

## Statistical Inference for Variable Importance

Mark J. van der Laan\*

\*Division of Biostatistics, School of Public Health, University of California, Berkeley,  
laan@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/ucbbiostat/paper188>

Copyright ©2005 by the author.

# 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 typical 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) carefully defining the wished 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 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 locally optimal estimators of marginal variable importance and covariate-adjusted variable importance, accompanied with p-values and statistical inference. We also propose a road map for statistical analysis based on this approach. Finally, we generalize this methodology to variable importance parameters for time-dependent variables.

# 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. to 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, in an AIDS application involving prediction of viral replication capacity based on the mutation profile of the HIV-virus, in spite of the fact that the actual 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 do 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 predictor (e.g., as in random forest Breiman (1999)) are in a very obvious way heavily affected by the curse of dimensionality. 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. 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 definition of

variable importance in prediction. At the end of this article we discuss and present a more general definition of variable importance and corresponding methodology.

To formalize the statistical problem of learning variable importance in prediction, 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 the 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  $W^*$  or a linear combination of all components of  $W^*$ . Let  $W^* = (A, W)$ , where we typically assume (to avoid special cases) that  $P(A = a | W) > 0$  and  $P(A = 0 | W) > 0$   $P_W$  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  $W^*$ .

In applications one would carry out the proposed methodology for a list of variables extracted from  $W^*$ , as outlined in our road map presentation later in this article. In this article we address, in particular, estimation and inference of the following real valued parameter of the predictor  $E_{P_0}(Y | A, W)$  on a model for  $P_0$  defined as

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

where  $P^* = P^*(P)$  is a distribution of  $O$  which is a known function of  $P$ . Note that this parameter is only well defined if  $P(A = a | W)P(A = 0 | W) > 0$ ,  $P^*$ -a.e. We will refer to this real valued parameter as *marginal variable importance* of the variable  $A$ . We also address estimation of  $W$ -adjusted variable importance:

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

where  $w$  can be any value in the set  $\{w : P(A = a | W = w)P(A = 0 | W = w) > 0\}$ . 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  $V$ -adjusted variable importance, where  $V$  can denote any subset of  $W$ , defined as

$$P \rightarrow \Psi(P)(a, v) \equiv E_{P^*}(E_P(Y | A = a, W) - E_P(Y | A = 0, W) | V = v).$$

This 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  $V = v$ ,  $P(A = a | W = w)P(A = 0 | W = w) > 0$ . 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.

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)  $P(A = a | (Y_a; a), W) = P(A = a | W)$  (no unmeasured confounding assumption), then for the case  $P^*(P) = P$

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

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  $E_P(Y_a - Y_0 | V)$  if  $A$  is discrete, and double robust locally efficient estimators of the unknown parameter  $\beta_0$  in a model

$E_P(Y_a - Y_0 | V) = m(a, V | \beta(P))$ . The latter model can be viewed as a semi-parametric marginal structural model  $E(Y_a | V) = m(a, V | \beta) + g(V)$ , 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  $E(Y_a | V) = g(a, V | \beta)$  for a user supplied parametrization  $g(\cdot | \beta)$ .

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 proposed them as important parameters in general (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 our definition of marginal variable importance  $P^*$  could represent a partly or completely known user supplied distribution, or it can be  $P$  itself. By allowing  $P^*$  to be any data generating distribution implied by  $P$ , we can cover a range of variable importance measures of interest. We illustrate this in the next subsection.

## 1.1 An example of variable importance as a causal (direct) effect.

Suppose that  $(A, A_2)$  represent a joint treatment/exposure variable, and that  $W_1$  denotes a set of pre-treatment variables, which are potential confounders of the joint treatment effect. A particularly interesting case is obtained by letting part or the whole of  $A_2$  be measured after  $A$ , so that the variable importance parameter defined below, which simply corresponds with a particular choice  $P^*(P)$ , can be interpreted as a direct effect parameter of  $A$  blocking its effect on  $Y$  through  $A_2$  (as in Robins and Greenland (1992), Pearl (2000), van der Laan and Petersen (2004)). Under the assumption that  $(A, A_2)$  is randomized, given  $W_1$ , and the missing data structure assumption  $O = (W_1, A, A_2, Y = Y(A, A_2))$ , where  $Y(a, a_2)$  is a treatment specific outcome (counterfactual), we have that  $E_{W_1}\{E_P(Y | A = a, A_2 = a_2, W_1) - E_P(Y | A = 0, A_2 = a_2, W_1)\} = EY(a, a_2) - Y(0, a_2)$  represents an interesting causal parameter. That is, it represent the causal effect of the first treatment component  $A = a$  at a fixed level  $A_2 = a_2$ . One might

