*The International Journal of Biostatistics*

*Volume* 2, *Issue* 1 2006 *Article* 2

# Statistical Inference for Variable Importance

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

**Recommended Citation:**
van der Laan, Mark J. (2006) "Statistical Inference for Variable Importance," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 2.
**DOI:** 10.2202/1557-4679.1008

# Statistical Inference for Variable Importance

Mark J. van der Laan

## Abstract

Many statistical problems involve the learning of an importance/effect of a variable for predicting an outcome of interest based on observing a sample of $n$ independent and identically distributed observations on a list of input variables and an outcome. For example, though prediction/machine learning is, in principle, concerned with learning the optimal unknown mapping from input variables to an outcome from the data, the typical reported output is a list of importance measures for each input variable. The approach in prediction has been to learn the unknown optimal predictor from the data and derive, for each of the input variables, the variable importance from the obtained fit. In this article we propose a new approach which involves for each variable separately 1) defining variable importance as a real valued parameter, 2) deriving the efficient influence curve and thereby optimal estimating function for this parameter in the assumed (possibly nonparametric) model, and 3) develop a corresponding double robust locally efficient estimator of this variable importance, obtained by substituting for the nuisance parameters in the optimal estimating function data adaptive estimators. We illustrate this methodology in the context of prediction, and obtain in this manner double robust locally optimal estimators of marginal variable importance, accompanied with p-values and confidence intervals. In addition, we present a model based and machine learning approach to estimate covariate-adjusted variable importance. Finally, we generalize this methodology to variable importance parameters for time-dependent variables.

**KEYWORDS:** causal effect, efficient influence curve, estimating function, prediction, variable importance, adjusted-variable importance

**Author Notes:** This research has been supported by NIH grants NIH-R01 GM071397 and NIH R01 GM67233.

# 1 Introduction

In many applications an important goal is the construction of a predictor of an outcome as a function of a collection of input variables based on a learning data set from a particular data generating distribution. One can define an optimal predictor as a parameter of the data generating distribution by defining it as the function of input variables which minimizes the expectation of a particular loss function (of the experimental unit, and the candidate regression) w.r.t. the true data generating distribution. If one selects the squared error loss function (i.e., the square of the difference between the outcome and predicted value), then this optimal predictor is the conditional mean of the outcome, given the input variables. In the statistical literature such location parameters of the conditional distribution of the outcome given the input variables are referred to as regressions: e.g., mean regression and median regression.

In many applications the number of input variables can be very large. As a consequence, assuming a fully parameterized regression model such as a linear regression model with only main terms and minimizing the empirical mean of the loss function (e.g., the sum of squared residual errors in the case of the squared error loss function) is likely to yield poor estimators, since the number of main terms will typically be too large (thereby resulting in over-fitting), and other functional forms of the input variables should be considered. That is, the current type of applications typically demand nonparametric regression estimators. Because of the curse of dimensionality, minimizing the empirical mean of the loss function, i.e., the empirical risk, over all allowed regression functions results in a predictor with perfect performance of the actual data set it was based upon, but poor performance on an independent sample. As a consequence, many estimators follow the sieve loss based estimation strategy. That is, 1) one selects a sequence of subspaces indexed by so called fine tuning parameters (i.e., a sieve), 2) one minimizes or locally minimizes the empirical risk over each subspace to obtain a subspace specific (minimum empirical risk) estimator, and 3) one selects the fine tuning parameter (i.e., the subspace) with an appropriate method aiming to trade off bias and variance. Examples of fine tuning parameters indexing constraints on the space of regression functions are an initial dimension reduction, the number of terms in the regression model, and the complexity of the allowed functional forms (e.g., basis functions). Each specification of the fine tuning parameters corresponds to a candidate estimator of the true underlying regression. In order to select among these candidate estimators (i.e., to select these fine tuning parameters) most algorithms either minimize a penalized empirical risk or minimize the so called cross-validated risk.

If one applies such "machine learning" algorithms to a data set it is very common that the actual resulting fit is very low dimensional. For example,

consider an AIDS application involving prediction of viral replication capacity based on the mutation profile of the HIV-virus (Birkner et al. (2005b)). In order to obtain an estimate of the regression function we applied the cross-validated DSA-algorithm (Sinisi and van der Laan (2004)) which fits the true regression with a linear combination of multi-way interactions, where the size of the model, the complexity of the interaction, and the dimension reduction were selected with cross-validation, and, given the choice of these fine tuning parameters, the algorithm uses deletion, substitution, and addition moves to search in the space of all linear combinations of multi-way interactions satisfying the constraints. In spite of the fact that our data adaptive D/S/A algorithm searched over a high dimensional space of regression functions, it ended up selecting a linear regression with two main terms and a single interaction (Birkner et al. (2005b)). Though such an estimator is based on a sensible trade off between bias and variance of the candidate predictors, the resulting fit is disappointing from two perspectives. Firstly, in most applications one believes that the true regression is a function of almost all variables, with many variables giving very small contributions to the regression. Secondly, a practitioner typically wishes to obtain a measure of variable importance for each variable, and such a low dimensional fit reflects zero importance for all variables not present in the obtained fit. It has been common practice to address the second issue by reporting many of the fits the algorithm has searched over, and to summarize these different fits in a particular manner. Initially, we also followed this approach, but came to the conclusion that the statistical interpretation of such a summary measure is unclear. Bootstrap-aggregation (Breiman (1996)) has been used to obtain a predictor which is high dimensional so that most variables actually contribute to the predictor. However, one is still confronted with the problem that a bias-variance trade off for a predictor (which is a high dimensional parameter) is typically the wrong bias-variance trade-off for a variable importance measure (a real valued parameter). Therefore, such estimates of variable importance based on a bootstrap-aggregated predictors (e.g., as in random forest Breiman (1999)) are typically more biased than necessary.

In addition, contrary to the methods proposed in this article, all these algorithms are not targeted at an estimate of the variable importance for a particular variable, and thereby invest much of their variability in parts of the regression function which do not affect the actual variable importance for the particular variable. 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. Our proposed methods assume models for the wished type of variable importance, and then use general estimating function methodology (Robins and Rotnitzky (1992), van der Laan and Robins (2002)) to determine the class of estimating

DOI: 10.2202/1557-4679.1008 

functions orthogonal to all nuisance parameters, so that the corresponding estimators are fully targeted at the parameter of interest (i.e., the particular variable importance).

We 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 this article we propose estimators of variable importance which are directly targeted at this parameter, thereby guaranteeing that for each variable we obtain a sensible estimate of variable importance, accompanied with a p-value and confidence interval. As a consequence, this methodology requires a separate estimation procedure for each choice of variable. Though our proposed methodology for variable importance can be applied to any definition of variable importance, for the sake of concreteness and presentation, we will focus on a particular definition of variable importance in prediction. In our discussion at the end of this article we discuss and present a more general definition of variable importance and corresponding methodology. In an accompanying follow up article (Birkner and van der Laan (2005)) we will apply and illustrate the methodology with the HIV-replication capacity example described above to obtain a inference for mutation-specific variable importance across the genome of the HIV-virus.

Since the proposed parameters of marginal and $V$-adjusted variable importance, as defined below, and the corresponding estimators correspond with estimators proposed in causal inference, the practical performance of our estimators of marginal and $V$-adjusted variable importance parameters is established in a variety of simulation studies (Wang and van der Laan (2004), Neugebauer and van der Laan (2004),Yu and van der Laan (2003a)). In particular, Wang and van der Laan (2004) considers simulations in the case that the nuisance parameters in the estimating functions defining our estimators are estimated with data adaptive (i.e., nonparametric) estimators. In the future we will also study the practical performance of these methods in the situation that the set of covariates is large relative to sample size, in particular, in comparison with the current regression based approach: by our reasoning above, one would expect that the performance will increase in favor of our methods, but, there is no free lunch in the sense that the second order terms in the expansion of our estimators will start dominating the performance due to the large dimension of the nuisance parameters: for example, our estimators of marginal variable importance might start converging to the true variable importance at a rate slower than $1/\sqrt{n}$. The latter phenomena is often referred to as the curse of dimensionality.

To formalize the statistical problem of estimating variable importance in prediction as addressed in this article, suppose that we observe $n$ i.i.d observations of a random vector $O = (W^*, Y) \sim P_0$, where $Y$ is an outcome of

interest and $W^*$ represent a vector of input variables which can be used to predict $Y$ such as baseline co-variables. Let $A = A(W^*)$ play the role of a particular extraction of $W^*$ for which we want to estimate the variable effect of $A = a$ relative to $A = 0$. We note that $A$ can be any function of $W^*$. For example, $A$ could simply be a component of a vector $W^*$, but it could also represent any other function such as a two-way interaction term corresponding with two components of $W^*$, a linear combination of all components of $W^*$, or a subvector of $W^*$. Let $W$ be such that $W^* = (A, W)$. In order to define the variable importance parameter we will need to assume that $P(A = a \mid W) > 0$ and $P(A = 0 \mid W) > 0$ $P_W$ a.e. In particular, this assumption holds for all $a$ if the support of $W$ is a cartesian product of a support for $A$ and support for $W^*$, and one could also change the target population (by sub-setting on $W$) to make this assumption hold. In applications one might carry out the proposed methodology for a list of variables extracted from $W^*$, as outlined in our road map presentation later in this article. For example, in the HIV-replication capacity example we would estimate the variable importance for each mutation in the HIV-virus. That is, for each mutation in the HIV-virus, we would let $A$ denote this mutation, $W$ denotes the set of other mutations, and $Y$ denote the replication capacity of the virus, and apply the methodology proposed in this article for estimating the variable importance of $A$.

In this article we address, in particular, estimation and inference of the following real valued parameter of the predictor $E_{P_0}(Y \mid A, W)$ on a model for $P_0$ defined as

$$P \to \Psi(P)(a) \equiv E_P(E_P(Y \mid A = a, W) - E_P(Y \mid A = 0, W)).$$

Note that this parameter is only well defined if $P(A = a \mid W)P(A = 0 \mid W) > 0$, $P_W$-a.e. We will refer to the real valued parameter $\Psi(P)(a)$ and the whole "curve" $\Psi(P) = (\Psi(P)(a) : a)$ as $a$-specific *marginal variable importance* and *marginal variable importance* of the variable $A$, respectively. We will also address estimation of ($a$-specific) $W$-adjusted variable importance:

$$P \to \Psi(P)(a, w) \equiv E_P(Y \mid A = a, W = w) - E_P(Y \mid A = 0, W = w),$$

where $w$ can be any value in the set $\{w : P(A = a \mid W = w)P(A = 0 \mid W = w) > 0)\}$. For example, if the data generating distribution $P$ corresponds with sampling a subject from a population, then this states that the subpopulation defined by $W = w$ should contain subjects with $A = a$ and subjects with $A = 0$.

