The International Journal of Biostatistics

Volume 2, Issue 1

2006

Article 6

Application of a Variable Importance Measure Method

Merrill D. Birkner, *University of California, Berkeley*
Mark J. van der Laan, *Division of Biostatistics, School of
Public Health, University of California, Berkeley*

Recommended Citation:

Birkner, Merrill D. and van der Laan, Mark J. (2006) "Application of a Variable Importance Measure Method," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 6.

DOI: 10.2202/1557-4679.1013

Application of a Variable Importance Measure Method

Merrill D. Birkner and Mark J. van der Laan

Abstract

Van der Laan (2005) proposed a targeted method used to construct variable importance measures coupled with respective statistical inference. This technique involves determining the importance of a variable in predicting an outcome. This method can be applied as inverse probability of treatment weighted (IPTW) or double robust inverse probability of treatment weighted (DR-IPTW) estimators. The variance and respective p-value of the estimate are calculated by estimating the influence curve. This article applies the Van der Laan (2005) variable importance measures and corresponding inference to HIV-1 sequence data. In this application, the method is targeted at every codon position. In this data application, protease and reverse transcriptase codon positions on the HIV-1 strand are assessed to determine their respective variable importance, with respect to an outcome of viral replication capacity. We estimate the DR-IPTW W-adjusted variable importance measure for a specified set of potential effect modifiers W . In addition, simulations were performed on two separate datasets to examine the DR-IPTW estimator.

KEYWORDS: variable importance, HIV

1 Introduction

In many genomic applications, the prediction of a phenotypic outcome by genetic markers is of biological importance. In particular, variable importance measures can be calculated for individual genetic components when predicting an outcome. Variable importance measures can be used to determine the effect of each variable with respect to an outcome of interest. In this article, we will apply the variable importance measure methodology, proposed by van der Laan (2005), to an HIV-1 sequence dataset, in addition to simulated datasets.

Current literature on prediction provides measures of variable importance by fitting the regression of an outcome on the covariates, and then calculating measures based on this fit. Bootstrap aggregation (random forest, Breiman (1999)) is a method for obtaining a high dimensional fit so that most variables are present in the fit. Contrary to the methods proposed in van der Laan (2005), all of these algorithms are not targeted at an estimate of the variable importance for a particular variable, but instead are targeted at constructing a predictor. For example, if the true regression is a linear combination of multi-way interactions, then every multi-way interaction not including the variable of interest does not affect our definition of variable importance. The van der Laan (2005) proposed methods assume models for the wished type of variable importance, and then use general estimating function methodology (Robins and Rotnitzky, 1992) to determine the class of estimating functions, so that the corresponding estimators are fully targeted at the parameter of interest (i.e., the particular variable importance). van der Laan (2005) also note that current machine learning algorithms do not provide a p -value or confidence interval for a reported measure of variable importance (see e.g., random forest in Breiman (1999)).

In the application presented in this article, the variable importance of each codon and respective significance are assessed with respect to their relationship to viral replication. We will also estimate the DR-IPTW W -adjusted variable importance measure for a specified set of potential effect modifiers W . In Section 5, we will present two simulations to examine the DR-IPTW estimator. These simulations will examine the performance of the estimator under different data generating distributions which corresponds with varying levels of information in the data.

2 Data Application

2.1 Biological Motivation

Sequencing a virus, such as HIV-1, could potentially give further insight into the genotype-phenotype associations of a virus. We focus on phenotype replication ability of the virus. In many cases, genetic mutations on the viral strand are associated with a change in the replication capacity of the virus. This in turn can change the virulence of the virus and/or cause resistance to previously effective antiretroviral drugs. The motivation behind this analysis is focused on determining the significant codons which are related to replication capacity. The data consists of codon positions which are coded as mutated or non-mutated. This is a biologically relevant question since antiretroviral medications are manufactured to target specific regions on the viral strand. Therefore determining the importance of specific regions or codons is vital when assessing the viral regions which must be targeted by antiretroviral medications.

2.2 Data Description

In this particular dataset, the key phenotype of interest is the replication capacity (RC) of HIV-1, as it reflects the severity of the disease. A measure of replication capacity may be obtained by monitoring viral replication in an ideal environment, with many cellular targets, no exogenous or endogenous inhibitors, and no immune system responses against the virus (Barbour et al., 2002; Segal et al., 2004).

Genotypes of interest correspond to codons in the protease and reverse transcriptase regions of the viral strand. The protease (PR) enzyme affects the reproductive cycle of the virus by breaking protein peptide bonds during viral replication. The reverse transcriptase (RT) enzyme synthesizes double-stranded DNA from the virus' single-stranded RNA genome, thereby facilitating integration into the host's chromosome. Since the PR and RT regions are essential to viral replication, many antiretrovirals (protease inhibitors and reverse transcriptase inhibitors) have been developed to target these specific genomic locations. Studying PR and RT genotypic variation involves sequencing the corresponding HIV-1 genome regions and determining the amino acids encoded by each codon (i.e., each nucleotide triplet).

The HIV-1 sequence dataset consists of $n = 317$ records, linking viral

replication capacity (RC) with protease (PR) and reverse transcriptase (RT) sequence data, from individuals participating in studies at the San Francisco General Hospital and Gladstone Institute of Virology (Segal et al., 2004). Protease codon positions 4 to 99 (i.e., *pr4* – *pr99*) and reverse transcriptase codon positions 38 to 223 (i.e., *rt38* – *rt223*) of the viral strand are studied in this analysis.

