Causal inference methods for continuous exposures

by

Iván Leonardo Díaz Muñoz

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark van der Laan, Chair
Professor Alan Hubbard
Professor Jasjeet Sekhon

Fall 2013

Causal inference methods for continuous exposures

Abstract

Causal inference methods for continuous exposures

by

Iván Leonardo Díaz Muñoz

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark van der Laan, Chair

This dissertation is concerned with the definition and optimal estimation of causal parameters in semiparametric models, focusing on parameters that measure the causal effect of a continuous exposure. It is divided in six self-contained chapters.

In chapter 1 we present a histogram-like estimator of a conditional density that uses super learner cross-validation to estimate the histogram probabilities, as well as the optimal number and position of the bins. The conditional density of the exposure given the confounders is the continuous analogous of the propensity score for binary exposures, whose consistent estimation is critical in causal inference problems. The proposed estimator is an alternative to kernel density estimators when the dimension of the problem is large. We demonstrate its applicability to estimation of Marginal Structural Model (MSM) parameters in which an initial estimator of the treatment mechanism is needed. MSM estimation based on the proposed density estimator results in less biased estimates, when compared to estimates based on a misspecified parametric model.

Estimating the causal effect of an intervention on a population typically involves defining parameters in a nonparametric structural equation model in which the treatment or exposure is deterministically assigned in a static or dynamic way. In chapters 2 and 3 we present two examples of the methodology of stochastic interventions, in which we define new causal parameters that take into account the fact that intervention policies can result in stochastically assigned exposures. In chapter 2 we present a parameter that measures the effect of a population intervention in which the exposure distribution is shifted. In chapter 3 we present a parameter measuring the effect of a truncation in the exposure distribution. The statistical parameters that identify the causal parameters of interest are established. Inverse probability of treatment weighting (IPTW), augmented IPTW (A-IPTW), and targeted maximum likelihood estimators (TMLE) are developed. Simulation studies are performed to demonstrate the properties of these estimators, which include the double robustness and efficiency of the A-IPTW and the TMLE, and application examples are presented.

Chapter 4 deals with estimation of the causal dose-response curve. In a non parametric model, if the treatment is continuous, the dose-response curve is not a pathwise differentiable parameter, and no $\sqrt{n}-$consistent estimator is available. However, the risk of a candidate algorithm for

estimation of the dose-response curve is a pathwise differentiable parameter, whose consistent and efficient estimation is possible. In this work, we review the cross-validated augmented inverse probability of treatment weighted estimator (CV A-IPTW) of the risk, and present a cross validated targeted minimum loss based estimator (CV-TMLE) counterpart. These estimators are proven consistent an efficient under certain consistency and regularity conditions on the initial estimators of the outcome and treatment mechanism. We also present a methodology that uses these estimated risks to select among a library of candidate algorithms. These selectors are proven optimal in the sense that they are asymptotically equivalent to the oracle selector under certain consistency conditions on the estimators of the treatment and outcome mechanisms. Because the CV-TMLE is a substitution estimator, it is more robust than the CV-AIPTW against empirical violations of the positivity assumption. This and other small sample size differences between the CV-TMLE and the CV-A-IPTW are explored in a simulation study.

Finally, In chapter 5 we present an application of some of the methods developed in this dissertation, related to prediction and variable importance (VIM) methods for longitudinal data sets containing both continuous and binary exposures subject to missingness. We demonstrate the use of these methods for prognosis of medical outcomes of severe trauma patients, a field in which current medical practice involves rules of thumb and scoring methods that only use a few variables and ignore the dynamic and high-dimensional nature of trauma recovery. Well principled prediction and VIM methods can thus provide a tool to make care decisions informed by the high-dimensional patients physiological and clinical history. Our VIM parameters can be causally interpreted (under causal and statistical assumptions) as the expected outcome under time-specific clinical interventions. The results of the analysis show effects whose size and significance would have been not been found using a naive parametric approach, as well as improvements of up to 0.07 in the AUROC.

Contents

List of Figures

List of Tables

Chapter 1

Conditional density estimation

Conditional probability estimation is one of the most important problems in statistics, as a parameter of interest itself, or as a nuisance parameter that must be estimated, for example in causal inference and missing data problems. The conditional density of a continuous exposure given the covariates is the equivalent of the propensity score for binary outcomes, which has historically played a central role in the estimation of causal effects in observational studies (see e.g., Rosenbaum and Rubin, 1983; R. Mansson, 2007). Therefore, it is not surprising that estimation of causal parameters for continuous exposures requires the specification of an initial estimator of the exposure mechanism. Two examples of causal parameters for continuous treatments that require the specification of an initial estimator for the conditional density of the exposure given the covariates are given by marginal structural models (Robins, Hernan, and Brumback, 2000; Neugebauer and van der Laan, 2007) and causal parameters corresponding to stochastic interventions (Díaz and van der Laan, 2011a). In this chapter we develop a machine learning estimator of a conditional density that can be used to estimate the exposure mechanism, and demonstrate its use in the estimation of MSM parameters. This estimator will be a critical input for most of the methods for causal inference presented in the following chapters.

Analogously to the case of the propensity score for binary outcomes, the consistency of the initial estimator of the exposure mechanism usually determines the consistency of the estimator of the causal parameters of interest. Therefore, the development of tools that provide consistent estimators of the exposure mechanism is of particular interest to the causal inference and biostatistics research community.