These measures are special cases of a ($a$-specific) $V$-adjusted variable importance, where $V$ can denote any subset of $W$, defined as

$$P \to \Psi(P)(a, v) \equiv E_P\left(E_P(Y \mid A = a, W) - E_P(Y \mid A = 0, W) \mid V = v)\right).$$

DOI: 10.2202/1557-4679.1008 

This $V$-adjusted variable importance parameter is only well defined under the assumption that for all $w$ in a support of the conditional distribution, $P_{W|V=v}$, of $W$, given $V = v$, $P(A = a \mid W = w)P(A = 0 \mid W = w) > 0$. Note that, if $V = W$, then this $V$-adjusted variable importance parameter equals the $W$-adjusted variable importance, but, in general, it equals the regression of the $W$-adjusted variable importance on $V$ evaluated at $V = v$. If $V$ is the empty set, then the $V$-adjusted variable importance equals the marginal variable importance. We decided to denote each of these parameters with the same notation $\Psi$, since they will be treated separately and the dependence on $a, (a, W)$ or $(a, V)$ identifies which of the three parameters is meant.

## Interpretation of variable importance parameter.

We remark here that these measures of variable importance are inspired by their analogues in causal inference. Specifically, if one assumes 1) that the observed data structure $(W, A, Y)$ is chronologically ordered (the time ordering assumption), 2) that it equals a missing data structure $(W, A, Y) = (W, A, Y_A)$ on a set of $a$-specific (so called counterfactual) outcomes $(Y_a : a)$ (consistency assumption), where $a$ varies over the support of $A$, and that 3) $P(A = a \mid (Y_a; a), W) = P(A = a \mid W)$ (no unmeasured confounding or randomization assumption), then

$$\begin{aligned} \Psi(P)(a) &= E_P(Y_a - Y_0) \\ \Psi(P)(a, W) &= E_P(Y_a - Y_0 \mid W) \\ \Psi(P)(a, V) &= E_P(Y_a - Y_0 \mid V). \end{aligned}$$

That is, under these additional assumptions our measures of variable importance can be interpreted as marginal or adjusted causal effects. Because of the fact that the above assumptions 1-3, defining the counterfactual causal inference framework, do not provide any restrictions on the data generating distribution (Gill and Robins (2001), Yu and van der Laan (2002), Gill et al. (1997)), our methods immediately apply to this causal inference model. In particular, under these additional assumptions 1-3 our results for $V$-adjusted variable importance yields nonparametric double robust locally efficient estimators of the causal effect $E_P(Y_a - Y_0 \mid V)$ if $A$ is discrete, and double robust locally efficient estimators of the unknown parameter $\beta_0$ in a model $E_P(Y_a - Y_0 \mid V) = m(a, V \mid \beta(P))$ with $m(0, V \mid \beta) = 0$ for all $\beta$ and $V$. The latter model can be viewed as a semi-parametric marginal structural model $E(Y_a \mid V) = m(a, V \mid \beta) + g(V)$, where $g$ is left unspecified. Here we remind the reader that marginal structural models as introduced by Robins (e.g., Robins (2000)) are models for $E(Y_a \mid V) = g(a, V \mid \beta)$ for a user supplied parametrization $g(\cdot \mid \beta)$.

On the other hand, these causal inference assumptions 1-3 are not necessary to provide meaning-full interpretations to the $V$-adjusted variable importance, and therefore we propose them as important parameters in general. In particular, we have used these measures of variable importance in prediction (e.g., Sinisi and van der Laan (2004)). These variable importance measures measure the importance of a variable adjusting for all other variables used to predict the outcome. In the case that the variable is a surrogate for an unmeasured confounder, then it is still important to establish this variable as important, and subsequent research/experiments could now determine till what degree the variable importance is causal.

In order to understand the relation of this model-free variable importance parameter, and main effects and interactions in the context of a linear regression model, we provide the following illustrations. If $E_P(Y \mid A, W) = \beta_0 + \beta_1 A + \beta_2 W$, then $\Psi(P)(a) = \Psi(P)(a, W) = \beta_1 a$. If $E_P(Y \mid A, W) = \beta_0 + \beta_1 A + \beta_2 AW_1 + \beta_3 W$, then $\Psi(P)(a, W) = \Psi(P)(a, W_1) = \beta_1 a + \beta_2 a W_1$, and $\Psi(P)(a) = (\beta_1 + \beta_2 E W_1)a$. In more generality, if $E_P(Y \mid A, W)$ is a linear combination of multi-way interactions, then $\Psi(P)(a, W)$ is the linear combination of all the multi-way interactions including $a$, obtained by simply deleting the multi-way interactions which do not include $a$. If $E(Y \mid A, W) = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_1 A_2 + \beta_4 W$, then $\Psi(P)(a_1, a_2) = \Psi(P)(a_1, a_2, W) = \beta_1 a_1 + \beta_2 a_2 + \beta_3 a_1 a_2$. In particular, this marginal variable importance curve identifies the interaction $\beta_3$ as $\Psi(P)(1, 1) - \Psi(P)(1, 0) - \Psi(P)(0, 1) = \beta_3$. That is, one could generalize the definition of a two way interaction in a linear regression model by defining a two way interaction as $\Psi(P)(1, 1) - \Psi(P)(1, 0) - \Psi(P)(0, 1)$, where $\Psi(P)(a_1, a_2) = E\{E_P(Y \mid A_1 = a_1, A_2 = a_2, W) - E_P(Y \mid A_1 = 0, A_2 = 0, W)\}$.

## 1.1 Summary and organization of article.

In Section 2 we provide methodology for estimation of and inference for $a$-specific marginal and $V$-adjusted variable importance for a discrete variable $A$ in a nonparametric model. In spite of the fact that a regression $E_P(Y \mid A, W)$ is not a path-wise differentiable parameter of $P$ in a nonparametric model, and thereby is not root-$n$ estimable (so that inference is problematic), we show that marginal variable importance $\Psi(P)$ is a path-wise differentiable parameter for discrete variables $A$ in the nonparametric model. By applying the general estimating function methodology (Robins and Rotnitzky (1992),van der Laan and Robins (2002)), this path-wise differentiability allows us to construct double robust locally efficient estimators of marginal variable importance, which are defined as a solution of an optimal estimating equation for $\psi_0$ indexed by estimators of two unknown nuisance parameters being the conditional probability distribution $\Pi$ of $A$, given $W$, and the regression $\theta(A, W) = E(Y \mid A, W)$.

DOI: 10.2202/1557-4679.1008 

This estimator of marginal variable importance is double robust w.r.t. misspecification of these nuisance parameters in the sense that the consistency of the estimator of marginal variable importance relies on one of these estimators being consistent. In addition, if one of these estimators succeeds in estimating the true nuisance parameter at a rate so that a certain second order term is $o_P(1/\sqrt{n})$, then it follows from general theorems 2.3 and 2.4 in van der Laan and Robins (2002) that this double robust estimator is locally efficient. Specifically, under specified empirical process and conditions on the nuisance parameter estimates, this estimator is asymptotically linear with an influence curve which equals the efficient influence curve if both nuisance parameters are correctly estimated. As a consequence, under these conditions necessary for establishing asymptotic linearity, we have asymptotically valid Wald-type or bootstrap based confidence intervals and p-values allowing us to test the null hypothesis $H_0 : \psi_0 = 0$. This is presented in Subsection 2.1.

In Subsection 2.2 we focus on estimation of $a$-specific $W$-adjusted variable importance. In order to deal with the curse of dimensionality, we redefine $a$-specific $W$-adjusted variable importance as the following parameter on the nonparametric model:

$$\Psi(P)(a, W) \equiv m(a, W \mid \beta(P)),$$

where $m(a, W \mid \beta(P))$ is the projection of $E_P(Y \mid A = a, W) - E_P(Y \mid A = 0, W)$ onto a working model $\{m(a, W \mid \beta) : \beta\}$. This nonparametric approach of dealing with the curse of dimensionality in the context of causal inference was introduced in Neugebauer and van der Laan (2005). In this manner, we created a smoothed version of the original measure of $a$-specific $W$-adjusted variable importance. We establish the wished path-wise differentiability of $\beta(P)$, and develop locally efficient estimators of $\beta_0 = \beta(P_0)$ which are double robust w.r.t. the specification of the conditional distribution $\Pi$ of $A$, given $W$, and the regression $\theta(A, W) = E(Y \mid A, W)$, as outlined above. We show that this double robust locally efficient estimator of the $a$-specific $W$-adjusted variable importance can be represented as a simple least squares estimator, which is thus practically very appealing.

In addition, we will provide a machine learning approach, involving model selection for $m(\cdot \mid \beta)$, for the purpose of learning the original (non-smoothed) $a$-specific $W$-adjusted variable importance $E(Y \mid A = a, W) - E(Y \mid A = 0, W)$.

Subsequently, for a subset $V \subset W$ of co-variables, in Subsection 2.2 we will generalize the above results to a $V$-adjusted variable importance parameter defined on the nonparametric model as

$$\Psi(P)(a, V) \equiv m(a, V \mid \beta(P)),$$

where $m(a, W \mid \beta(P))$ is the projection of $E(E_P(Y \mid A = a, W) - E_P(Y \mid A = 0, W) \mid V)$ onto a working model $\{m(a, V \mid \beta) : \beta\}$.

The above results provide a comprehensive nonparametric methodology for statistical inference for marginal and $V$-adjusted $a$-specific variable importance (for any $V \subset W$) in the case that the variable $A$ is discrete.

In Section 3 we present methodology for estimation of and inference for marginal and $V$-adjusted variable importance based on modelling the $W$-adjusted variable importance, which now applies to general (i.e., $A$ can be continuous or discrete) variables $A$. Contrary to the methodology for discrete $A$ in Section 2, in this case the methodology is not $a$-specific. In Section 3 we assume a model $\{m(\cdot \mid \beta) : \beta\}$ for the true $W$-adjusted variable importance $E(Y \mid A, W) - E(Y \mid A = 0, W)$. We show that, in this smaller model for $P_0$, the parameter $P \to \beta(P)$ (and the corresponding $W$-adjusted variable importance parameter) is now path-wise differentiable for general (e.g., continuous) variables $A$. Again, this allows us now to develop a class of double robust locally efficient estimators. The nuisance parameters in the double robust estimating functions are a regression of a function of $(A, W)$ on $W$ (so that estimation does not necessarily require estimating $\Pi$) and the regression $\theta$. We also show that these estimates of $W$-adjusted variable importance can be mapped into corresponding estimates of marginal variable importance, and $V$-adjusted variable importance.