The outcome/phenotype of interest is the natural logarithm of a continuous measure of replication capacity, ranging from 0.261 to 151. The M covariates correspond to the $M = 282$ codon positions in the PR and RT regions, with the number of possible codons ranging from one to ten at any given location. A majority of patients typically exhibit one codon at each position. Codons are therefore recoded as binary covariates, with value of **zero** (or “wild-type”) corresponding to the most common codon among the $n = 317$ patients and value of **one** (or “mutation”) for all other codons. Previous biological research was used to confirm mutations and hence provide accurate PR and RT codon genotypes for each patient (hivdb.stanford.edu/cgi-bin/RTMut.cgi).

We will formally define the data as n i.i.d. observations of a random vector $O = (W^*, Y) \sim P_0$. In this case, Y is replication capacity, or the outcome of interest. W^* represents a vector of input variables which are used to predict Y , such as baseline variables, or codons in this specific case. We will let $A = A(W^*)$ refer to a part of W^* for which we want to estimate the variable effect of $A = a$ relative to $A = 0$. In our specific application, we will be estimating the individual variable effect of a mutation versus no mutation at each individual codon position. Therefore, we will let $W^* = (A, W)$ equal a vector of input variables which can be used to predict Y . The marginal variable importance of the variable A can be defined as:

$$P \rightarrow \Psi(P)(1) \equiv E_P^* \{E_P(Y|A = 1, W) - E_P(Y|A = 0, W)\} \quad (1)$$

where $P^* = P^*(P)$ is the distribution of O which is a known function of P . Since A is binary, we will be dealing with the following W -adjusted variable importance: $E(Y|A = 1, W) - E(Y|A = 0, W)$. The W -adjusted variable importance can be defined as:

$$P \rightarrow \Psi(P)(w) \equiv E_P(Y|A = 1, W = w) - E_P(Y|A = 0, W = w) \quad (2)$$

We will define w as any value in the set $w : P(A = 1|W = w)P(A = 0|W = w) > 0$. If the data generating distribution, P , corresponds with sampling an

individual from a population, then this states that the subpopulation defined by $W = w$ should include subjects with $A = 1$ and subjects with $A = 0$. The specific case when $\Psi(P)(1) = 0$ corresponds with the codon having no association with the outcome, on average over all strata of W .

3 Methods

The methods presented in van der Laan (2005) are applied to the HIV-1 and simulated data. Before beginning this analysis, several codons were removed from the analysis based on an inclusion criteria. The criteria includes firstly applying a multiple testing procedure to the set of test statistics built from the marginal association of each codon against the outcome of replication capacity. We used the False Discovery Rate (FDR) procedure which controls the expectation of the ratio of false positives to total rejections at a 0.05 level (Benjamini and Hochberg, 1995). In total, 16 mutations were chosen with an FDR adjusted p -value less than 0.05. Of these 16 mutations, there are three codons which were predicted perfectly by *pr29* (*pr31*, *pr44*, and *pr52*). Therefore, *pr31*, *pr44*, and *pr52* are removed as codons that we are interested in assessing since their significance should be identical to that of *pr29*. In addition, the remaining codons ($282 - 16$) were assessed to determine the $P(A_i = 1)$, which corresponds to the probability that the codon A_i is mutated. Only those codons with $P(A_i = 1) \geq 0.1$ were chosen to obtain variable importance measures. This resulted in 24 additional codons. In conclusion, $(24 + 13) = 37$ codon positions were assessed to determine the variable importance and respective significance measures.

In order to calculate the importance value, we must take into account potential confounders of the relationship between the codon and the outcome of replication capacity. The set W is defined as this set of confounders, and consists of 36 codons, described above (the set of 37 codons minus the codon of interest).

We would like to note that we have noticed that there is a violation of the experimental treatment assignment (ETA) assumption in this dataset. This causal inference assumption states that treatments are not deterministically assigned according to the observed covariate profile. In our specific case the treatment is analogous to the mutation status of the codon. We will define $\Pi(A|W)$ as the probability of each mutation conditional on the set of confounders W , where A is a single codon of interest. The ETA assumption

requires $0 < \Pi(A|W) < 1$ and without this assumption the variable importance measure is not identifiable from data. If Π is close to 0 or 1 then this will make the estimation very difficult and subject the estimator to high levels of variability. In this dataset, we calculated the $\Pi(A|W)$ value for each codon and determined that many of the codon specific probabilities are close to 0 or 1. In this case we have decided to implement the DR-IPTW method, since it is more accurate in the case that the ETA assumption is violated.