now wish to obtain a summary measure of these  $a_2$ -specific causal effects by taking an average over all  $a_2$ -values w.r.t. to a user supplied distribution of  $A_2$ , given  $W_1$ . That is, one might define the parameter of interest as

$$E\left(\sum_{a_2}(Y_{a,a_2} - Y_{0,a_2})g(a_2 | W_1)\right)$$

for a given distribution  $g$  of  $A_2$ , given  $W_1$ . Now, we note that this parameter can be represented as

$$E\left(\sum_{a_2}(Y_{a,a_2} - Y_{0,a_2})g(a_2 | W_1)\right) = E_{P^*}\{E_P(Y | A = a, A_2, W_1) - E_P(Y | A = 0, A_2, W_1)\},$$

where the marginal distribution of  $P^* = P^*(P)$  of  $W_1$  equals the marginal distribution of  $W_1$  under  $P$ , and the conditional distribution of  $A_2$ , given  $W_1$ , of  $P^*$  equals  $g(\cdot | W_1)$ . Thus this causal parameter equals our marginal variable importance parameter for this particular choice of  $P^*$  and setting  $W = (A_2, W_1)$ . If one sets  $g(a_2 | W_1)$  equal to the conditional distribution of  $A_2$ , given  $W_1$ , under  $P$ , then  $P^* = P$ . If  $A_2$  represents an intermediate variable between  $A$  and the outcome  $Y$ , then this parameter can be interpreted as a so called direct effect (of  $A$ , not mediated through  $A_2$ ) parameter by setting  $g(\cdot | W_1)$  equal to a distribution of  $A_2$  it would have followed in the world where  $A = 0$  (i.e.,  $g(A_2 | W_1) = P(A_2 | A = 0, W_1)$ ): see van der Laan and Petersen (2004).

Though the causal inference model defined by the time-ordering assumptions, randomization assumption, and missing data structure assumption (1-3 above) allows one to give a causal interpretation to the definition of variable importance, we note that the definitions of variable importance also have a meaning-full interpretation without these extra assumptions.

## 1.2 Summary and organization of article:

In spite of the fact that a regression  $E_P(Y | A, W)$  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 will 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 in van der Laan and Robins (2002), this path-wise



differentiability allows us to construct double robust 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 of  $A$ , given  $W$ , and the regression  $E(Y | A, W)$ . This estimator of marginal variable importance is double robust w.r.t. to miss-specification 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, 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 Section 3.

In Section 4 we focus on estimation of  $W$ -adjusted variable importance. In order to deal with the curse of dimensionality, we redefine  $W$ -adjusted variable importance as the following parameter on the nonparametric model defined as

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

where  $m(a, W | \beta(P))$  is the projection of  $E_P(Y | A = a, W) - E_P(Y | A = 0, W)$  onto a working model  $\{m(a, W | \beta) : \beta\}$ . In this manner, one creates a smoothed version of the original measure of  $W$ -adjusted variable importance. For discrete  $A$ , we establish the wished path-wise differentiability of  $\beta(P)$ , and develop double robust locally efficient estimators of  $\beta_0 = \beta(P_0)$ . We show that this double robust locally efficient estimator of  $W$ -adjusted variable importance can be represented as a simple least squares estimator, which is thus practically very appealing. In addition, as we will show, this insight teaches us also that, given a user supplied machine learning/ data adaptive regression algorithm, it provides us immediately with a data adaptive learning methodology, involving model selection for  $m(\cdot | \beta)$ , for the purpose of learning the original  $W$ -adjusted variable importance  $E(Y | A = a, W) - E(Y | A = 0, W)$ .

