hat{\mu}(x) \leftarrow$ Fit linear regression with full X vector
2. $\hat{\mu}_{-s}(x) \leftarrow$ Fit linear regression with either X_1 or X_2 removed
3. Calculate difference in R^2
4. Bootstrap-based confidence intervals

Experiment: results, linear regression



Variable importance: extensions?

When pursuing variable importance more generally:

- what if the truth is a complex linear model?

Variable importance: extensions?

When pursuing variable importance more generally:

- what if the truth is a complex linear model?
- what if the truth is not a linear model at all?

Variable importance: extensions?

When pursuing variable importance more generally:

- what if the truth is a complex linear model?
- what if the truth is not a linear model at all?
- what if your collaborator wants to fit a flexible algorithm?

Variable importance: extensions?

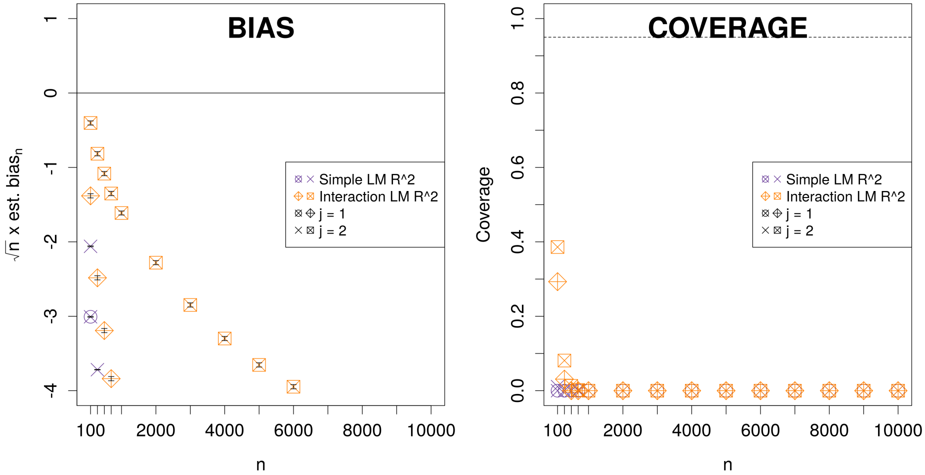
When pursuing variable importance more generally:

- what if the truth is a complex linear model?
- what if the truth is not a linear model at all?
- what if your collaborator wants to fit a flexible algorithm?

Fitting simple linear regression estimators (even including interactions) may not be sufficient!

New experiment:
$X = (X_1, X_2)$ independent, $Y \mid X = x \sim N((x_1 + x_2)^4, 1)$

Experiment (interaction): results, linear regression



Variable importance: flexible estimators?

Issues when fitting a flexible estimator (e.g., random forests):

- bias-variance tradeoff for conditional mean

Variable importance: flexible estimators?

Issues when fitting a flexible estimator (e.g., random forests):

- bias-variance tradeoff for conditional mean
- algorithm-specific importance may not be comparable

Variable importance: flexible estimators?

Issues when fitting a flexible estimator (e.g., random forests):

- bias-variance tradeoff for conditional mean
- algorithm-specific importance may not be comparable
- inference on this importance difficult

Variable importance: flexible estimators?

Issues when fitting a flexible estimator (e.g., random forests):

- bias-variance tradeoff for conditional mean
- algorithm-specific importance may not be comparable
- inference on this importance difficult

To handle these issues, we typically:

1. specify a population-based importance measure

Variable importance: flexible estimators?

Issues when fitting a flexible estimator (e.g., random forests):

- bias-variance tradeoff for conditional mean
- algorithm-specific importance may not be comparable
- inference on this importance difficult

To handle these issues, we typically:

1. specify a population-based importance measure
2. correct for excess bias inhereted from flexible estimator

Variable importance: population

How might we measure importance if we could predict perfectly?

Variable importance: population

How might we measure importance if we could predict perfectly?