3.1 DR-IPTW Method

Before describing the DR-IPTW estimator, we will outline the IPTW estimator since it has components which are utilized in the DR-IPTW estimator. The IPTW estimator is an estimator which assesses the difference in the mean replication capacity between the mutated and non-mutated individuals, weighting by the probability of each mutation conditional on the set of confounders W , $\Pi(A|W)$. We want to note that throughout the following sections, A will be defined as a single codon. In the cases in which we are determining the variable importance of one of the codons in the set W , we will condition on W omitting the particular codon of interest. Therefore, no codon has its estimate conditioned on its own value. In the IPTW case, we will initially define $D(O_i|\Pi_n)$ as:

$$D(O_i|\Pi_n) = Y_i \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W_i)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W_i)} \right)$$

The estimator Ψ_n is defined in terms of $D(O_i|\Pi_n)$ and can be written as:

$$\Psi_n = \frac{1}{n} \sum_{i=1}^n D(O_i|\Pi_n) = \frac{1}{n} \sum_{i=1}^n Y_i \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W_i)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W_i)} \right)$$

In this equation, $\Pi_n(A = 1|W_i)$ corresponds to the estimated probability obtained from fitting a logistic regression, which regresses the binary A values on the W values, or potential confounders. In the case of this analysis the logistic regression was estimated with the POLYCLASS function in R.

POLYCLASS is an exploratory, data-adaptive, or black box regression technique used to predict categorical or binary outcomes. This classification

method, uses forward addition and backward deletion, searches through a series of models defined by main effects, splines and cross-products to create a logistic regression model. The procedure uses cross-validation to choose the complexity (number of basis functions) of the model. This method therefore attempts to balance the variance/bias of the classification error. This data-adaptive logistic regression technique combines stepwise (hierarchical) addition and deletion of variables and finds a linear combination of variables that provides a better predictor of the outcome event.

The double robust method (DR-IPTW) differs from the IPTW method in that it incorporates a regression of $E(Y|A, W)$, which is referred to as $\theta(A, W)$. The quantity $E(Y|A = 1, W) - E(Y|A = 0, W)$ is computed. This quantity is referred to as $\theta_n(A = 1, W) - \theta_n(A = 0, W)$, which corresponds to the difference in effect of $A = 1$ and $A = 0$. In this case θ_n is an estimator of the true regression function $\theta_0(A, W) = E_0(Y|A, W)$. In the DR-IPTW case, we will initially define $D(O_i|\Pi_n, \theta_n)$ as:

$$D(O_i|\Pi_n, \theta_n) = Y_i \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W_i)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W_i)} \right) - \theta(A, W_i) \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W_i)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W_i)} \right) + \theta(1, W_i) - \theta(0, W_i)$$

The estimator Ψ_n is defined in terms of $D(O_i|\Pi_n, \theta_n)$ and can be written as:

$$\Psi_n = \frac{1}{n} \sum_{i=1}^n D(O_i|\Pi_n, \theta_n) = \frac{1}{n} \sum_{i=1}^n Y_i \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W_i)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W_i)} \right) - \theta(A, W_i) \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W_i)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W_i)} \right) + \theta(1, W_i) - \theta(0, W_i)$$

When calculating the DR-IPTW estimator, $\theta(A, W)$ is calculated in this HIV-1 example by using a 37-dimensional main effect model. This is similar to what will be illustrated in the simulation section. The linear regression method always selects all of the variables and therefore is a better method for

the purpose of estimating the variable importance. One wants to adjust richly within each strata A when calculating $\theta(A, W)$. It is also more important in the cases where we have ETA violations because we cannot rely on the Π function. The advantage of the DR-IPTW method is that the estimator remains consistent if either Π_n or θ_n is modeled correctly.

3.2 Inference of Ψ_n

Once the IPTW or DR-IPTW estimator is computed, one is often interested in the inference and therefore statistical significance of the respective variable importance measures. In order to determine the inference on the estimator the influence curve is calculated. The conservative influence curve $IC_1(O) \equiv D(O|\psi_0, \Pi_0)$ is used, therefore yielding a corresponding conservative estimate of the asymptotic variance. We would like to note that the double robust influence curve is conservative in the setting where Π_0 is correctly specified, as shown in van der Laan and Robins (2003). In these analyses, we use this for our variance estimate because of its computational efficiency. However, to obtain a more accurate estimate one should use the bootstrap. Since our HIV-1 results were hardly significant, we did not repeat the analysis using the bootstrap. We will define the variance of the DR-IPTW estimator and note that the IPTW estimator is a special case in which $\theta_n = 0$. The asymptotic variance of $\sqrt{n}(\Psi_n - \Psi)$ can be conservatively estimated with:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (D(O_i|\theta_n, \Pi_n) - \Psi_n)^2$$

In both the IPTW and DR-IPTW cases, the test statistic for the estimator is compared to a $N(0, 1)$ distribution and is defined as:

$$T_n = \frac{\Psi_n}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \sim N(0, 1)$$

3.3 DR-IPTW Estimator of W-adjusted Variable Importance

We are also interested in determining the W-adjusted variable importance $\theta_0(a, W) - \theta_0(0, W)$ which can be represented as $E(D(\theta, \Pi)|W)$ for $\theta = \theta_0$