Subsequently, for a subset  $V \subset W$  of co-variables, in Section 4 we will

obtain the analogue results for the more general  $V$ -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 methodology for statistical inference for marginal and  $V$ -adjusted variable importance (for any  $V \subset W$ ) in the case that  $A$  is discrete. In Section 5 we actually 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  $P \rightarrow \beta(P)$  (and the corresponding  $W$ -adjusted variable importance parameter) is now path-wise differentiable for general (i.e, continuous) variables  $A$ . Again, this allows us now to develop the class of all double robust locally efficient estimators, and provide the test of the null hypothesis  $H_0 : \theta_0(a, W) = 0$ , and inference. Again, we will generalize these results to  $V$ -adjusted variable importance.

In Section 6 we provide a road map for prediction, variable importance, and multiple testing for variable importance corresponding with our newly introduced methodology. 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 inspired by the analogues of causal effects of time-dependent treatment variables in causal inference.

In the next section we describe the current approach (applied to our definition of variable importance) for obtaining variable importance in prediction based on substitution estimators, so that it can be contrasted to our proposed methodology.

## 2 Nonparametric model: Likelihood based estimator of variable importance.

The definition of marginal variable importance suggests a substitution estimator

$$\psi_n(a) = E_{P_n^*} \{\theta_n(a, W) - \theta_n(0, W)\},$$

where  $\theta_n$  is an estimator of the true regression function  $\theta_0(A, W) = E_0(Y | A, W)$ , and  $P_n^*$  is the empirical estimate of  $P_0^* = P^*(P_0)$ , or, if  $P_0^*$  is known, then  $P_n^* = P_0^*$ . Similarly,  $\theta_n$  provides directly an estimate of  $W$ -adjusted variable importance:

$$\psi_n(a, W) = \theta_n(a, W) - \theta_n(0, W),$$

and  $V$ -adjusted variable importance

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

where  $\hat{E}$  stands for the estimated regression of  $\psi_n(a, W_i)$  on  $V_i$ ,  $i = 1, \dots, n$ . We refer to this estimator as a likelihood based estimator because these estimators have the same form a maximum likelihood estimator of  $\psi_0$  would have. A maximum likelihood estimator of  $\psi_0$  would involve substituting a maximum likelihood estimator of  $P_{Y|A,W}$  based on the likelihood  $\prod_i P(Y_i | A_i, W_i)$  (according to a model) and map this maximum likelihood estimator in the estimator  $\theta_n$  and thereby  $\psi_n$ .

Though maximum likelihood estimation is efficient in parametric models, in a nonparametric model, estimation of  $\theta_0$  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 with minimal mean squared error w.r.t.  $\theta_0$ . Because  $\Psi(P)(a)$  is actually a path-wise differentiable parameter of  $P$  (as shown in the next section for discrete variables  $A$ ) and  $\Theta(P)(A, W) = E_P(Y | A, W)$  is a very non-smooth parameter, a bias-variance trade-off for the purpose of estimation of  $\theta_0$  implies a completely wrong bias-variance trade-off for estimation of  $\psi_0(a)$ : specifically, one should use an over-fitted  $\theta_n$  so that the squared-bias of  $\psi_n$  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  $\theta_0$  is problematic, and requires non-standard model selection techniques as presented in van der Laan and Rubin (2005). In the latter article we present so called estimating function based cross-validation methodology. This methodology would use the optimal estimating function for  $\psi_0$  derived from the efficient influence curve of  $\Psi$  at  $P_0$ , which is presented in the next section and used to directly construct a double robust locally efficient estimator of  $\psi_0$ .

For the interested reader, we present here the details of this method 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  $\psi_0$  as presented in the next section, which is indexed by nuisance parameter values  $\theta$  for  $E(Y \mid A, W)$  and  $\Pi$  for  $P(A \mid W)$ . In addition, given a collection of candidate estimators  $\hat{\Theta}_h$  indexed by fine tuning parameter  $h$ , let

$$\hat{\Psi}_h(P_n) = E_{P_n^*} \{ \hat{\Theta}_h(P_n)(a, W) - \hat{\Theta}_h(P_n)(0, W) \},$$

be the corresponding substitution estimator of  $\theta_0$ . 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  $\psi_0$ :

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

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  $(P_{n, B_n}^0)$  (i.e.,  $O_i$  is an element of training sample if  $B_n(i) = 0$ ) and validation sample  $(P_{n, B_n}^1)$  (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  $\hat{\Psi}_h(P_{n, B_n}^0)$  to  $\psi_0$  w.r.t.  $h$ .