In Section 4 we assume a model $\{m(\cdot \mid \beta) : \beta\}$ for the true $V$-adjusted variable importance $E_{W|V}E(Y \mid A = a, W) - E(Y \mid A = 0, W)$. We develop a class of double robust estimating functions of the unknown parameter $\beta$ for general variables $A$. The estimating functions for $V$-adjusted variable importance now rely on an inverse weighting by the conditional density $\Pi$ of $A$, given $W$, while the estimating functions in the previous section only depend on this density through a regression. Off course, this is due to the fact that modelling $W$-adjusted variable importance is a more parametric approach than modelling $V$-adjusted variable importance. We also provide a nonparametric machine learning type method for estimating the true $V$-adjusted variable importance involving model selection.

In Section 5 we summarize our methods into a road map for data analysis involving prediction, nonparametric estimation of $V$-adjusted variable importance, marginal variable importance, and multiple testing for marginal variable importance.

Finally, we end this article with a discussion in which we present the generalization of the definition of variable importance to measures of variable importance for time-dependent variables based on longitudinal data structures, again immediately implied by the analogues of causal effects of time-dependent treatment variables in causal inference.

DOI: 10.2202/1557-4679.1008 

# 2 Discrete variable, nonparametric model, $a$-specific variable importance.

## 2.1 Marginal variable importance.

Firstly, we establish path-wise differentiability and a closed form expression for the efficient influence curve/canonical gradient of the marginal variable importance parameter.

**Theorem 1** *Suppose that $O = (A, W, Y) \sim P_0$, where $A$ is a discrete random variable with finite support. Assume the identifiability condition $P(A = a \mid W)P(A = 0 \mid W) > 0$, $P_0$-a.e.*

*Consider the nonparametric model for $P_0$, and let $\Psi(P)(a) = E_P\{E_P(Y \mid A = a, W) - E_P(Y \mid A = 0, W)\}$ be the parameter of interest. Let $\psi_0 = \Psi(P_0)$. The efficient influence curve/canonical gradient at $P_0$ of this parameter is given by:*

$$\begin{aligned} IC^*(O \mid P_0) &= (\theta_0(a, W) - \theta_0(0, W)) - \psi_0(a) \\ &+ \left\{ \frac{I(A = a)}{\Pi_0(a \mid W)}(Y - \theta_0(a, W)) - \frac{I(A = 0)}{\Pi_0(0 \mid W)}(Y - \theta_0(0, W)) \right\}, \end{aligned}$$

*where $\theta_0(a, W) \equiv E(Y \mid A = a, W)$ and $\Pi_0(a \mid W) = P(A = a \mid W)$.*

For notational convenience, we will denote $\psi_0(a)$ with $\psi_0$. This result can be explicitly verified by showing that the path-wise derivative of the parameter at $P_0$ along a 1-dimensional sub-model through $P_0$ with score $s$ in the nonparametric model can be represented as an inner product $\langle IC^*, s \rangle_{P_0} \equiv E_{P_0} IC^*(O)s(O)$. We actually established the formula for the canonical gradient by deriving the influence curve of the substitution estimator of $\psi_0$, since in a nonparametric model there exists only one influence curve which thus equals the efficient influence curve. Since it is straightforward we will not provide the proof of this result.

Consider now the estimating function for $\psi_0$ based on the efficient influence curve given by

$$\begin{aligned} (O, \psi, \theta, \Pi) &\rightarrow D(O \mid \psi, \theta, \Pi) \\ &\equiv \theta(a, W) - \theta(0, W) - \psi \\ &+ \left\{ \frac{I(A = a)}{\Pi(a \mid W)}(Y - \theta(a, W)) - \frac{I(A = 0)}{\Pi(0 \mid W)}(Y - \theta(0, W)) \right\}, \end{aligned}$$

We have the following double robustness result for the unbiasedness of this estimating function.

**Result 1** *Assume* $P(A=a \mid W)P(A=0 \mid W) > 0$, $P_0$*-a.e. We have*

$$E_{P_0} D(O \mid \psi_0, \theta, \Pi) = 0 \text{ if either } \theta = \theta_0 \text{ or } \Pi = \Pi_0.$$

**Proof.** We have

$$\begin{aligned} E_0 D(\cdot \mid \psi_0, \theta, \Pi) &= E_0 \left(\theta(a,W) - \theta(0,W)\right) - E_0 \psi_0(W) \\ &+ \int_W \left( \frac{\Pi_0(a \mid W)}{\Pi(a \mid W)} (\theta_0(a,W) - \theta(a,W)) - \frac{\Pi_0(0 \mid W)}{\Pi(0 \mid W)} (\theta_0(0,W) - \theta(0,W)) \right) dP_0(W). \end{aligned}$$

This completes the proof. □

We can construct a double robust locally efficient estimators by solving the corresponding estimating equation at estimated nuisance parameters. Specifically, given estimators $\Pi_n, \theta_n$ of $\Pi_0, \theta_0$, we estimate $\psi_0$ with

$$\psi_n = \hat{\Psi}(P_n) = P_n D(\cdot \mid \theta_n, \Pi_n),$$

where

$$\begin{aligned} D(O \mid \theta, \Pi) &\equiv \theta_0(a,W) - \theta_0(0,W) \\ &+ \left\{ \frac{I(A=a)}{\Pi_0(a \mid W)} (Y - \theta_0(a,W)) - \frac{I(A=0)}{\Pi_0(0 \mid W)} (Y - \theta_0(0,W)) \right\}. \end{aligned} \tag{1}$$

Here we use the notation $Pf \equiv \int f(x) dP(x)$ for the expectation operator. We note that the latter estimator can be written as:

$$\begin{aligned} \psi_n &= \frac{1}{n} \sum_{i=1}^n Y_i \left( \frac{I(A_i=a)}{\Pi_n(a \mid W_i)} - \frac{I(A_i=0)}{\Pi_n(0 \mid W_i)} \right) \\ &- \frac{1}{n} \sum_{i=1}^n \theta_n(A_i, W_i) \left( \frac{I(A_i=a)}{\Pi_n(a \mid W_i)} - \frac{I(A_i=0)}{\Pi_n(0 \mid W_i)} \right) \\ &+ \frac{1}{n} \sum_{i=1}^n \theta_n(a, W_i) - \theta_n(0, W_i). \end{aligned}$$

If one is willing to assume a correctly specified model for $\Pi_0$, then one could set $\theta_n = 0$, which results in the following estimator

$$\psi_n^0 = \frac{1}{n} \sum_{i=1}^n Y_i \left( \frac{I(A_i=a)}{\Pi_n(a \mid W_i)} - \frac{I(A_i=0)}{\Pi_n(0 \mid W_i)} \right).$$

### 2.1.1 Asymptotic properties.

Under the conditions of Theorem 2.5 in van der Laan and Robins (2002) we have that the double robust estimator $\psi_n$ is consistent and asymptotically linear if either $\Pi_n$ converges to $\Pi_0$ or $\theta_n$ converges to $\theta_0$, where the convergence

needs to be at a rate so that certain second order terms are $o_P(1/\sqrt{n})$. For example, typically it suffices that $\Pi_n$ and $\theta_n$ converge at a rate $o_P(n^{-1/4})$ to their limits, where one of these limits need to be correct. If one is willing to assume a correctly specified parametric model for $\Pi_0$, and $\Pi_n$ is an asymptotically efficient estimator (for the precise statement we refer to van der Laan and Robins (2002), since it only needs to be efficient for a smooth function of $\Pi_0$), then under these regularity conditions, $\psi_n$ is consistent and asymptotically linear with influence curve

$$IC(O \mid P_0) = D(O \mid \psi_0, \theta^*, \Pi_0) - \Pi(D(\cdot \mid \psi_0, \theta^*, \Pi_0) \mid T_\Pi(P_0)),$$

where $\theta^*$ denotes the possibly misspecified limit of $\theta_n$, $T_\Pi(P_0) \subset L_0^2(P_0)$ is the closure of the linear span of the scores of the model for $\Pi_0$, and $\Pi(\cdot \mid T_\Pi(P_0))$ denotes the projection operator onto $T_\Pi(P_0)$ in the Hilbert space $L_0^2(P_0)$ endowed with inner product $\langle h_1, h_2 \rangle_{P_0} = E_{P_0} h_1(O) h_2(O)$. In particular, this shows that under these conditions $\psi_n$ is asymptotically efficient if $\theta^* = \theta_0$: in that case, the projection on the tangent space $T_\Pi(P_0)$ equals 0. Therefore, we refer to this estimator $\psi_n$ as a locally efficient double robust estimator: it is efficient if both working models for $\Pi_0$ and $\theta_0$ are correctly specified and of low enough dimension so that their estimators result in negligible second order terms, and it is consistent and asymptotically linear if one of the working models is correctly specified.

### 2.1.2 Inference and testing for marginal variable importance.

Consequently, in the case that one assumes a correctly specified model for $\Pi_0$, and the needed regularity conditions of Theorem 2.4 in van der Laan and Robins (2002), then one can use as conservative influence curve $IC_1(O) \equiv D(O \mid \psi_0, \theta^*, \Pi_0)$. A corresponding conservative Wald-type based estimate of the asymptotic variance of $\sqrt{n}(\psi_n - \psi_0)$ is thus given by

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{IC}_1(O_i)^2,$$

where $\hat{IC}_1(O) \equiv D(O \mid \psi_n, \theta_n, \Pi_n)$. A corresponding asymptotically conservative Wald-type 0.95-confidence interval is defined as $\psi_n \pm 1.96\sigma_n/\sqrt{n}$. One can test the null hypothesis $H_0 : \psi_0 = 0$ with the test-statistic $T_n = \sqrt{n}\psi_n/\sigma_n$ whose asymptotic distribution is $N(0,1)$ if the null hypothesis is true. In order to avoid a computer intensive bootstrap, we suggest that this approach is also reasonable in the double robust model, though strictly speaking there is no guarantee that the above influence curve is conservative. The actual influence curve in the double robust model, that is, the model which assumes that either $\Pi_0$ is correctly modelled or $\theta_0$ is correctly modelled, is provided in Theorem 2.5 in van der Laan and Robins (2002).