Table 1: Marginal Test Statistics, Variable Importance Measures, p -values

Codon	Test Statistic	p -value	G-Computational	DR-IPTW Ψ_n	p -value
pr10	-3.7900	0.0002	0.0334	-0.0199	0.8138
pr13	1.4431	0.1489	0.0229	0.0332	0.6982
pr14	-2.4362	0.0148	-0.0081	0.0330	0.5888
pr15	1.6171	0.1058	0.0259	-0.0294	0.7363
pr35	1.3597	0.1138	0.0227	-0.0299	0.7518
pr36	1.2917	0.1964	0.0552	0.0372	0.5777
pr37	1.4975	0.1342	-0.0100	0.0035	0.9711
pr41	-0.3640	0.7158	-0.0206	-0.0638	0.5644
pr43	-6.1184	<0.00001	-0.0631	0.0068	0.9145
pr46	-5.2239	<0.00001	-0.0366	-0.0078	0.8773
pr54	-5.5391	<0.00001	-0.0384	-0.0225	0.6106
pr57	1.5776	0.1146	0.0331	-0.0146	0.9161
pr62	0.4136	0.6791	0.0166	-0.0009	0.9914
pr63	1.5628	0.1181	-0.0254	0.0088	0.9337
pr64	0.6550	0.5124	-0.0097	-0.1089	0.3872
pr71	-3.3939	0.0007	0.0520	0.0732	0.2135
pr77	-1.8474	0.0646	-0.0213	-0.0874	0.3974
pr82	-4.5206	<0.00001	0.0239	0.0173	0.6963
rt83	2.458	0.0139	0.0113	0.0113	0.7866
pr90	-6.2378	<0.00001	-0.1028	-0.0961	0.0964
pr93	0.1422	0.8868	0.0284	-0.0348	0.6764

or $\Pi = \Pi_0$. Therefore, we can estimate this W-adjusted variable importance by building regression models using POLYMARS to predict the outcome $D(O_i|\theta_n, \Pi_n)$.

4 Results

As a result of the ETA violation, we have realized that the IPTW estimator is unreliable and therefore we only report the DR-IPTW measures and p -values. We also report the marginal test statistics and p -values, and the G-computational likelihood based estimator based on the representation $E_W(\theta(A = 1, W) - \theta(A = 0, W))$. This is illustrated in Table 1 and Table 2. These estimates are the W -adjusted variable importance measure for a

Table 2: Marginal Test Statistics, Variable Importance Measures, p -values (continued)

Codon	Test Statistic	p -value	G-Computational	DR-IPTW Ψ_n	p -value
rt41	-5.2253	<0.00001	-0.0773	-0.0728	0.4028
rt102	-3.0874	0.00201	-0.0090	0.0259	0.6948
rt103	-1.3612	0.1734	-0.0031	0.0644	0.3716
rt121	-4.0703	0.0001	-0.0374	0.0453	0.5004
rt122	-1.3460	0.1782	0.0168	0.0228	0.8162
rt135	-2.8680	0.0041	-0.1106	-0.0982	0.3512
rt162	1.1116	0.2663	0.0424	-0.0534	0.5806
rt177	2.0754	0.0379	0.0369	0.0503	0.5312
rt178	-1.8435	0.0652	0.0860	0.0661	0.5408
rt184	-6.1626	<0.00001	-0.1250	-0.1298	0.2953
rt196	-0.1872	0.8514	0.0082	-0.0370	0.7249
rt200	-1.3211	0.1864	0.0104	0.0544	0.6550
rt207	0.6699	0.5029	0.0227	0.0028	0.9727
rt211	2.3058	0.0211	0.0453	0.1130	0.2374
rt214	2.1518	0.00314	0.0483	0.0131	0.8784
rt215	-4.4794	<0.00001	0.1012	-0.0594	0.5259

specified set of potential effect modifiers W . The W -adjusted variable importance measure can be interpreted as the difference in mean replication capacity, holding W constant, between those with a mutation at that specific codon versus those with no mutation at that specific codon. Therefore this variable importance measure is the average impact of the codon within strata of W . There are no significant importance measures in this case, and therefore after adjusting for the other codons, these codons were not significant.

4.1 Data Adaptive Estimation of the W-Adjusted Variable Importance Measure

Tables 3 and 4 display the models assessing the DR-IPTW variable importance measure for each codon. In order to provide an interpretation of these models we will choose for example, the first codon $pr46$, produces the following model: $E(Y|pr46 = 1, W) - E(Y|pr46 = 0, W) = -0.0483 - 0.6129(pr43) + 0.3509(pr62)$. In this case, the codons $pr43$ and $pr62$ best predict the variable importance measure $pr46$. Therefore, this fit states that the impact of $pr46$ is negative with regards to replication capacity among strata where there is a mutation in $pr43$. The effect is only positive among strata where there is a mutation at $pr62$ and no mutation at $pr43$. Another example, in terms of interpretation, is with regards to $pr54$. In this case the following model was obtained: $E(Y|pr54 = 1, W) - E(Y|pr54 = 0, W) = 0.0136 - 0.5196(pr43)$. Therefore this fit states that the impact of $pr54$ is positive with regards to replication capacity among strata where $pr43$ is not mutated and negative when $pr43$ is mutated.