### 3 Discrete variable, nonparametric model, 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  $W = (W_1, W_2)$  and let  $dP_0^* = dP_{W_2, 0} dP_{W_1|W_2}^*$  be a distribution of  $W$  whose marginal of  $W_2$  equals the true marginal distribution of  $W_2$ , and the conditional distribution of  $W_1$ , given  $W_2$  equals a conditional distribution  $P_{W_1|W_2}^*$  which is possibly known or equal to the true conditional distribution of  $W_1$ , given  $W_2$ . Suppose that  $A$  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  $\Psi(P)(a) = E_{P^*}\{E_P(Y | A = a, W) - E_P(Y | A = 0, W)\}$  be the parameter of interest. If  $P_{W_1|W_2}^*$  is known, then the efficient influence curve/canonical gradient of this parameter is given by:

$$IC^*(O | P_0) = \int \{\theta_0(a, W) - \theta_0(0, W)\} dP^*(W_1 | W_2) - \psi_0 \\ + \frac{dP^*(W_1 | W_2)}{dP_0(W_1 | W_2)} \left\{ \frac{I(A = a)}{\Pi_0(a | W)} (Y - \theta_0(a, W)) - \frac{I(A = 0)}{\Pi_0(0 | W)} (Y - \theta_0(0, W)) \right\},$$

where

$$\theta_0(a, W) \equiv E_{P_0}(Y | A = a, W) \\ \Pi_0(a | W) = P(A = a | W).$$

If  $P_0^* = P_0$ , then the efficient influence curve/canonical gradient of this parameter is given by:

$$IC^*(O | P_0) = (\theta_0(a, W) - \theta_0(0, W)) - \psi_0 \\ + \left\{ \frac{I(A = a)}{\Pi_0(a | W)} (Y - \theta_0(a, W)) - \frac{I(A = 0)}{\Pi_0(0 | W)} (Y - \theta_0(0, W)) \right\}.$$

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

$$(O, \psi, \theta, \Pi, \eta) \rightarrow D(O | \psi, \theta, \Pi, \eta) \\ \equiv \int \{\theta(a, W) - \theta(0, W)\} dP^*(W_1 | W_2) - \psi \\ + \frac{dP^*(W_1 | W_2)}{\eta(W_1 | W_2)} \left\{ \frac{I(A = a)}{\Pi(a | W)} (Y - \theta(a, W)) - \frac{I(A = 0)}{\Pi(0 | W)} (Y - \theta(0, W)) \right\},$$

where  $\eta$  represents a candidate (i.e., parameter value) for the conditional distribution of  $W_1$ , given  $W_2$ . Let  $D(O | \psi, \theta, \Pi)$  be the estimating function

for the case that  $P_0^* = P_0$  which does thus not depend on the nuisance parameter  $P_{W_1|W_2}$ . We have the following double robustness result for the unbiasedness of these two estimating functions.

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

$$E_{P_0}D(O | \psi_0, \theta, \Pi, \eta) = 0 \text{ if either } \theta = \theta_0 \text{ or } (\Pi = \Pi_0, \eta = \eta_0).$$

Similarly, in the case  $P_0^* = P_0$ , we have

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

**Proof.** We have

$$\begin{aligned} P_0D(\cdot | \psi_0, \theta, \Pi, \eta) &= \int \{\theta(a, W) - \theta(0, W)\}dP_0^*(W) - \psi_0 \\ &+ \int \frac{dP^*(W_1 | W_2)}{\eta(W_1 | W_2)} \times \\ &\left( \frac{\Pi_0(a | W)}{\Pi(a | W)}(\theta_0(a, W) - \theta(a, W)) - \frac{\Pi_0(0 | W)}{\Pi(0 | W)}(\theta_0(0, W) - \theta(0, W)) \right) dP_0(W). \end{aligned} \tag{1}$$

If  $\eta = \eta_0$ , then the last expression (1) can be written as:

$$\begin{aligned} &\int \{\theta(a, W) - \theta(0, W)\}dP_0^*(W) - \psi_0 + \\ &\int \left( \frac{\Pi_0(a|W)}{\Pi(a|W)}(\theta_0(a, W) - \theta(a, W)) - \frac{\Pi_0(0|W)}{\Pi(0|W)}(\theta_0(0, W) - \theta(0, W)) \right) dP_0^*(W). \end{aligned}$$