In general, the bootstrap will provide asymptotically valid confidence intervals under the regularity conditions needed to establish the asymptotic linearity of $\psi_n$.

## 2.2 $V$-adjusted variable importance.

In this subsection we present methods for estimation and inference of $a$-specific $V$-adjusted variable importance for a subset of variables $V$ of the complete set of observed covariates $W$. Firstly, we consider estimation of a smoothed $a$-specific $V$-adjusted variable importance parameter. Subsequently we show that our findings imply also a data adaptive machine learning methodology for estimation of the actual $a$-specific $V$-adjusted variable importance.

### 2.2.1 Smoothed V-adjusted variable importance.

Consider a working model $\{m(V \mid \beta) : \beta\}$ for $\psi_0(a,V) \equiv E_0(E_0(Y \mid A = a, W) - E_0(Y \mid A = 0, W) \mid V)$, indexed by a Euclidean parameter $\beta$. Let

$$\beta(P) \equiv \arg\min_{\beta} E_P \left(\Psi(P)(a,V) - \Psi(P)(0,V) - m(V \mid \beta)\right)^2$$

be the parameter of interest, and let the model for $P_0$ be nonparametric. We note that $m(V \mid \beta_0)$ defines a working model based projection of the true $V$-adjusted $a$-specific variable importance on the working model, where $\beta_0 = \beta(P_0)$ denotes the true parameter value.

The following theorem shows that $\beta(P)$ is a path-wise differentiable parameter with a closed-form efficient influence curve/canonical gradient in the nonparametric model.

**Theorem 2** *Assume $P(A = a \mid W)P(A = 0 \mid W) > 0$, $P_0$-a.e. The canonical gradient for $P \to \beta(P)$ at $P_0$ in the nonparametric model is given by:*

$$IC^*(O \mid P_0) \;=\; -c^{-1}(P_0)\frac{d}{d\beta_0} m(V \mid \beta_0)\left(D(O \mid \theta_0, \Pi_0) - m(V \mid \beta_0)\right),$$

*where $D(O \mid \theta_0, \Pi_0)$ is defined in (1), and*

$$\begin{aligned} c(P_0) \;\equiv\; & P_0 \frac{d^2}{d\beta_0^2} m(\cdot \mid \beta_0)(D(O \mid \theta_0, \Pi_0) - m(V \mid \beta_0) \\ & - P_0 \frac{d}{d\beta_0} m(\cdot \mid \beta_0) \frac{d}{d\beta_0} m(\cdot \mid \beta_0)^\top. \end{aligned}$$

Note that, if $m(V \mid \beta)$ is linear in $\beta$, then the second derivative matrix (first term) in the expression for $c(P_0)$ equals zero so that $c(P_0)$ reduces to

$-P_0 \frac{d}{d\beta_0} m(\cdot \mid \beta_0) \frac{d}{d\beta_0} m(\cdot \mid \beta_0)^\top$. We derived this expression for the efficient influence curve by deriving the influence curve of the estimator $\arg\min_\beta P_n(D(\cdot \mid \theta_n, \Pi_0) - m(\cdot \mid \beta))^2$ with $\theta_n$ being a nonparametric estimator, using the fact that in a nonparametric model an influence curve of a regular asymptotically linear estimator equals the efficient influence curve. That this expression is indeed the canonical gradient can be verified explicitly by establishing the path-wise derivative of $\beta(P)$.

Consider now the estimating function for $\beta_0$ based on the efficient influence curve given by

$$\begin{aligned}(O, \beta, \theta, \Pi) &\rightarrow D(O \mid \beta, \theta, \Pi) \\ &\equiv \frac{d}{d\beta} m(V \mid \beta) \left(D(O \mid \theta, \Pi) - m(V \mid \beta)\right),\end{aligned}$$

where we did not include the standardizing derivative matrix $c(P_0)$ in the definition of the estimating function. We have the following double robustness result for this estimating function.

**Result 2** *Assume $P(A = a \mid W)P(A = 0 \mid W) > 0$ $P_0$-a.e. We have*

$$E_{P_0} D(O \mid \beta_0, \theta, \Pi) = 0 \text{ if } \theta = \theta_0 \text{ or } \Pi = \Pi_0.$$

**Proof.** We have

$$\begin{aligned}&E_0(d/d\beta_0 m(V \mid \beta_0)(D(O \mid \theta, \Pi) - m(V \mid \beta_0)) \\ &= E_0(d/d\beta_0 m(V \mid \beta_0)(E_0(D(O \mid \theta, \Pi) \mid W) - m(V \mid \beta_0)) \\ &= E_0(d/d\beta_0 m(V \mid \beta_0)(\psi_0(a, W) - m(V \mid \beta_0)) \\ &= E_0(d/d\beta_0 m(V \mid \beta_0)(E - 0(\psi_0(a, W) \mid V) - m(V \mid \beta_0)) \\ &= E_0(d/d\beta_0 m(V \mid \beta_0)(\psi_0(a, V) - m(V \mid \beta_0)) \\ &= 0,\end{aligned}$$

by definition of $\beta_0$. Note that we used that $E_0(D(O \mid \theta, \Pi) \mid W) = \psi_0(a, W)$, if either $\Pi = \Pi_0$ or $\theta = \theta_0$. This completes the proof. □

We can construct a double robust locally efficient estimator by solving the corresponding estimating equation at estimated nuisance parameters. Specifically, given estimators $\Pi_n, \theta_n$ of $\Pi_0, \theta_0$, we estimate $\beta_0$ with the solution $\beta_n$ of the estimating equation

$$0 = P_n D(\cdot \mid \beta_n, \theta_n, \Pi_n).$$

We note that this estimator can be represented as a least squares estimator:

$$\beta_n = \arg\min_\beta \sum_{i=1}^n \left(D(O_i \mid \theta_n, \Pi_n) - m(V_i \mid \beta)\right)^2.$$

Thus, $\beta_n$ can be computed with standard least squares regression by regressing the outcome $D(O_i \mid \theta_n, \Pi_n)$ on $V_i$ using the model $m(V \mid \beta)$. In particular, if one is willing to assume a correctly specified model for $\Pi_0$ and one sets $\theta_n = 0$, this estimator is given by

$$\beta_n^0 = arg \min_{\beta} \sum_{i=1}^{n} \left( \left\{ \frac{I(A_i = a)}{\Pi_n(a \mid W_i)} - \frac{I(A_i = 0}{\Pi_n(0 \mid W_i)} \right\} Y_i - m(V_i \mid \beta) \right)^2.$$

Since $\beta_n$ is defined as a solution of an estimating equation, statistical inference for $\beta_0$ and testing proceeds in the same manner as outlined in the previous subsection.

### 2.2.2 Data adaptive estimation of $V$-adjusted variable importance.

The representation of our double robust locally efficient estimator of the model based projection of the $V$-adjusted variable importance immediately suggests to apply any given data adaptive regression algorithm to the imputed data set $(D(O_i \mid \theta, \Pi), V_i)$, $i = 1, \ldots, n$, treating $D(O_i \mid \theta, \Pi)$ as outcome.

To formalize the rational behind such an approach, we note the following result.

**Result 3** *If* $P(A = a \mid W)P(A = 0 \mid W) > 0$, $P_0$*-a.e., then*

$$E_0(D(O \mid \theta, \Pi) \mid V) = E_0(E_0(Y \mid A = a, W) - E_0(Y \mid A = 0, W) \mid V),$$

*if either* $\Pi = \Pi_0$ *or* $\theta = \theta_0$.

This is shown by first noting that $E_0(D(O \mid \theta, \Pi) \mid W) = E_0(Y \mid A = a, W) - E_0(Y \mid A = 0, W)$ if either $\Pi = \Pi_0$ or $\theta = \theta_0$. This suggests that one can indeed apply an available machine learning algorithm to the imputed data set to obtain a data adaptive fit of the true $V$-adjusted variable importance. However, we note that if such an algorithm is based on (say) cross-validation to select among candidate estimators applied to a training sample (i.e., a part of the imputed data set), then the cross-validation might not be completely honest since the candidate estimators based on a training sample will still be functions of the validation sample. This is due to the fact that, for an observation $O_i$ in the training sample, $D(O_i \mid \theta_n, \Pi_n)$ depends on the whole sample through $\theta_n, \Pi_n$. It remains to be investigated if this dependence can cause significant bias in the cross-validation method.

Because of this issue, we like to also point out that Result 3 allows us to apply the unified loss based estimation methodology in van der Laan and Dudoit (2003) to the parameter $\Psi(P)(a, V) = E_P(E_P(Y \mid A = a, W) - E(Y \mid A = 0, W) \mid V)$ by defining the loss function for $\psi$ as

$$L(O, \psi \mid \Pi, \theta) \equiv (D(O \mid \Pi, \theta) - \psi(V))^2.$$

DOI: 10.2202/1557-4679.1008 

This loss function is indexed by unknown nuisance parameters $\Pi_0, \theta_0$. By the previous result 3, we have that the parameter of interest is indeed a minimizer of the risk corresponding with this loss function:

$$\psi_0 = \Psi(P_0) = \arg\min_{\psi} E_0 L(O \mid \psi, \Pi, \theta),$$

if either $\Pi = \Pi_0$ or $\theta = \theta_0$. The unified loss based estimation methodology provides now a road map for construction of data adaptive estimators, which is grounded by theory (for formal results we refer to van der Laan and Dudoit (2003)).

This road map is defined by the following steps: 1) develop estimators $\hat{\Pi}(P_n)$, $\hat{\Theta}(P_n)$ of the nuisance parameters in the loss function, 2) define a sequence of subspaces $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ of the parameter space $\mathbf{\Psi}$ for $\Psi$ (i.e., functions of $V$), 3) compute subspace specific estimators such as

$$\hat{\Psi}_s(P_n) \approx \arg\min_{\psi \in \mathbf{\Psi}_s} P_n L(\cdot, \psi \mid \hat{\Theta}(P_n), \hat{\Pi}(P_n)),$$

which aim to minimize the empirical risk over the subspace, and 4) given such candidate estimators $P_n \rightarrow \hat{\Psi}_s(P_n)$ indexed by $s$, we select $s$ with loss-based cross-validation

$$\hat{S}(P_n) = \arg\min_{s} E_{B_n} P^1_{n,B_n} L(\cdot, \hat{\Psi}_s(P^0_{n,B_n}) \mid \hat{\Theta}(P^0_{n,B_n}), \hat{\Pi}(P^0_{n,B_n})),$$

and 5) one estimates $\psi_0 = \Psi(P_0)$ with $\hat{\Psi}_{\hat{S}(P_n)}(P_n)$. We note that this cross-validation selector $\hat{S}(P_n)$ now indeed compares candidate estimators which are only functions of the training sample $P^0_{n,B_n}$, and are thus independent of the validation sample. We refer to Sinisi and van der Laan (2004) for a detailed and concrete description (and implementation) of this loss based estimation methodology, where the calculation of subspace specific estimators involves searching among candidate linear regression models indexed by subsets of basis functions, thereby providing a more aggressive variable selection strategy than forward selection and forward/backward selection, as commonly used in the literature.