Oracle prediction functions:

- $\mu^*(x) := E_{P_0}(Y \mid X = x)$
- $\mu^*_{-s}(x) := E_{P_0}(Y \mid X_{(-s)} = x)$

Variable importance: population

How might we measure importance if we could predict perfectly?

Oracle prediction functions:

- $\mu^*(x) := E_{P_0}(Y \mid X = x)$
- $\mu^*_{-s}(x) := E_{P_0}(Y \mid X_{(-s)} = x)$

Population importance defined in terms of μ^*, μ^*_{-s}!

Variable importance: population

Both R^2 and ANOVA involve the MSE:

$$MSE_{P_0}(f^*) := E_{P_0}\{Y - f^*(X)\}^2;$$

Variable importance: population

Both R^2 and ANOVA involve the MSE:

$$MSE_{P_0}(f^*) := E_{P_0}\{Y - f^*(X)\}^2;$$

$$R^2_{P_0}(\mu^*) := 1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}$$

Variable importance: population

Both R^2 and ANOVA involve the MSE:

$$MSE_{P_0}(f^*) := E_{P_0}\{Y - f^*(X)\}^2;$$

$$R^2_{P_0}(\mu^*) := 1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}$$

$$\text{ANOVA}_{P_0} \equiv R^2_{P_0}(\mu^*) - R^2_{P_0}(\mu^*_{-s}) := \left[1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}\right] - \left[1 - \frac{MSE_{P_0}(\mu^*_{-s})}{var_{P_0}(Y)}\right]$$

Variable importance: population

Both R^2 and ANOVA involve the MSE:

$$MSE_{P_0}(f^*) := E_{P_0}\{Y - f^*(X)\}^2;$$

$$R^2_{P_0}(\mu^*) := 1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}$$

$$\text{ANOVA}_{P_0} \equiv R^2_{P_0}(\mu^*) - R^2_{P_0}(\mu^*_{-s}) := \left[1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}\right] - \left[1 - \frac{MSE_{P_0}(\mu^*_{-s})}{var_{P_0}(Y)}\right]$$

The MSE is a risk: large values imply poor performance.

Variable importance: population

Both R^2 and ANOVA involve the MSE:

$$MSE_{P_0}(f^*) := E_{P_0}\{Y - f^*(X)\}^2;$$

$$R^2_{P_0}(\mu^*) := 1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}$$

$$\text{ANOVA}_{P_0} \equiv R^2_{P_0}(\mu^*) - R^2_{P_0}(\mu^*_{-s}) := \left[1 - \frac{MSE_{P_0}(\mu^*)}{var_{P_0}(Y)}\right] - \left[1 - \frac{MSE_{P_0}(\mu^*_{-s})}{var_{P_0}(Y)}\right]$$

The MSE is a risk: large values imply poor performance.

Variable importance: the best-case, population comparison of risks!

Experiment in a linear model

$X = (X_1, X_2)$ independent, $Y \mid X = x \sim N(3x_1 + x_2, 1)$.

Estimation procedure:

- $\hat{\mu}(x) \leftarrow$ Fit loess smoother with full X vector
- $\hat{\mu}_{-s}(x) \leftarrow$ Fit loess smoother with either X_1 or X_2 removed
- plug in to calculate difference in R^2, ANOVA
- Influence function-based confidence intervals

Experiment in a linear model

$X = (X_1, X_2)$ independent, $Y \mid X = x \sim N(3x_1 + x_2, 1)$.

Estimation procedure:

- $\hat{\mu}(x) \leftarrow$ Fit loess smoother with full X vector
- $\hat{\mu}_{-s}(x) \leftarrow$ Fit loess smoother with either X_1 or X_2 removed
- plug in to calculate difference in R^2, ANOVA
- Influence function-based confidence intervals

Question: do we need to correct the plug-in estimator?

Experiment in a linear model

$X = (X_1, X_2)$ independent, $Y \mid X = x \sim N(3x_1 + x_2, 1)$.