Now, note that, if  $\Pi = \Pi_0$ , then this equals

$$\int \{\theta_0(a, W) - \theta_0(0, W)\}dP_0^*(W) - \psi_0 = 0.$$

This proves the unbiasedness if  $(\eta = \eta_0, \Pi = \Pi_0)$ . Suppose now  $\theta = \theta_0$  only. Then (1) equals 0 as well. This proves the unbiasedness at  $\theta = \theta_0$ . This completes the proof.  $\square$

We can construct a double robust locally efficient estimators by solving the corresponding estimating equation at estimated nuisance parameters. Specifically, given estimators  $\Pi_n, \eta_n, \theta_n$  of  $\Pi_0, P_{W_1|W_2,0}$ , and  $\theta_0$ , we estimate  $\psi_0$  with

$$\hat{\Psi}(P_n) = P_nD(\cdot | \theta_n, \Pi_n, \eta_n),$$

where

$$\begin{aligned}
 D(O \mid \theta, \Pi, \eta) &\equiv D(O \mid \psi, \theta, \Pi, \eta) + \psi \\
 &= \int \{\theta(a, W) - \theta(0, W)\} dP^*(W_1 \mid W_2) \\
 &+ \frac{dP^*(W_1 \mid W_2)}{\eta(W_1 \mid W_2)} \left\{ \frac{I(A = a)}{\Pi(a \mid W)} (Y - \theta(a, W)) - \frac{I(A = 0)}{\Pi(0 \mid W)} (Y - \theta(0, W)) \right\}
 \end{aligned}$$

is simply the estimating function without  $\psi$  subtracted. Similarly, in the case that  $P_0^* = P_0$ , the solution of the estimating equation is given by

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

where

$$\begin{aligned}
 D(O \mid \theta, \Pi) &\equiv \theta_0(a, W) - \theta_0(0, W) \\
 &+ \left\{ \frac{I(A = a)}{\Pi_0(a \mid W)} (Y - \theta_0(a, W)) - \frac{I(A = 0)}{\Pi_0(0 \mid W)} (Y - \theta_0(0, W)) \right\}. \quad (2)
 \end{aligned}$$

We note that the latter estimator can be written as:

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

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

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

### 3.1 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  $\theta_0$ . 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 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

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

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  $\Pi(\cdot | T_{\Pi}(P_0))$  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 it is consistent and asymptotically linear if one of the working models is correctly specified.

### 3.2 Inference and testing for 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  $IC_1(O) \equiv D(O | \psi_0, \theta^*, \Pi_0)$ . 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  $\hat{IC}_1(O) \equiv D(O | \psi_n, \theta_n, \Pi_n)$ . 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  $\psi_n$ .

## 4 Discrete $A$ , nonparametric model, $V$ -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  $W$ . In order to deal with the curse of dimensionality one could either assume a model for  $V$ -adjusted variable importance, or define the parameter of interest as a projection of the true  $V$ -adjusted variable importance on a working model. In this section we describe the latter methodology, while the model based approach is presented in the next section. We recall that  $W$ -adjusted variable importance is a special case of the  $V$ -adjusted variable importance by setting  $V = W$ .

### 4.1 $V$ -adjusted variable importance.

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

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

be the parameter of interest, and let the model for  $P_0$  be nonparametric. We note that  $m(V | \beta_0)$  defines a working model based projection of the true  $V$ -adjusted 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  $P(A = a | W)P(A = 0 | W) > 0$ ,  $P_0$ -a.e. The canonical gradient for  $P \rightarrow \beta(P)$  at  $P_0$  in the nonparametric model is given by:*

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

where  $D(O | \theta_0, \Pi_0)$  is defined in (2), and

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

Note that, if  $m(V | \beta)$  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  $-P_0 \frac{d}{d\beta_0} m(\cdot | \beta_0) \frac{d}{d\beta_0} m(\cdot | \beta_0)^\top$ . We derived this expression for the efficient influence curve by deriving the influence curve of the estimator  $\arg \min_{\beta} P_n(D(\cdot | \theta_n, \Pi_0) - m(\cdot | \beta))^2$  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

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

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  $P(A = a | W)P(A = 0 | W) > 0$   $P_0$ -a.e. We have

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

**Proof.** We have

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

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

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

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

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

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

Thus,  $\beta_n$  can be computed with standard least squares regression by regressing the outcome  $D(O_i | \theta_n, \Pi_n)$  on  $V_i$  using the model  $m(V | \beta)$ . 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

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

## 4.2 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  $(D(O_i | \theta, \Pi), V_i)$ ,  $i = 1, \dots, n$ , treating  $D(O_i | \theta, \Pi)$  as outcome.