# 3 General $A$, modelling $W$-adjusted variable importance.

In this section we present methods for estimation and inference of $V$-adjusted variable importance for a subset of variables $V$ of the complete set of observed covariates $W$, which applies to discrete as well as continuous $A$. In order to deal with the curse of dimensionality we assume a model for $W$-adjusted variable importance. The estimates of marginal variable importance and $V$-adjusted variable importance for $V \subset W$ are now directly derived from the obtained fit for $W$-adjusted variable importance.

## 3.1 W-adjusted variable importance.

We now actually assume a model $\{m(A,W \mid \beta) : \beta\}$, indexed by a Euclidean parameter $\beta$, for $\Psi(P_0)(A,W) \equiv E_0(Y \mid A,W) - E_0(Y \mid A=0,W)$. Let $\beta(P)$ be defined by the equality

$$m(A,W \mid \beta(P)) = \Psi(P)(A,W) = E_P(Y \mid A,W) - E_P(Y \mid A=0,W),$$

and let $\beta_0 = \beta(P_0)$ be the true parameter value. The model $m(\cdot \mid \beta)$ should satisfy $m(0,W \mid \beta) = 0$ for all $\beta$ and $W$. Contrary to the $\beta(P)$ in the $a$-specific model $m(W \mid \beta)$ we used for discrete $A$ in the previous section, in this case, $\beta_0$ identifies the $W$-adjusted variable importance $\psi_0(a,W)$ for each $a$ in the support of $A$.

The model for $P_0$ defined by the restriction $E(Y \mid A,W) - E(Y \mid A = 0,W) = m(A,W \mid \beta_0)$ for some $\beta_0$ can be represented as a generalized semi-parametric regression model

$$E(Y \mid A,W) = m(A,W \mid \beta) + g(W), \tag{2}$$

where $g(W)$ is left unspecified, and $m(0,W \mid \beta) = 0$ for all $W, \beta$. The semi-parametric regression model $E(Y \mid A,W) = m(A \mid \beta) + g(W)$ has been studied in the literature: Newey (1995); Rosenbaum and Rubin (1983); Robins et al. (1992); Robins and Rotnitzky; Yu and van der Laan (2003b). Yu and van der Laan (2003b) note that this model can be trivially generalized to the model we need here: $E(Y \mid A,W) = m(A,W \mid \beta) + g(W)$. The latter three articles derive the orthogonal complement of the nuisance tangent space (i.e., the set of all gradients of the path-wise derivative, which defines all estimating functions of interest, van der Laan and Robins (2002)), the efficient influence curve/canonical gradient, and establish the wished double robustness of the corresponding estimating functions. In particular, for our purpose we refer to Theorem 2.1 and 2.2 in Yu and van der Laan (2003b), and state here the same result for convenience.

**Theorem 3** *(Yu and van der Laan (2003)) Consider the parameter $P \to \beta(P)$ in the model for $P_0$*

$$P_0 \in \{P : E_P(Y \mid A,W) - E_P(Y \mid A=0,W) = m(A,W \mid \beta) \text{ for some } \beta\}.$$

*It is also assumed that the (density at 0) $\Pi(0 \mid W) > 0$ $P_0$-a.e. so that $E_{P_0}(Y \mid A,W) - E_P(Y \mid A=0,W)$ is well defined. The orthogonal complement of the nuisance tangent space at $P_0$ of this parameter $P \to \beta(P)$ is given by:*

$$T_{nuis}^{\perp}(P_0) = \{(h(A,W) - E_0(h(A,W) \mid W))(Y - m(A,W \mid \beta_0) - \theta_0(0,W)) : h\},$$

*where* $\theta_0(W) = E_0(Y \mid A = 0, W)$. *The efficient influence curve/canonical gradient is given by*

$$IC^*(O \mid P_0) = \{h_{opt}(A,W) - E_0(h_{opt}(A,W) \mid W)\}\,(Y - m(A,W) \mid \beta_0) - \theta_0(0,W)),$$

*where*

$$h_{opt}(A,W) = u_0(A,W) - \frac{E_0(u_0(A,W) \mid W)}{E_0(c_0(A,W) \mid W)} \times c_0(A,W), \tag{3}$$

$u_0(A,W) \equiv \frac{1}{c_0(A,W)} \frac{d}{d\beta_0} m(A, W \mid \beta_0)$, *and* $c_0(A,W) \equiv VAR(Y \mid A, W))$. *In particular, if* $VAR(Y \mid A, W) = c_0(W)$ *only depends on* $W$, *then*

$$h_{opt}(A,W) = \frac{1}{c_0(W)} \left( \frac{d}{d\beta_0} m(A, W \mid \beta_0) - E\left( \frac{d}{d\beta_0} m(A, W \mid \beta_0) \mid W \right) \right).$$

Consider now the class of estimating functions for $\beta_0$ indexed by a choice $h$ based on the orthogonal complement of the nuisance tangent space presented in Theorem 3:

$$\begin{aligned} (O, \beta, \theta, \Pi) &\rightarrow D_h(O \mid \beta, \theta, \Pi) \\ &\equiv \{h(A,W) - E_\Pi(h(A,W) \mid W)\}\,(Y - m(A, W \mid \beta) - \theta(0,W)), \end{aligned}$$

where $\Pi$ denotes a candidate for the conditional distribution of $A$, given $W$, and $\theta$ is a candidate for the regression $E(Y \mid A, W)$. The optimal choice $h_{opt}$ is given in (3). A simple candidate for $h$, we recommend for practical use, is

$$h^*(A,W) = \frac{d}{d\beta_0} m(A, W \mid \beta_0) - E\left( \frac{d}{d\beta_0} m(A, W \mid \beta_0) \mid W \right).$$

As shown in the above references (e.g., Theorem 2.2 in Yu and van der Laan (2003b)), it follows straightforwardly that these estimating functions are double robust in the following sense.

**Result 4** *Assume* $\Pi_0(0 \mid W) > 0$ $P_0$*-a.e. We have*

$$E_{P_0} D_h(O \mid \beta_0, \theta, \Pi) = 0 \text{ if either } \theta = \theta_0 \text{ or } \Pi = \Pi_0.$$

We can construct a double robust locally efficient estimator of $\beta_0$ by solving the corresponding estimating equation at estimated nuisance parameters. Specifically, given estimators $\Pi_n, \theta_n, h_n$ of $\Pi_0$, $\theta_0$, $h_{opt}$ (or $h^*$), we estimate $\beta_0$ with the solution $\beta_n$ of the estimating equation

$$0 = P_n D_{h_n}(\cdot \mid \beta, \theta_n, \Pi_n).$$

We note that there is no need to estimate $\Pi_n$, since we only need an estimate of $E_\Pi(h(A,W) \mid W)$ for some given function $h$. Therefore, we recommend

to simply estimate $E_\Pi(h(A,W) \mid W)$ by regressing (an estimate of) $h(A,W)$ on $W$ according to a model or using a particular data adaptive regression algorithm. If $\beta \to m(A,V \mid \beta)$ is linear, then the estimating equation is itself linear in $\beta$ so that the solution $\beta_n$ exists in closed form. That is, in this case, one can rewrite the estimating equation as $B_n(\beta) = c_n$, so that $\beta_n = B_n^{-1}(c_n)$, where $B_n$ is a $k \times k$ matrix ($k$ dimension of $\beta$), and $c_n$ is a $k$-dimensional vector (both deterministic functions of the data $P_n$).

### 3.1.1 Asymptotic properties.

Under the regularity conditions of Theorem in van der Laan and Robins (2002) we have that the double robust estimator $\beta_n$ is consistent and asymptotically linear if either $\Pi_n$ converges to $\Pi_0$ or $\theta_n$ converges to $\theta_0$ ($h_n$ can converge to any $h$), and it is asymptotically efficient if both estimators are consistent, and $h_n$ converges to $h_{opt}$. If one is willing to assume a correctly specified model for $\Pi_0$, and $\Pi_n$ is an asymptotically efficient estimator (for the precise statement we refer to Theorem 2.4 in van der Laan and Robins (2002), since it only needs to be efficient for a smooth function of $\Pi_0$), then under these regularity conditions, $\beta_n$ is consistent and asymptotically linear with influence curve

$$IC(O \mid P_0) = -c_{h,0}^{-1} D_h(O \mid \beta_0, \theta^*, \Pi_0) - \Pi(c_{h,0}^{-1} D_h(\cdot \mid \psi_0, \theta^*, \Pi_0) \mid T_\Pi(P_0)),$$

where $\theta^*$ denotes the possibly misspecified limit of $\theta_n$, $h$ denotes the limit of $h_n$, $T_\Pi(P_0) \subset L_0^2(P_0)$ is the closure of the linear span of the scores of the model for $\Pi_0$, and $\Pi(\cdot \mid T_\Pi(P_0))$ denotes the projection operator onto $T_\Pi(P_0)$ in the Hilbert space $L_0^2(P_0)$ endowed with inner product $\langle h_1, h_2 \rangle_{P_0} = E_{P_0} h_1(O) h_2(O)$.