4.2 HIV-1 Results

The procedure identified the variable importance and respective p -values of several codon positions. As previously mentioned there was a violation of the ETA assumption and therefore we reported the DR-IPTW measures and p -values. An interpretation of the results would focus on the fact that after adjusting for the codons in set W there are no significant codons. The marginal association and significance of the individual codons is actually representative of a group of codons, therefore the respective confounders. This lack of significance of the specific codons can be interpreted as a situation

Table 3: DR-IPTW Models (D represents the fit of $E(Y|\text{Codon}=1, W) - E(Y|\text{Codon}=0, W)$)

Codon	Model
pr10	$D = -0.019$
pr13	$D = 0.033$
pr14	$D = 0.033$
pr15	$D = 0.107 - 0.687(pr41)$
pr35	$D = -0.029$
pr36	$D = 0.037$
pr37	$D = 0.003$
pr41	$D = -0.064$
pr43	$D = 0.007$
pr46	$D = -0.048 - 0.613(pr43) + 0.351(pr62)$
pr54	$D = 0.014 - 0.519(pr43)$
pr57	$D = -0.015$
pr62	$D = -0.001$
pr63	$D = -0.139 + 0.840(pr64)$
pr64	$D = -0.143 - 0.148(rt121) - 0.074(pr43) + 14.198(rt121 * pr43) - 0.109(pr10) - 12.702(rt121 * pr10)$
pr71	$D = 0.188 - 0.434(rt200)$
pr77	$D = -0.087$
pr82	$D = 0.053 - 0.605(rt121)$
pr83	$D = 0.011$
pr90	$D = -0.096$
pr93	$D = -0.035$
rt41	$D = 0.010 - 0.414(pr14) + 0.937(pr71) - 0.833(rt215) - 10.597(rt215 * pr14) - 0.712(rt102) + 3.022(rt102 * rt215) - 2.181(pr46) + 7.051(pr14 * pr46) + 2.377(pr46 * rt215)$

Table 4: DR-IPTW Models (continued) (D represents the fit of $E(Y|\text{Codon}=1, W) - E(Y|\text{Codon}=0, W)$)

Codon	Model
rt102	$D = 0.078 - 0.876(rt121)$
rt103	$D = 0.136 - 0.097(pr15) - 0.055(pr77) - 3.095(pr15 * pr77)$
rt121	$D = -0.069 + 0.239(rt177) + 0.113(pr15) + 2.188(pr15 * rt177)$
rt122	$D = 0.023$
rt135	$D = -0.098$
rt162	$D = -0.053$
rt177	$D = 0.198 - 0.497(rt196) - 0.468(pr15) - 6.406(rt196 * pr15)$
rt178	$D = 0.066$
rt184	$D = -0.129$
rt196	$D = -0.037$
rt200	$D = 0.459 - 0.629(pr64) + 0.639(pr46) - 1.057(pr71) + 0.206(pr43) - 6.661(pr43 * pr64) - 0.713(rt162) + 4.55(pr46 * rt162)$
rt207	$D = 0.002$
rt211	$D = 0.113$
rt214	$D = 0.013$
rt215	$D = 0.059$

where many of the other codons are contributing small amounts of information when we adjust for all the variables. Therefore when we adjust for the group of confounders, or codons, the codon of interest does not appear significant. Biologically, in the case of the HIV-1 virus, this could be of importance. Instead of one mutation greatly influencing the replication capacity of the virus, many mutations in combination could confound this effect and create an increase or decrease of replication potential and resistance. For example, many studies show that the addition of *Mpr46I* and *Ipr47V* improves the replication of *Ipr50V* mutant viruses (Prado et al., 2002). In addition, the *Gpr48V/Lpr90M* double mutation has shown delayed viral replication, whereas *Lpr90M* alone had a higher replication capacity (Sune et al., 2004). In regards to resistance, position *pr90* has an impact on the substrate cleft of the virus and *L90M* causes resistance to Saquinavir when combined with various other mutations (Shafer et al., 1998). Therefore the fact that the individual codon positions appear insignificant after adjusting for respective confounding codons could be biologically plausible in the case of HIV-1, as well as in other studies of viral genetics. Further investigation into these insignificant results could determine the confounders for each specific codon and therefore examine the groups of related codons.

In addition, the models presented in Tables 3 and 4 present fitted models for the *W*-adjusted variable importance for all the listed codons. In some cases, the fits include as effect modifiers both protease and reverse transcriptase codons. The current HIV-1 literature does not mention interactions of this nature (protease and reverse transcriptase), so we cannot relate it to the known literature, but they might represent interesting findings. In addition, there are several interactions of only protease or reverse transcriptase, which are not popular mutations in the literature.

Among the codons that have lower *p*-values in Table 1 and Table 2, *pr71*, *pr90*, *rt184*, *rt211* are all codons with a *p*-value less than 0.3. We have decided to focus on these mutations when exploring the biological importance. Mutations in *pr71* and *pr90* have been found to have an impact on replication and resistance of the virus. An example of a protease mutation is position *pr10*, where *Lpr10I/F/V/R*, one of the most common mutations, is associated with resistance to all protease inhibitors when present with another mutation. Position *pr90* has an impact on the substrate cleft of the virus and *L90M* causes resistance to specific drugs when combined with various other mutations (Shafer et al., 1998). In addition, *Mrt184V* reduces replication capacity by reducing the ability of the reverse transcriptase to

process correctly. Viruses that contained the *M184V* mutation were not able to undergo compensatory mutagenesis and reestablish wild-type replication kinetics, viruses that did not contain *M184V* mutated extensively. We refer the reader to Birkner et al. (2005) and Segal et al. (2004) for additional biological information regarding the HIV-1 codon mutations.

5 Simulations

In addition to the HIV-1 data application, analyses on simulated datasets were also performed to investigate the DR-IPTW estimator under various underlying data generating distributions. In the following simulation cases, A denotes the variable of interest, and W is the set of 30 covariates, omitting A . The first simulated dataset refers to a situation in which there is a violation of the experimental treatment assignment (ETA) assumption as described previously. The second simulated dataset creates data in a slightly different manner and therefore does not result in violations to the ETA assumption.