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

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

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

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

This is shown by first noting that  $E_0(D(O | \theta, \Pi) | W) = E_0(Y | A = a, W) - E_0(Y | A = 0, W)$  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,  $D(O_i | \theta_n, \Pi_n)$  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 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  $\Psi(P)(a, V) = E_P(E_P(Y | A = a, W) - E(Y | A = 0, W) | V)$  by defining the loss function for  $\psi$  as

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

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:

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

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  $\Psi_s \subset \Psi$  of the parameter space  $\Psi$  for  $\Psi$  (i.e., functions of  $V$ ), 3) compute subspace specific estimators such as

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

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

$$\hat{S}(P_n) = \arg \min_s E_{B_n} P_{n, B_n}^1 L(\cdot, \hat{\Psi}_s(P_{n, B_n}^0) | \hat{\Theta}(P_{n, B_n}^0), \hat{\Pi}(P_{n, B_n}^0)),$$

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  $P_{n,B_n}^0$ , 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.

### 4.3 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  $\theta_0$ . 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 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

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

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  $\Pi(\cdot | T_{\Pi}(P_0))$  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 it is consistent and asymptotically linear if one of the working models is correctly specified.

### 4.4 Inference and testing for 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  $\psi_n$ .

## 5 General $A$ , model based $V$ -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  $W$ . In order to deal with the curse of dimensionality one could either assume a model for  $V$ -adjusted variable importance, or define the parameter of interest as a projection of the true  $V$ -adjusted variable importance on a working model. In this section we describe the model based approach, while the other approach was presented in the previous section. For pedagogical purposes, we will start with presenting the methodology for  $W$ -adjusted variable importance, though this is a special case of the  $V$ -adjusted variable importance.

COBRA  
A BEPRESS REPOSITORY

Collection of Biostatistics  
Research Archive

## 5.1 W-specific variable importance.

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

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

and let  $\beta_0 = \beta(P_0)$  be the true parameter value. The model  $m(\cdot | \beta)$  should satisfy  $m(0, W | \beta) = 0$  for all  $\beta$  and  $W$ . Contrary to the  $\beta(P)$  in the  $a$ -specific model  $m(W | \beta)$  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  $E(Y | A, W) - E(Y | A = 0, W) = m(A, W | \beta_0)$  for some  $\beta_0$  can be represented as a generalized semi-parametric regression model

$$E(Y | A, W) = m(A, W | \beta) + g(W), \quad (3)$$

where  $g(W)$  is left unspecified, and  $m(0, W | \beta) = 0$  for all  $W, \beta$ . The semi-parametric regression model  $E(Y | A, W) = m(A | \beta) + g(W)$  has been studied in the literature: Newey (1995); Rosenbaum and Rubin (1983); Robins et al. (1992); Robins and Rotnitzky (2001a,b); Yu and van der Laan (2003). Yu and van der Laan (2003) note that this model can be trivially generalized to the model we need here:  $E(Y | A, W) = m(A, W | \beta) + g(W)$ . 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), 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 (2003), and state here the same result for convenience.

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

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

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

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

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

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

where

$$h_{opt}(A, W) = u_0(A, W) - \frac{E_0(u_0(A, W) | W)}{E_0(c_0(A, W) | W)} \times c_0(A, W), \quad (4)$$

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

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

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:

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

where  $\Pi$  denotes a candidate for the conditional distribution of  $A$ , given  $W$ , and  $\theta$  is a candidate for the regression  $E(Y | A, W)$ . The optimal choice  $h_{opt}$  is given in (4). A simple candidate for  $h$  we recommend for practical use is

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

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

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

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

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

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

As pointed out to me by Dan Rubin (UC Berkeley, Biostatistics), if  $\beta \rightarrow m(A, V | \beta)$  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  $B_n(\beta_n) = c_n$ , 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.