### 3.1.2 Inference and testing.

Consequently, in the case that one assumes a correctly specified model for $\Pi_0$, one can use as conservative influence curve $IC_1(O) \equiv -c(h_0)^{-1} D_h(O \mid \beta_0, \theta^*, \Pi_0)$. A conservative estimate of the asymptotic covariance matrix of $\sqrt{n}(\beta_n - \beta_0)$ is thus given by

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^n \hat{IC}_1(O_i) \hat{IC}_1(O_i)^\top,$$

where $\hat{IC}_1(O) = D(O \mid \beta_n, \theta_n, \Pi_n)$. An asymptotically conservative 0.95-confidence interval for $\beta_0(j)$ is thus given by $\beta_n(j) \pm 1.96\sqrt{\Sigma_n(j,j)}/\sqrt{n}$, and one can test the null hypothesis $H_0 : \beta_0(j) = 0$ with the test-statistic $T_n(j) = \sqrt{n}\beta_n(j)/\sqrt{\Sigma_n(j,j)}$ whose asymptotic marginal distribution is $N(0,1)$ under the null hypothesis $H_0$. In order to avoid a computer intensive bootstrap, we suggest that this approach is also reasonable in the double robust model,

though strictly speaking there is no guarantee that the above influence curve is conservative. As explained at the end of this section, one can also map this influence curve in an influence curve for variable importance $\psi_0$, and carry out an analogue inference and testing procedure.

## 3.2 Data adaptive estimation of $V$-adjusted variable importance.

Given an estimator $\psi_n(a, W)$ of $W$-adjusted variable importance, the definition of marginal variable importance suggests a substitution estimator

$$\psi_n(a) = E_{P_n}\psi_n(a, W) = \frac{1}{n}\sum_{i=1}^{n}\psi_n(a, W_i),$$

where $P_n$ is the empirical distribution. Similarly, an estimate of $W$-adjusted variable importance implies an estimate of $V$-adjusted variable importance given by

$$\psi_n(a, V) = \hat{E}(\psi_n(a, W) \mid V),$$

where $\hat{E}$ stands for the estimated regression of $\psi_n(a, W_i)$ on $V_i$, $i = 1, \ldots, n$ according to a model for $V$-adjusted variable importance, or based on a data adaptive regression algorithm.

We consider these estimators of marginal and $V$-adjusted variable importance as likelihood based estimators because these estimators have the same form a maximum likelihood estimator would have. A maximum likelihood estimator of $\psi_0(a, W)$ would involve substituting a maximum likelihood estimator of $P_{Y|A,W}$ based on the likelihood $\prod_i P(Y_i \mid A_i, W_i)$ (according to a model) and map this maximum likelihood estimator into a corresponding estimator of $W$-adjusted variable importance. A maximum likelihood estimator of $V$-adjusted variable importance also requires substitution of a maximum likelihood estimator of the conditional distribution of $V$, given $W$: e.g., for marginal variable importance, the maximum likelihood estimator of the distribution of $W$ in the nonparametric model for $P_W$ is the empirical distribution of $W$.

If one is willing to assume a parametric model for $\psi_0(a, W)$ or $\theta_0$, then the above likelihood based estimators of marginal and $V$-adjusted variable importance are sensible. However, if one is concerned with data adaptive estimation of $\psi_0(a, W)$, then an important question is how to estimate $\psi_n(a, W)$ to obtain a good estimator of marginal variable importance (or more general, $V$-adjusted variable importance).

In a nonparametric model, estimation of $\psi_0(a, W))$ requires smoothing or sieve-based estimation, and thereby data adaptive selection among candidate estimators indexed by fine tuning parameters. This data adaptive selection involves trading off the bias and variance of the candidate estimators and

aims to select an estimator of $W$-adjusted variable importance with minimal mean squared error w.r.t. $\psi_0(a, W)$. Because $\Psi(P)(a)$ is actually a path-wise differentiable parameter of $P$ if $A$ is discrete or if we assume $\Psi(P)(a) = \beta a$ (as shown in the next section), and $E_P(Y \mid A, W)$ is a very non-smooth parameter, a bias-variance trade-off for the purpose of estimation of $\psi_0(a, W)$ implies a completely wrong bias-variance trade-off for estimation of $\psi_0(a)$: specifically, one should use an over-fitted $\psi_n(a, W)$ so that the squared-bias of $\psi_n(a)$ is comparable with its variance (which is supposed to be $O(1/n)$ for a path-wise differentiable parameter). Therefore, the actual data adaptive selection of the fine tuning parameter indexing the candidate estimators of $\theta_0$ (or $\psi_0(a, W)$) is problematic, and requires non-standard model selection techniques as presented in van der Laan and Rubin (2005).

In the latter article we present a new estimating function based cross-validation methodology. This methodology would use the optimal estimating function for $\psi_0(a)$ derived from the efficient influence curve of $\Psi$ at $P_0$ in the nonparametric model (or only assuming $\psi_0(a) = \beta_0 a$).

For the interested reader, we present here the details of this methodology as presented in general in van der Laan and Rubin (2005). Firstly, let $D(O \mid \psi, \theta, \Pi)$ be this optimal estimating function for $\psi_0(a)$ as identified by the efficient influence curve, which is indexed by nuisance parameter values $\theta$ for $E(Y \mid A, W)$ and $\Pi$ for $P(A \mid W)$. The actual formula for this estimation function is presented in the previous section for discrete $A$. For continuous $A$, assuming a linear model $\psi_0(a) = \beta a$, the estimating functions are presented in the next section. Secondly, consider a collection of candidate estimators $\hat{\Psi}_h(a, W)$ of $W$-adjusted variable importance indexed by fine tuning parameter $h$. Let

$$\hat{\Psi}_h(P_n) = E_{P_n} \hat{\Psi}_h(P_n)(a, W)$$

be the corresponding substitution estimator of marginal variable importance. Given initial estimators $\hat{\Pi}(P_n)$ and $\hat{\Theta}(P_n)$ of $\Pi_0$ and $\theta_0$, one would now select $h$ by minimizing the cross-validated (say) Euclidean norm of the estimating equation for $\psi_0(a)$:

$$h_n = \arg\min_h E_{B_n} \left\{ \sum_{i:B_n(i)=1} D(O_i \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{\Theta}(P^0_{n,B_n}), \hat{\Pi}(P^0_{n,B_n})) \right\}^2,$$

where $B_n \in \{0, 1\}^n$ denotes a random vector defining a random split of the learning sample $(P_n)$ in a training sample $(P^0_{n,B_n})$ (i.e., $O_i$ is an element of training sample if $B_n(i) = 0$) and validation sample $(P^1_{n,B_n})$ (i.e., $O_i$ is an element of validation sample if $B_n(i) = 1$). Since $P_0 D(O \mid \psi, \theta, \Pi) = -(\psi - \psi_0)$ if either $\theta = \theta_0$ or $\Pi = \Pi_0$, this cross-validated risk is a criterion for selecting $h$ which aims to minimize the distance of $\hat{\Psi}_h(P^0_{n,B_n})$ to $\psi_0(a)$.

DOI: 10.2202/1557-4679.1008 

# 4 General $A$, Modelling V-adjusted variable importance.

Given $V \subset W$, we now assume a model $\{m(A,V \mid \beta) : \beta\}$, indexed by a Euclidean parameter $\beta$, for $\Psi(P_0)(a,V) \equiv E_0(E_0(Y \mid A=a,W) - E_0(Y \mid A=0,W) \mid V)$. That is, we now only model the actual wished $V$-adjusted variable importance, and consider the estimation problem in this model. Let $\beta(P)$ be defined by the equality

$$m(a,V \mid \beta(P)) = \Psi(P)(a,V) = E_P(E_P(Y \mid A=a,W) - E_P(Y \mid A=0,W) \mid V),$$

and let $\beta_0 = \beta(P_0)$ be the true parameter value.

Based on general results in Chapter 2 in van der Laan and Robins (2002) and Theorem 3 one can establish the following result for this model. This is proved by assuming the causal inference framework, copying the proof of the analogue results in Chapter 6 in van der Laan and Robins (2002) for the structural nested models, and noting that the causal inference assumptions do not affect the model for the data generating distribution. This proof is not repeated here.

**Theorem 4** *Consider the parameter $P \rightarrow \beta(P)$ in the model for $P_0$*

$$P_0 \in \{P : E_P(E_P(Y \mid A=\cdot,W) - E_P(Y \mid A=0,W) \mid V) = m(\cdot,V \mid \beta) \text{ for some } \beta\}.$$

*It is also assumed that $\Pi_0(0 \mid W) > 0$ $P_0$-a.e. so that $E_{P_0}(Y \mid A,W) - E_P(Y \mid A=0,W)$ is well defined. The orthogonal complement of the nuisance tangent space at $P_0$, $T_{nuis}^{\perp}(P_0)$, of this parameter $P \rightarrow \beta(P)$ can be represented as follows. Define for a $h = (h_1, h_2, \Pi^*)$*

$$\begin{array}{l} D_h(O \mid \beta_0, \theta_0, \Pi_0) \\ = \frac{\Pi^*(A|V)}{\Pi_0(A|W)}\{h_1(A,V) - E_{\Pi^*}(h_1(A,V) \mid V)\}(Y - m(A,V \mid \beta_0) + h_2(V)) \\ -\frac{\Pi^*(A|V)}{\Pi_0(A|W)}\{h_1(A,V) - E_{\Pi^*}(h_1(A,V) \mid V)\}(\theta_0(A,W) - m(A,V \mid \beta_0) + h_2(V)) \\ + \sum_a \Pi^*(a \mid V)\{h_1(a,V) - E_{\Pi^*}(h_1(A,V) \mid V)\}(\theta_0(a,W) - m(a,V \mid \beta_0) + h_2(V)). \end{array}$$

*Here $\Pi^*$ can be any conditional density of $A$, given $V$. We have*

$$T_{nuis}^{\perp}(P_0) \supset \{D_h(\cdot \mid \beta_0, \theta_0, \Pi_0) : h\}.$$

Under regularity conditions it is possible to achieve equality in the latter statement. Consider now the class of estimating functions for $\beta_0$ indexed by a choice $h$ based on the representation of the orthogonal complement of the nuisance tangent space as presented in Theorem 4:

$$(O, \beta, \theta, \Pi) \rightarrow D_h(O \mid \beta, \theta, \Pi),$$

where $\Pi$ denotes a candidate for the conditional distribution of $A$, given $W$, and $\theta$ is a candidate for the regression $E(Y \mid A, W)$. The optimal choice $h_{opt}$ does not exist in closed form. We recommend the following choice $h^*$:

$$\begin{aligned} h_1^*(A,V) &= \frac{d}{d\beta_0} m(A, V \mid \beta_0) \\ h_2^*(V) &= E(E(Y \mid A=0, W) \mid V) \\ \Pi^*(A \mid V) &= P_0(A \mid V). \end{aligned}$$

It follows straightforwardly that these estimating functions are double robust.