5.1 Simulation I

We will define the first five covariates as X_1, X_2, X_3, X_4, X_5 . In this simulation the number of observations, n , is equal to 500. In order to produce twenty-five correlated variables, we were interested in producing groups of five variables correlated to each of the five initial variables, respectively. This will be described in terms of one of the original X values, though this can be repeated for each of the variables. We simulated X_1 as random $U(0, 1)$ variables. We are interested in simulating $X_6, X_7, X_8, X_9, X_{10}$ as variables correlated with X_1 . We define X_6 , for example, as follows: $X_6 = X_1 * c_1$, where $c_1 \sim U(0.7, 1.3)$. If these variables are greater than $c_2 = 0.6$ the variable specific individual is coded as '1' and otherwise as '0'. This same process occurs for X_7, X_8, X_9, X_{10} . Furthermore, we will simulate X_2 as a variable that is correlated with $X_{11}, X_{12}, X_{13}, X_{14}, X_{15}$ using the same process. In total, this results in 30 total variables. Finally, the outcome, Y , was calculated as follows: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_0, \beta_1, \dots, \beta_5 \sim U(1, 3)$.

We will fit Π with POLYCLASS and the function θ is fit in two different ways, 1) apply POLYMARS to fit $\theta(1, W)$ and $\theta(0, W)$ separately using samples with $A_i = 1$ and $A_i = 0$ respectively, and 2) fit $\theta(A, W)$ with linear

regression with 30 main terms. In the first stratification method, due to the fact that POLYMARS focuses on prediction performance as any other prediction algorithm, it often occurs that the actual variable in the true regression correlated with A is not included since it adds little to the predictive power of the fit. On the other hand, the linear regression method always selects the appropriate variables and therefore is a better method for the purpose of estimating the variable importance, especially in the case of ETA violations since one cannot rely on the Π function.

5.2 Simulation without ETA violations

We will define the first five covariates as X_1, X_2, X_3, X_4, X_5 which are independent binary variables. In this simulation the number of observations, n , is equal to 500. In order to produce twenty-five correlated variables, we were again interested in producing groups of five variables correlated to each of the five initial variables, respectively. This will be described with one of the original X values, though this is repeated for each of the variables. We will choose a random Bernoulli variable ($B(p^*)$) with probability equal to $p^* = (0.3, 0.6, 0.8)$. In each simulation we used a different probability from this set. In the case of X_1 for example, for $j = 1, \dots, n$ observations, if $X_1(j) \neq B(p^*)$ then $X_6(j) = X_1(j)$ otherwise we changed $X_6(j)$ to the opposite value (therefore a '1' to a '0' or a '0' to a '1'). This was repeated for a different draw of the Bernoulli, with the same probability, for each of the variables $X_6, X_7, X_8, X_9, X_{10}$. In addition, X_2 is correlated with $X_{11}, X_{12}, X_{13}, X_{14}, X_{15}$ using the same process. In total, this results in 30 total variables. Again, the outcome Y is generated identical to the previous simulation and the function Π and θ are fit in the same manner as the previous simulation.

5.3 Results

The advantage of this variable importance measure is that we can obtain a targeted ranked list of variables and their respective significance. Table 5 displays the 30 variables and their respective DR-IPTW estimator and p -value in a scenario in which we have ETA violations. Two methods were performed, which deal with the construction of the $\theta(A, W)$ function. The importance measures are not accurate when $\theta(A, W)$ is misspecified in combination with ETA violations (Table 5, columns 2,3). The p -values appear

Table 5: A=Stratification by variable, B=Linear Model with 30 terms

Variable	Estimate ^A	p-value ^A	Estimate ^B	p-value ^B
X ₁	1.3484	<0.00001	0.4894	<0.00001
X ₂	1.8999	<0.00001	0.8511	<0.00001
X ₃	2.2160	<0.00001	1.0946	<0.00001
X ₄	2.7739	<0.00001	1.2681	<0.00001
X ₅	1.1672	<0.00001	0.8468	<0.00001
X ₆	0.5056	0.00003	-0.0695	0.5627
X ₇	0.8051	<0.00001	-0.0293	0.6172
X ₈	0.4607	<0.00001	-0.3154	0.0011
X ₉	0.4643	<0.00001	0.2105	0.0281
X ₁₀	0.5412	<0.00001	0.2105	0.0281
X ₁₁	0.7710	<0.00001	-0.0389	0.7087
X ₁₂	0.7845	<0.00001	-0.0140	0.8636
X ₁₃	0.7306	<0.00001	-0.6233	<0.00001
X ₁₄	0.7308	<0.00001	-0.0989	0.2693
X ₁₅	0.7717	<0.00001	0.7233	<0.00001
X ₁₆	1.3605	<0.00001	-0.0031	0.9600
X ₁₇	0.4654	<0.00001	-0.0240	0.5833
X ₁₈	0.8790	<0.00001	0.1994	0.0322
X ₁₉	0.4827	<0.00001	-0.2434	<0.00001
X ₂₀	0.8638	<0.00001	-0.1586	0.0968
X ₂₁	1.5360	<0.00001	-0.1394	0.0099
X ₂₂	0.9365	<0.00001	-0.2189	0.0196
X ₂₃	1.5854	<0.00001	-0.0895	0.1715
X ₂₄	0.9508	<0.00001	0.1476	0.1247
X ₂₅	1.7680	<0.00001	0.0421	0.5152
X ₂₆	0.1080	0.1292	-0.0971	0.1481
X ₂₇	-0.2009	0.00005	-0.2055	0.0002
X ₂₈	0.8195	<0.00001	0.0517	0.4208
X ₂₉	0.2721	0.0041	-0.0442	0.6206
X ₃₀	0.6502	<0.00001	-0.1282	0.0467