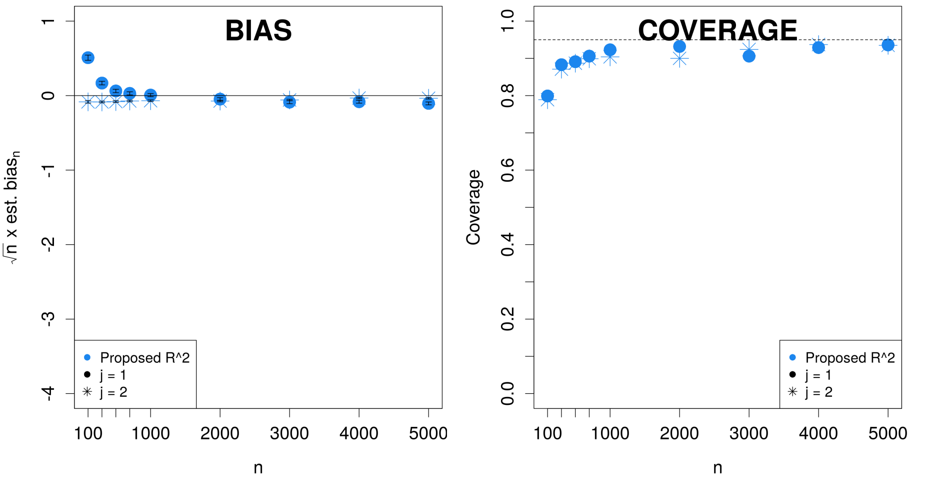
Estimation procedure:

- $\hat{\mu}(x) \leftarrow$ Fit loess smoother with full X vector
- $\hat{\mu}_{-s}(x) \leftarrow$ Fit loess smoother with either X_1 or X_2 removed
- plug in to calculate difference in R^2, ANOVA
- Influence function-based confidence intervals

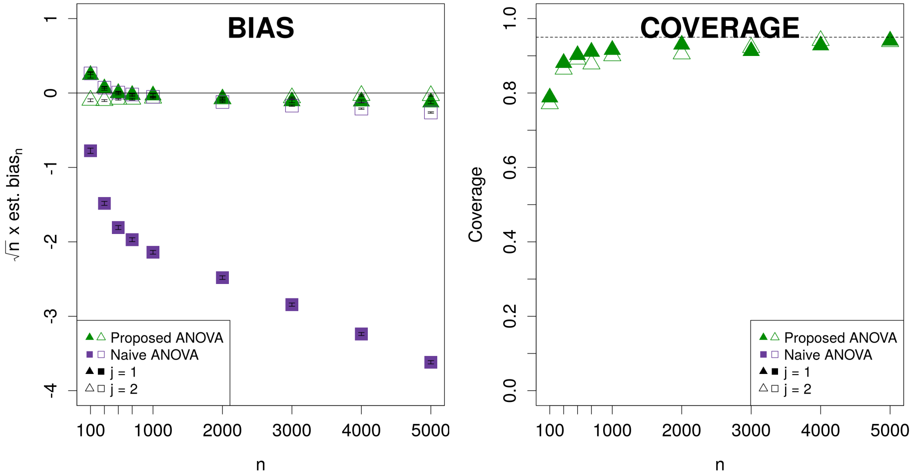
Question: do we need to correct the plug-in estimator?

No, for R^2; Yes, for ANOVA (using the influence function).

Experiment: results, flexible estimators (R^2)



Experiment: results, flexible estimators (ANOVA)



Conclusions

Population variable importance may be thought of as the best-case, population comparison of risks.

Conclusions

Population variable importance may be thought of as the best-case, population comparison of risks.

Asymptotically valid CIs based on plug-in estimators for:

- difference in R^2
- difference in AUC
- cross-entropy (deviance)

even when using flexible estimation techniques are used.

Conclusions

Population variable importance may be thought of as the best-case, population comparison of risks.

Asymptotically valid CIs based on plug-in estimators for:

- difference in R^2
- difference in AUC
- cross-entropy (deviance)

even when using flexible estimation techniques are used.

We also have results in studies with missing data; here, some correction is necessary!