**Result 5** *Assume* $\Pi_0(0 \mid W) > 0$ $P_0$*-a.e. For any function* $h(A,V) = (h_1(A,V), h_2(V), \Pi^*(A \mid V))$*, we have*

$$E_{P_0} D_h(O \mid \beta_0, \theta, \Pi) = 0 \text{ if either } \theta = \theta_0 \text{ or } \Pi = \Pi_0.$$

We can construct a double robust locally efficient estimator of $\beta_0$ by solving the corresponding estimating equation at estimated nuisance parameters. Specifically, given estimators $\Pi_n, \theta_n$ and $h_n$ of $\Pi_0$, $\theta_0$, and $h_{opt}$ (or $h^*$), we estimate $\beta_0$ with the solution $\beta_n$ of the estimating equation

$$0 = P_n D_{h_n}(\cdot \mid \beta_n, \theta_n, \Pi_n).$$

As remarked earlier, if $m(A, V \mid \beta)$ is linear in $\beta$, then the estimating equation can be written as $B_n(\beta_n) = c_n$, so that $\beta_n = B_n^{-1}(c_n)$, where $B_n$ is a $k \times k$ matrix ($k$ dimension of $\beta$), and $c_n$ is a $k$-dimensional vector.

## 4.1 Data adaptive estimation of $V$-adjusted variable importance.

Given a model for $\Psi(P)(a,V) = m(a, V \mid \beta(P))$, we provided a (closed form) double robust estimator of $\beta_0$. One could now search among many candidate models, each time compute the corresponding double robust estimator of the unknown parameters, and use a particular criteria to guide the search for an optimal fit of the true $V$-adjusted variable importance.

We follow the general estimating function based learning approach described in van der Laan and Rubin (2005). Firstly, in order to form an appropriate criteria for $\psi_0$, we re-define the estimating function in $\beta$ as a function in a candidate $\psi(a,V)$ in the parameter space consisting of all functions of $(a,V)$ which equal 0 at $a=0$:

$$\begin{aligned} &D_h(O \mid \psi, \theta, \Pi) \\ &= \tfrac{\Pi^*(A|V)}{\Pi(A|W)} \{h_1(A,V) - E_{\Pi^*}(h_1(A,V) \mid V)\}(Y - \psi(A,V) + h_2(V)) \\ &- \tfrac{\Pi^*(A|V)}{\Pi_0(A|W)} \{h_1(A,V) - E_{\Pi^*}(h_1(A,V) \mid V)\}(\theta_0(A,W) - \psi(A,V) + h_2(V)) \\ &+ \textstyle\sum_a \Pi^*(a \mid V) \{h_1(a,V) - E_{\Pi^*}(h_1(A,V) \mid V)\}(\theta_0(a,W) - \psi(a,V) + h_2(V))). \end{aligned}$$

DOI: 10.2202/1557-4679.1008 

We note that $P_0 D_h(O \mid \psi_0, \theta, \Pi) = 0$ for all $h$ if either $\Pi = \Pi_0$ or $\theta = \theta_0$. Therefore, given a countable collection of basis functions $h_j(A, V)$, $j = 1, \ldots$, we define

$$\Theta(\psi, P \mid \theta, \Pi) = \sum_j P^2 D_{h_j}(\cdot \mid \psi, \theta, \Pi) q_j,$$

as a weighted Euclidean norm of the vector-estimating function $(D_{h_j} : j)$ based on a list of weights $(q_j : j)$. Indeed, this provides a valid criteria for $\psi$ since $\psi_0 = \arg\min_\psi \Theta(\psi, P_0 \mid \theta, \Pi)$ if either $\theta = \theta_0$ or $\Pi = \Pi_0$. The weights could also be data based $q_{jn}$ converging to some fixed weight $q_j$ for $n$ converging to infinity.

Subsequently, one applies the sieve based estimation methodology based on this criteria, which provides a road map for construction of data adaptive estimators, which is grounded by theory (see van der Laan and Rubin (2005) for formal results). This road map is described by the following steps: 1) develop estimators $\hat{\Pi}(P_n)$, $\hat{\Theta}(P_n)$ of the nuisance parameters in the estimating function criterion, 2) define a sequence of subspaces $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ of the parameter space $\mathbf{\Psi}$ for $\Psi$ (i.e., functions of $a, V$ which equal 0 at $a = 0$), 3) compute subspace specific estimators such as

$$\hat{\Psi}_s(P_n) \approx \arg\min_{\psi \in \mathbf{\Psi}_s} \Theta(\psi, P_n \mid \hat{\Theta}(P_n), \hat{\Pi}(P_n)),$$

which aims to minimize the empirical criteria over the subspace, and 4) given such candidate estimators $P_n \to \hat{\Psi}_s(P_n)$ indexed by $s$, we select $s$ with the estimating function based cross-validation selector (van der Laan and Rubin (2005))

$$\hat{S}(P_n) = \arg\min_s E_{B_n} \Theta(\hat{\Psi}_s(P^0_{n,B_n}), P^1_{n,B_n} \mid \hat{\Theta}(P^0_{n,B_n}), \hat{\Pi}(P^0_{n,B_n})),$$

and 5) one estimates $\psi_0 = \Psi(P_0)$ with $\hat{\Psi}_{\hat{S}(P_n)}(P_n)$.

For example, if the subspace $\mathbf{\Psi}_s$ consists of all linear combinations of maximally $s_1$ basis functions with maximal complexity $s_2$, then the calculation of the subspace-specific estimators $\hat{\Psi}_s(P_n)$ involves searching over candidate linear models (say $m(\mid \beta)$), and calculating the corresponding double robust estimator of the unknown coefficients (i.e., $\beta_n$), as presented in the previous subsection.

# 5 A road map for data analysis involving prediction.

Consider a sample $(W_i^*, Y_i)$, $i = 1, \ldots, n$. Let $\psi_{j0}(a) = E_0\{E_0(Y \mid A_j = a, W_j^*) - E(Y \mid A_j = 0, W_j^*)\}$ be a marginal variable importance parameter

corresponding with a variable $A_j$ extracted from $W^*$, such as a component of $W^*$, where $W^* = (A_j, W_j^*)$. In order to avoid identifiability issues, let's assume that the support of $W^*$ is a Cartesian product of supports for $A_j$ and $W_j^*$, $j = 1, \ldots, J$. Similarly, for a user supplied specification of $V^* \subset W^*$, let $\psi_{j0}(a, V^*) = E_0(E_0(Y \mid A_j = a, W_j^*) - E_0(Y \mid A_j = 0, W_j^*) \mid V_j^*)$, where again $V^* = (A_j, V_j^*)$ is a decomposition of $V^*$ in $A_j$ and a remaining $V_j^*$, $j = 1, \ldots, J$.

We refer to our accompanying article Birkner and van der Laan (2005) for an illustration of the proposed type of data analysis in order to relate mutations in the HIV-virus to replication capacity of the virus. In this case, $A_j$ denotes the $j$-th mutation, and one is concerned with inference regarding the marginal importance of the $j$-th mutation, and how the variable importance of the $j$-th mutation is modified by the absence or presence of other mutations.

**Prediction:** Using a particular machine learning algorithm one can obtain a fit of the optimal predictor $E(Y \mid W^*)$ with corresponding performance assessment using cross-validation. This predictor will typically have very poor performance due to the curse of dimensionality. But, as argued in this article, by going after specific features of $E(Y \mid W^*)$, such as marginal variable importance parameters, it is still possible to learn a lot about this predictor.

**Marginal variable importance:** Report estimators of $(\psi_{jn}(a) : a)$ which provide insight in the importance of variable $A_j$, $j = 1, \ldots, J$. For each $a$, one can accompany $\psi_{jn}(a)$ with a standard error estimate $\sigma_{jn}(a)$, and p-value $\text{pval}_j(a)$ for the null hypothesis $H_{0j}(a) : \psi_j(a) = 0$ as implied by the observed value of test-statistic $T_{jn}(a) = \psi_{jn}(a)/\sigma_{jn}(a)$, and the assumption that this test-statistic is distributed $N(0, 1)$ under $H_{0j}(a)$ (which is supposedly true asymptotically by the central limit theorem). This yields now a list $(\psi_{jn}(a), \sigma_{jn}(a), \text{pval}_j(a) : a)$, $j = 1, \ldots, J$.

**Multiple testing for marginal variable importance:** Consider the test-statistic indexed by the possible $a$ values and $j = 1, \ldots, J$. Given the list of corresponding marginal p-values $(\text{pval}_j(a) : a, j)$, one can use multiple testing methods based on marginal p-values to control the number of false positives. For example, one can apply the multiple testing method of Benjamini and Hochberg (1995) which controls the False Discovery rate (FDR).

One can also apply the general re-sampling based multiple testing methodology controlling a user supplied Type-I error rate (e.g., Family wise error, Tail Probability of proportion of false positives among rejections) as presented in Pollard and van der Laan (2004), and generalized in our subsequent articles Dudoit et al. (2004); van der Laan et al. (2004b) and van der Laan et al. (2004a). The latter methodology aims to estimate

DOI: 10.2202/1557-4679.1008 

the mean zero centered multivariate Gaussian joint limit distribution of the test statistic vector $T_n = (T_{jn}(a) : a, j)$ under the true data generating distribution, in order to be less conservative than methods which are only based on marginal p-values (and thus need to be necessarily conservative in order to be valid under (e.g.) independence). This estimate of the joint distribution of $T_n$ can be obtained with the bootstrap or one can use the vector influence curve $IC(O \mid P_0)$ of $(\psi_{jn}(a) : a, j)$ to estimate the asymptotic covariance matrix with the empirical covariance matrix of $\hat{IC}(O_i)$, $i = 1, \ldots, n$ (see Pollard and van der Laan (2004)).

**$V^*$-adjusted variable importance.** Given a collection of user supplied values for $V^*$, report estimators estimators $(\psi_{jn}(a, V^*) : a)$ which provide insight in the importance of variable $A_j$ at co-variate profile $V^*$. Firstly, consider the case that the estimate $\psi_{jn}(a, V^*)$ was based on a model $m_j(a, V^* \mid \beta_j)$ so that one can assume that the estimate is asymptotically linear and thereby normally distributed. For each $a$ and $V^*$, one can accompany $\psi_{jn}(a, V^*)$ with a standard error estimate $\sigma_{jn}(a, V^*)$, and p-value $\text{pval}_j(a, V^*)$ for the null hypothesis $H_{0j}(a, V*) : \psi_j(a, V^*) = 0$ as implied by the observed value of test-statistic $T_{jn}(a, V^*) = \psi_{jn}(a, V^*)/\sigma_{jn}(a, V^*)$. This p-value can be calculated under the assumption that the test-statistic is distributed $N(0, 1)$ under $H_{0j}(a, V^*)$ (which is supposedly true asymptotically by the central limit theorem). This yields now, for each $V^*$, a list $(\psi_{jn}(a, V^*), \sigma_{jn}(a, V^*), \text{pval}_j(a, V^*) : a)$, $j = 1, \ldots, J$.