much more reasonable when $\theta(A, W)$ is specified correctly (full main term regression). The DR-IPTW estimator remains consistent when the $\theta(A, W)$ function is or is not misspecified under the assumption that ETA is not violated. In both of the cases, the first five variables remain most significant [Tables 6, 7, and 8].

6 Summary

This article is based on the technique proposed by van der Laan (2005). This method is a targeted method used to construct variable importance measures coupled with respective statistical inference. It involves determining the importance of a variable in predicting an outcome. This method can be applied as inverse probability of treatment weighted (IPTW) or double robust inverse probability of treatment weighted (DR-IPTW) estimators. The variance and respective p -value of the estimate are calculated by estimating the influence curve. The DR-IPTW method was applied to an HIV-1 dataset in which the variable importance and respective statistical significance is reported for each codon of interest. In addition, simulations were performed to examine the performance of the DR-IPTW estimator in both situations in which the experimental treatment assignment (ETA) is and is not violated. The simulations have illustrated the sensitivity of the variable importance estimate to the method used for estimating $\theta(A, W)$ in the case when ETA is violated. Therefore, it is of interest to obtain a data adaptive method for estimating the nuisance parameters θ and Π which takes into account that the variable importance is the parameter of interest. We plan to investigate this important issue in more detail in future research.

Table 6: $p^* = 0.3$; A=Stratification by variable, B=Linear Model with 30 terms

Variable	Estimate ^A	p -value ^A	Estimate ^B	p -value ^B
X_1	1.6407	<0.00001	0.7911	<0.00001
X_2	1.4192	<0.00001	0.6526	<0.00001
X_3	1.7002	<0.00001	0.8322	<0.00001
X_4	1.2425	<0.00001	0.6286	<0.00001
X_5	1.7952	<0.00001	0.8794	<0.00001
X_6	-0.0908	0.3808	-0.0514	0.6218
X_7	0.0549	0.5900	0.0175	0.8616
X_8	-0.1210	0.2369	-0.0545	0.5858
X_9	0.1278	0.1932	0.0679	0.4841
X_{10}	0.0471	0.6280	0.0189	0.8406
X_{11}	0.0084	0.9293	0.0034	0.9708
X_{12}	-0.0257	0.7965	-0.0111	0.9096
X_{13}	-0.0550	0.5674	-0.0418	0.6609
X_{14}	-0.2058	0.03521	-0.0691	0.4760
X_{15}	-0.0162	0.8701	-0.0066	0.9457
X_{16}	-0.0566	0.5505	-0.0143	0.8791
X_{17}	-0.0822	0.3861	-0.0221	0.8135
X_{18}	0.0269	0.7723	0.0062	0.9455
X_{19}	0.1047	0.2672	0.0270	0.7746
X_{20}	0.0442	0.6555	0.0056	0.9553
X_{21}	-0.0309	0.7457	-0.0121	0.8979
X_{22}	-0.0627	0.5385	-0.0358	0.7253
X_{23}	-0.0826	0.4002	-0.0318	0.7460
X_{24}	-0.0124	0.9010	-0.0104	0.9154
X_{25}	0.1024	0.2818	0.0604	0.5164
X_{26}	-0.0107	0.9120	0.0002	0.9982
X_{27}	0.0582	0.5479	0.0475	0.6260
X_{28}	0.0176	0.8520	-0.0044	0.9616
X_{29}	0.1164	0.2138	0.0396	0.6678
X_{30}	0.0688	0.4938	0.0478	0.6329

Table 7: $p^* = 0.6$; A=Stratification by variable, B=Linear Model with 30 terms

Variable	Estimate ^A	p -value ^A	Estimate ^B	p -value ^B
X_1	1.9821	<0.00001	0.9651	<0.00001
X_2	1.4110	<0.00001	0.6614	<0.00001
X_3	1.8145	<0.00001	0.8798	<0.00001
X_4	2.1625	<0.00001	1.0794	<0.00001
X_5	1.7301	<0.00001	0.8536	<0.00001
X_6	0.1101	0.2448	0.0524	0.5697
X_7	-0.0569	0.5297	-0.0220	0.8035
X_8	-0.0752	0.4163	-0.0514	0.5668
X_9	-0.0617	0.5037	-0.0252	0.7822
X_{10}	-0.0216	0.8203	0.0103	0.9114
X_{11}	0.0021	0.9814	0.0083	0.9267
X_{12}	0.0570	0.5304	0.0001	0.9987
X_{13}	-0.0422	0.6444	-0.0190	0.8340
X_{14}	-0.0135	0.8893	-0.0245	0.8009
X_{15}	0.0387	0.6592	0.0265	0.7604
X_{16}	-0.0261	0.7743	-0.0224	0.8010
X_{17}	-0.0077	0.9303	-0.0093	0.9152
X_{18}	0.2734	0.0028	0.1686	0.0618
X_{19}	-0.2540	0.0063	-0.1743	0.0572
X_{20}	-0.0434	0.6335	-0.0155	0.8615
X_{21}	-0.1595	0.0759	-0.1092	0.1229
X_{22}	0.0302	0.7398	0.0004	0.9963
X_{23}	-0.0017	0.9845	-0.0049	0.9559
X_{24}	-0.0706	0.4367	-0.0552	0.5357
X_{25}	-0.0005	0.9951	-0.0056	0.9484
X_{26}	0.0491	0.5829	0.0259	0.7750
X_{27}	-0.0692	0.4639	-0.0198	0.8291
X_{28}	0.0862	0.3519	0.0278	0.7604
X_{29}	0.0453	0.6197	0.0079	0.9294
X_{30}	0.0440	0.6295	0.0228	0.7993