## 5.2 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 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

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

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  $\Pi(\cdot | T_{\Pi}(P_0))$  denotes the projection operator onto  $T_{\Pi}(P_0)$  in the Hilbert space  $L_0^2(P_0)$  endowed with inner product  $\langle h_1, h_2 \rangle_{P_0} = E_{P_0} h_1(O) h_2(O)$ .

## 5.3 Inference and testing.

Consequently, in the case that one assumes a correctly specified model for  $\Pi_0$ , one can use as conservative influence curve  $IC_1(O) \equiv -c(h_0)^{-1} D_h(O | \beta_0, \theta^*, \Pi_0)$ . 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.

#### 5.4 V-specific variable importance.

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

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

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

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  $P_0$*

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

*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  $h = (h_1, h_2, \Pi^*)$

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

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

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

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  $h_{\text{opt}}$  does not exist in closed form. We recommend the following choice  $h^*$ :

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

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  $h_{\text{opt}}$  (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.

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

Given a model for  $\Psi(P)(a, V) = m(a, V | \beta(P))$ , 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  $a = 0$ :

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

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

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

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  $\psi_0 = \arg \min_{\psi} \Theta(\psi, P_0 | \theta, \Pi)$  if either  $\theta = \theta_0$  or  $\Pi = \Pi_0$ .

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  $\Psi_s \subset \Psi$  of the parameter space  $\Psi$  for  $\Psi$  (i.e., functions of  $a, V$  which equal 0 at  $a = 0$ ), 3) compute subspace specific estimators such as

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

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

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

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

For example, if the subspace  $\Psi_s$  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  $m(\cdot \mid \beta)$ ), and calculating the corresponding double robust estimator of the unknown coefficients (i.e.,  $\beta_n$ ), as presented in the previous subsection.

## 5.6 Marginal variable importance.

We note that an estimator of  $\beta_n$  results immediately in a corresponding estimator of marginal variable importance

$$\Psi(P)(a) = E_{P^*} m(a, V \mid \beta(P)).$$

Specifically, given an estimator  $\beta_n$  of  $\beta_0$ , and the empirical estimate  $P_n^*$  of  $P_0^*$ , we have

$$\psi_n(a) = E_{P_n^*} m(a, V \mid \beta_n).$$

In addition, by a simple delta-method it also follows that asymptotic linearity of  $\beta_n$  implies the asymptotic linearity of  $\psi_n$  with known influence curve. For example, if  $P_n^*$  is the empirical distribution of  $P_0$  (i.e.,  $P_0^* = P_0$ ), then

$$\psi_n(a) - \psi_0(a) \approx (P_n - P_0)m(a, \cdot \mid \beta_0) + P_0 \frac{d}{d\beta_0} m(a, \cdot \mid \beta_0)^\top (\beta_n - \beta_0),$$

so that the influence curve of  $\psi_n(a)$  equals

$$IC_a(O \mid P_0) = m(a, V \mid \beta_0) - \psi_0(a) + P_0 \frac{d}{d\beta_0} m(a, \cdot \mid \beta_0)^\top IC_\beta(O \mid P_0),$$

where  $IC_\beta(O \mid P_0)$  denotes the influence curve of  $\beta_n$ .

## 6 A road map for data analysis involving prediction.

Consider a sample  $(W_i^*, Y_i)$ ,  $i = 1, \dots, n$ . Let  $\psi_{j0}(a) = E_0\{E_0(Y | A_j = a, W_j^*) - E_0(Y | A_j = 0, W_j^*)\}$  be a marginal variable importance parameter corresponding with a variable  $A_j$  extracted from  $W^*$ , such as a component of  $W^*$ , where  $W^* = (A_j, W_j^*)$  and, in order to avoid identifiability issues, let's assume that the support of  $W^*$  is a Cartesian product of supports for  $A_j$  and  $W_j^*$ ,  $j = 1, \dots, J$ . Similarly, for a user supplied specification of  $V^* \subset W^*$ , let  $\psi_{j0}(a, V^*) = E_0(E_0(Y | A_j = a, W_j^*) - E_0(Y | A_j = 0, W_j^*) | V_j^*)$ , where again  $V^* = (A_j, V_j^*)$  is a decomposition of  $V^*$  in  $A_j$  and a remaining  $V_j^*$ ,  $j = 1, \dots, J$ .