If $A$ is discrete, then in this article we also presented data adaptive regression estimates for $\psi_{jn}(a, V^*)$ involving regression of imputed outcomes on $V^*$. This suggests using the permutation distribution (under which the imputed outcome is independent of $V^*$) for this imputed data set to obtain a null distribution for $\psi_{jn}(a, V^*)$ and obtain a p-value $\text{pval}_j(a, V^*)$ for $\psi_{jn}(a, V^*)$ under this permutation null distribution. We refer to Birkner et al. (2005a) for details on such permutation based testing methods using a test statistic derived from a data adaptive regression estimator.

We also suggest to still report standard error estimates and p-values based on the actual selected model, treating the data adaptively selected model as given a priori, thereby ignoring the variability due to the model selection. The latter measures can now be interpreted as lower bounds for the variance and significance of our reported data adaptive estimators $\psi_{jn}(a, V^*)$.

**Multiple testing for $V^*$-adjusted variable importance:** Consider the test-statistic indexed by the possible $a$ values and $j = 1, \ldots, J$. Given the list of corresponding marginal p-values $(\text{pval}_j(a, V^*) : a, j)$, one can use mul-

tiple testing methods based on marginal p-values to control the number of false positives. For example, one can apply the method of Benjamini and Hochberg (1995) controlling the False Discovery rate (FDR).

In the case that the estimates were model based so that asymptotic linearity can be assumed, one could apply the general re-sampling based multiple testing methodology. On the other hand, if the estimates of $V^*$-adjusted variable importance were based on data adaptive learning algorithms, then there is no guarantee that this methodology is asymptotically valid (e.g., conservative).

We remark that one could use the ranking of the p-values for marginal variable importance to obtain an ordered list of variables, and obtain candidate regression fits by applying a machine learning algorithm to the top $k$ variables. The dimension reduction $k$ can now be selected with cross-validation. In this manner our methods for variable importance can be used to generate candidate dimension reductions, which provide a potentially useful alternative to dimension reductions based on, for example, marginal associations or principal components.

# 6 Generalization and conclusion

We end this article with pointing out the generalization to statistical inference for time-dependent variables, and we conclude by stating the general message behind this article.

## 6.1 Variable importance for time-dependent variables.

Let $W = (W(0), \ldots, W(K)), Y$ be a longitudinal data structure collected over time at time points indexed by $k = 0, \ldots, K+1$, where $Y$ is the final outcome measured at time $K+1$. Decompose $W(k) = (L(k), A(k))$, where $A = (A(k) : k = 0, \ldots, K)$ represents a time-dependent component of $(W(k) : k = 0, \ldots, K)$ for which we want to determine a particular measure of importance in affecting $Y$. If one would define variable importance of $A$ in terms of $E(Y \mid W)$, then this might become a hard to interpret parameter, since it corresponds with adjustment by variables on the pathway from $A(k)$ to the future. Therefore, we propose to define measures of variable importance for a time-dependent variable implied by their analogues in a causal inference counterfactual model. In order to define such measures we refer to the so called $G$-computation formula in causal inference (e.g., Robins (2000)) which, under the consistency assumption and sequential randomization assumption (i.e., no unmeasured time-dependent confounding assumption) identifies the distribution of the observed data structure if one intervenes by setting

$A = a = (a(0), \ldots, a(K))$, but leaves the remainder of the data generating distribution in tact. It is obtained by factorizing the density of $O$ as a product over time of conditional probabilities, given the past, erasing the conditional probabilities for $A(k)$, and setting $A(k) = a(k)$ in the conditioning events of the remaining conditional probabilities:

$$P_a(\bar{L}, Y) \equiv \prod_{k=0}^{K} P(L(k) \mid \bar{L}(k-1), \bar{A}(k-1) = \bar{a}(k-1)) P(Y \mid \bar{L}(k), \bar{A}(k) = \bar{a}(k)).$$

Here we used the common notation $\bar{X}(k) = (X(0), \ldots, X(k))$ to denote the history of a time dependent process $X()$ up till time $k$. Note that this $G$-computation formula represent a parameter of the data generating distribution $P$ of $(W, Y)$. Let $V$ be a specified subset of the baseline co-variables $L(0)$, and let $P_{a,Y|V}$ be the conditional distribution of $Y$, given $V$, under $P_a$.

Now, we define $V$-adjusted variable importance of the time-dependent variable $A$ at $a = (a(0), \ldots, a(K))$ as

$$\Psi(P)(a, V) \equiv \Phi(P_{a,Y|V}) - \Phi(P_{0,Y|V}) \in \mathbb{R},$$

where $\Phi$ denotes a real value parameter defined on the set of all conditional distributions of $Y$, given $V$. For example, $\Phi(P_{a,Y|V})(a, V) = E_{P_{a,Y|V}}(Y \mid V)$ is the conditional mean. The latter definition of $\Psi(P)(a, V) = E_{P_{a,Y|V}}(Y) - E_{P_{0,Y|V}}(Y)$ for $K = 0$ reduces to our definition of $V$-adjusted variable importance used in this article. Our results in this article can be generalized to this more general definition of variable importance, since they simply rely on the general estimating function methodology as presented in van der Laan and Robins (2002). The analogues of the estimating functions and efficient influence curve calculations as presented in this article can now also be used in the causal inference framework for time dependent treatment to estimate a semi-parametric marginal structural model $E(Y_a - Y_0 \mid V) = m(a, V \mid \beta)$. For the sake of space, we will not present here the analogues of our formulas. Clearly, this framework allows many other interesting definitions of variable importance for time-independent (using the analogues of other definitions of causal effects) and time dependent variables. For example, in our accompanying technical report we present a generalization of $V$-adjusted marginal variable importance which includes, as special case, direct causal effects in the causal inference framework.

## 6.2 Conclusion

To conclude, the overall message of this article can be summarized as follows. Firstly, the causal inference literature provides definitions of causal effects for treatment variables which also imply interesting definitions of variable importance parameters of the data generating distributions not relying on the

consistency and randomization assumptions needed to define these causal effects. Secondly, estimators developed for such measures of variable importance are also estimators of the corresponding causal effects if the additional causal inference assumptions hold, and visa versa, estimators developed for causal effects (i.e., $E(Y_a - Y_0 \mid V)$) in the causal model are estimators of the corresponding definition of variable importance in general. In either model, the general estimating function methodology in van der Laan and Robins (2002) yields double robust locally efficient estimators *targeted* at each specific variable importance. Finally, the approach of 1) identifying a large list of specific parameters of a high dimensional parameter such as a regression or whole density of the data, and 2) applying the estimating function methodology, combined with data adaptive estimation to estimate the nuisance parameters in these estimating functions, to each of these parameters *separately*, provides a road map for data analysis which allows one to potentially learn more about the high dimensional parameter than one would learn with a substitution based approach, as currently applied in the machine learning literature.

## Acknowledgement

I would like to thank the referees for their helpful and insightful comments. In addition, I thank James Robins for stimulating discussions. This research was funded by NIH grant R01 GM071397.

# References

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. Birkner, M.J. van der Laan, and A. Hubbard. Data adaptive pathway testing. Technical report, Division of Biostatistics, Universtiy of California, Berkeley, 2005a. URL `www.bepress.com/ucbbiostat/`.

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. *Journal of Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005b.

M.D. Birkner and M.J. van der Laan. Application of a variable importance measure method to HIV-1 sequence data. Technical report, Division of Biostatistics, University of California, Berkeley, 2005.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman. Random forests - random features. Technical Report 567, Department of Statistics, University of California, Berkeley, 1999.

S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, 2004. URL `www.bepress.com/sagmb/vol3/iss1/art13`.

R. Gill and J.M. Robins. Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.*, 29(6), 2001.

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–94, New York, 1997. Springer Verlag.

R. Neugebauer and M.J. van der Laan. Why prefer double robust estimates. *Journal of Statistical Planning and Inference*, 2004.

R. Neugebauer and M.J. van der Laan. Locally efficient estimation of nonparametric causal effects on mean outcomes in longitudinal studies. Technical Report 134, Division of Biostatistics, University of California, Berkeley, 2005.

W.K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 1(4):335–341, 1995. ISSN 1350-7265.

K. S. Pollard and M. J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125(1–2):85–100, 2004.

James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 2000.

J.M. Robins, S.D Mark, and W.K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

J.M Robins and A. Rotnitzky. Comment on Inference for semiparametric models: some questions and an answer, by Bickel, P.J. and Kwon.

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *Aids Epidemiology, Methodological Issues*. Birkhauser, Boston, 1992.

P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

S. Sinisi and M.J. van der Laan. The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Journal of Statistical Methods in Molecular Biology*, 3(1), 2004.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, 2004a. URL `www.bepress.com/sagmb/vol3/iss1/art15`.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 14, 2004b. URL `www.bepress.com/sagmb/vol3/iss1/art14`.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon net estimator: Finite sample oracle inequalities and examples. Technical report 130, Division of Biostatistics, University of California, Berkeley, Nov. 2003.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2002.

M.J. van der Laan and D. Rubin. Estimating function based cross-validation and learning. Technical Report 180, Division of Biostatstics, University of California, Berkeley, 2005. URL `www.bepress.com/ucbbiostat/paper180/`.

Y. Wang and M.J. van der Laan. Data adaptive estimation of the treatment specific mean. Technical report, Division of Biostatistics, University of California, Berkeley, 2004.

Z. Yu and M.J. van der Laan. Construction of counterfactuals and the g-computation formula. Technical report, Division of Biostatistics, UC Berkeley, 2002.

Z. Yu and M.J. van der Laan. Double robust estimation in longitudinal marginal structural models. Technical report, Division of Biostatistics, University of California, Berkeley, 2003a.

Z. Yu and M.J. van der Laan. Measuring treatment effects using semiparametric models. Technical Report 136, Division of Biostatstics, University of California, Berkeley, 2003b. URL www.bepress.com/ucbbiostat/paper136/.