Table 8: $p^* = 0.8$; A=Stratification by variable, B=Linear Model with 30 terms

Variable	Estimate ^A	p -value ^A	Estimate ^B	p -value ^B
X_1	1.5014	<0.00001	0.7455	<0.00001
X_2	2.2608	<0.00001	1.0620	<0.00001
X_3	2.7094	<0.00001	1.3130	<0.00001
X_4	1.7810	<0.00001	0.8927	<0.00001
X_5	2.3074	<0.00001	1.1250	<0.00001
X_6	0.0633	0.5173	0.0182	0.8493
X_7	0.0291	0.7934	0.0475	0.6749
X_8	-0.0322	0.7467	-0.0744	0.4503
X_9	0.0648	0.6053	0.0092	0.9199
X_{10}	0.0633	0.5490	0.0341	0.7492
X_{11}	0.1393	0.1753	0.09581	0.3544
X_{12}	-0.1939	0.0330	-0.1683	0.0682
X_{13}	0.1554	0.1073	0.0585	0.5519
X_{14}	-0.0789	0.3820	-0.0827	0.3561
X_{15}	0.0916	0.3634	0.08576	0.3918
X_{16}	-0.0574	0.5468	-0.0549	0.5616
X_{17}	-0.0017	0.9866	0.0122	0.9017
X_{18}	0.0133	0.8961	0.0324	0.7422
X_{19}	0.1517	0.1465	0.1384	0.1828
X_{20}	-0.0504	0.6027	-0.0891	0.3585
X_{21}	-0.0747	0.4464	-0.0204	0.8303
X_{22}	-0.1149	0.2069	-0.0312	0.7323
X_{23}	-0.0291	0.7821	-0.0032	0.9751
X_{24}	-0.1404	0.1568	-0.1049	0.2947
X_{25}	0.2892	0.1253	0.0612	0.5585
X_{26}	-0.0591	0.5895	-0.0152	0.8870
X_{27}	0.0621	0.5313	0.0414	0.6730
X_{28}	-0.1837	0.1519	-0.0735	0.4287
X_{29}	0.0939	0.3575	0.0630	0.5232
X_{30}	0.0003	0.9969	0.0100	0.9124

References

- Jason D. Barbour, Terri Wrin, Robert M. Grant, Jeffrey N. Martin, Mark R. Segal, Christos J. Petropoulos, and Steven G. Deeks. Evolution of Phenotypic Drug Susceptibility and Viral Replication Capacity during Long-Term Virologic Failure of Protease Inhibitor Therapy in Human Immunodeficiency Virus-Infected Adults. *Journal of Virology*, 76(21):11104–11112, 2002.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.
- M. D. Birkner, S. E. Sinisi, and M. J. van der Laan. Multiple testing and data adaptive regression: An application to HIV-1 sequence data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. URL
- L. Breiman. Random forests - random features. Technical Report 567, Department of Statistics, University of California, Berkeley, 1999.
- Julia G. Prado, Terri Wrin, Jeff Beauchaine, Lidia Ruiz, Christos J. Petropoulos, Simon D.W. Frost, Bonaventura Clotet, Richard T. D’Aquila, and Javier Martinez-Picado. Amprenavir-Resistant HIV-1 Exhibits Lopinavir Cross-Resistance and Reduced Replication Capacity. *AIDS*, 16(7):1009–1017, 2002.
- J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology, Methodological issues*. Birkhäuser, 1992.
- Mark R. Segal, Jason D. Barbour, and Robert M. Grant. Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL Article 2.
- Robert W. Shafer, Mark A. Winters, Sarah Palmer, and Thomas C. Merigan. Multiple Concurrent Reverse Transcriptase and Protease Mutations and Multidrug Resistance of HIV-1 Isolates from Heavily Treated Patients. *Annals of Internal Medicine*, 128(11):906–911, 1998.

Carlos Sune, Lyndall Brennan, David R. Stover, and Thomas Klimkait. Effect of Polymorphisms on the Replication Capacity of Protease Inhibitor-Resistant HIV-1 Variants under Drug Pressure. *Clinical Microbiology and Infection*, 10(2):119–129, 2004.

Mark J. van der Laan. Statistical Inference for Variable Importance. Technical Report 188, Division of Biostatistics, University of California, Berkeley, 2005. URL www.bepress.com/ucbbiostat/paper188.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, 2003.