**Prediction:** Using a particular machine learning algorithm one can obtain a fit of the optimal predictor  $E(Y | W^*)$  with corresponding performance assessment using cross-validation. This object will typically have very poor performance due to the curse of dimensionality. But, as argued in this article, by going after specific features of  $E(Y | W^*)$  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  $A_j$ ,  $j = 1, \dots, J$ . 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  $(\psi_{jn}(a), \sigma_{jn}(a), \text{pval}_j(a) : a)$ ,  $j = 1, \dots, J$ .

**Multiple testing for marginal variable importance:** Consider the test-statistic indexed by the possible  $a$  values and  $j = 1, \dots, J$ . 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 method of Benjamini and Hochberg (1995) controlling 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., Fam-

ily 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 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  $IC(O | P_0)$  of  $(\psi_{jn}(a) : a, j)$  to estimate the asymptotic covariance matrix with the empirical covariance matrix of  $\hat{IC}(O_i)$ ,  $i = 1, \dots, n$ .

**$V^*$ -adjusted variable importance.** Given a collection of user supplied values for  $V^*$ , report 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  $m_j(a, V^* | \beta_j)$  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  $(\psi_{jn}(a, V^*), \sigma_{jn}(a, V^*), \text{pval}_j(a, V^*) : a), j = 1, \dots, J$ .

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  $j = 1, \dots, J$ . Given the list of corresponding marginal p-values ( $\text{pval}_j(a, V^*) : a, j$ ), one can use multiple testing methods based on marginal p-values to control the number of false positives. For example, one can apply the 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 might function better than dimension reductions based on marginal associations or principal components.

## 7 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.



## 7.1 Variable importance for time-dependent variables.

Let  $W = (W(0), \dots, W(K)), Y$  be a longitudinal data structure collected over time at time points indexed by  $k = 0, \dots, K + 1$ , where  $Y$  is the final outcome measured at time  $K + 1$ . Decompose  $W(k) = (L(k), A(k))$ , where  $A = (A(k) : k = 0, \dots, K)$  represents a time-dependent component of  $(W(k) : k = 0, \dots, K)$  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  $E(Y | W)$ , 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 might again 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  $A = a = (a(0), \dots, a(K))$ , 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:

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

Here we used the common notation  $\bar{X}(k) = (X(0), \dots, X(k))$  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  $a = (a(0), \dots, a(K))$  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,  $\Phi(P_{a,Y|V}) = E_{P_{a,Y|V}}(Y | V)$  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 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  $E(Y_a - Y_0 | V) = m(a, V | \beta)$ . 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 a time dependent variable.

## 7.2 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.,  $E(Y_a - Y_0 | V)$ ) 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.

## 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, University of California, Berkeley, 2005a. URL [www.bepress.com/ucbbiostat/](http://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. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 8, 2005b. URL [www.bepress.com/sagmb/vol4/iss1/art8](http://www.bepress.com/sagmb/vol4/iss1/art8).
- 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](http://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.
- W.K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 1(4):335–341, 1995. ISSN 1350-7265.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.
- 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 and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(0):143–155, 1992.
- 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, j. and kwon, jaimyoung. *Statistica Sinica*, 11:920–935, 2001a.
- J.M. Robins and A. Rotnitzky. Double robustness in statistical models. *Scandinavian Journal of Statistics*, To appear, 2001b.
- 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](http://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](http://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 M. Petersen. Estimation of direct and indirect causal effects in longitudinal studies. Technical report 155, Division of Biostatistics, University of California, Berkeley, August 2004.

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 Biostatistics, University of California, Berkeley, 2005. URL [www.bepress.com/ucbbiostat/paper180/](http://www.bepress.com/ucbbiostat/paper180/).

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. Measuring treatment effects using semiparametric models. Technical Report 136, Division of Biostatistics, University of California, Berkeley, 2003. URL [www.bepress.com/ucbbiostat/paper136/](http://www.bepress.com/ucbbiostat/paper136/).



COBRA  
A BEPRESS REPOSITORY

Collection of Biostatistics  
Research Archive