Parametric models such as generalized linear models intend to estimate the conditional density of a variable given a set of predictors by assuming a functional form that is known up to a finite-dimensional vector of real parameters. If the assumptions made about the functional form of the conditional density reflect characteristics of the true data generating mechanism, maximum likelihood methods usually yield consistent and efficient estimators of the parameters of the model and consequently of the conditional density (van der Vaart, 1998). However, it is common to find applications in fields such as epidemiology and social studies in which little information about the true data generating mechanism is known, and the researcher does not have enough scientific knowledge to assume a functional form for the conditional density. For such cases, non parametric estimators such as kernel density estimators, which do not assume a pre-specified functional form have been proposed. Kernel estimation was introduced by Rosenblatt (1969), and has been extensively studied in the statistics literature since then. As a remarkable property, under certain conditions on the true density, the univariate kernel density estimator has been proven to have mean integrated square (MISE) error of order $n^{-4/5}$, which is only $n^{-1/5}$ times larger than the MISE of a correctly specified MLE in a parametric parametric model (van der Vaart, 1998). A comprehensive description of univariate and multivariate kernel density estimators and their statistical properties can be found in Wand and Jones (1995) and Scott (1992). The multivariate kernel density estimator can be used to find estimates of the joint densities involved in the definition of the conditional density and compute a plug-in estimator. Nevertheless, unless the number of covariates is very small (Wand and Jones (1995) suggest less than 6) or the sample size is extremely large, these estimators suffer from the curse of dimensionality, and the resulting estimates are highly biased. This is an important issue in causal inference, since the number of confounders is often large.

Cross validation selection from a library of candidates of estimators has been proven to have optimal properties in terms of the risk of the resulting estimator (van der Vaart, Dudoit, and van der Laan, 2006). In particular, the super learner (van der Laan, Polley, and Hubbard, 2007) is a machine learning technique that uses cross-validated risks to choose an optimal estimator among a library defined by the convex hull of a user-supplied list of estimators. Simulations and analytic results about the super learner can be found in van der Laan, Dudoit, and Keles (2004) and van der Laan and Dudoit (2003). One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution).

In section 1.1 we propose a conditional density estimator that starts with a list of histogram-like density estimators indexed by the number of bins, their position, and the choice of an estimator for the histogram probabilities, and then uses the super learner to find the optimal estimator in the library given by the convex hull of this list of candidate estimators. We use the super learner itself to estimate the histogram probabilities of each of the estimators in the initial list.

A review of marginal structural models as described by Neugebauer and van der Laan (2007) is provided in section 1.2, as well as three MSM estimators that require an initial estimator of the exposure mechanism. The performance of our method in the context of MSM estimation is assessed through a simulation study in which the three estimators described in section 1.2 are computed under three different estimators of the exposure mechanism: a correctly specified parametric model, an incorrectly specified parametric model and our super learner based estimator. The results of this simulation are presented in section 4.4, and section 1.4 includes an application example of in which the causal effect of physical activity on all cause mortality is defined through an MSM. Finally, section 3.5 provides some concluding remarks and directions of future research.

1.1 Density estimator

Let A be a random variable representing an exposure of interest, and let W be a random vector containing a set of covariates confounding the causal relationship between A and an outcome Y. We are interested in finding an estimator of the exposure mechanism $g_0(A|W)$ (i.e., the true conditional density function of A given W). Such estimator will be used in the next sections to compute different estimators of causal effects defined by marginal structural models.

As explained in the introduction, we will use the super learner to choose a convex combination of estimators among a library of candidates consisting of histogram density estimators defined by hazard functions. In the following subsections we will define the super learner, the candidate estimators in the library, and present the cross validated estimator of the conditional density.

Super learner

Consider the usual setting in which we observe n identically distributed copies $O_i,\ i=1,\dots,n$ of the random variable $O=(W,A,Y)\sim P_0$. Super learner deals with estimation of parameters $\psi_0(O)$ defined as the minimizer of a loss function $L(O,\psi)$ over some parameter space Ψ. This is $\psi_0 = \arg\min_{\psi\in\Psi} E_0 L(O,\psi)$. For example, regression ($\psi_0(O)=E_0(Y|A,W)$) and conditional density estimation ($\psi_0(O)=g_0(A|W)$) problems can be formulated in this way by using loss functions $L(O,\psi)=\{Y-\psi(A,W)\}^2$ and $L(O,\psi)=-\log\{\psi(A,W)\}$, respectively.

An estimator $\hat{\Psi}$ of ψ_0 can be seen as a mapping that takes the empirical distribution P_n and maps it into an estimate. $\hat{\Psi}(P_n)$ is then the estimator based on the entire sample, and its risk can be computed as

$$R(\hat{\Psi},P_0)=\int L\{o,\hat{\Psi}(P_n)\}dP_0(o).$$

The true risk of an estimator depends on P_0, and is therefore an unknown quantity that needs to be estimated. A first option is to use a plug-in estimator in which P_n is used instead of P_0. If the space Ψ is very large, this plug-in estimator of the risk will favor estimators $\hat{\Psi}$ that over-fit the data. Instead, super learner provides an algorithm that uses a v-fold cross validated risk estimate to choose the best estimator of ψ_0.

Let $s\in\{1,\dots,S\}$ index a random sample split into a validation sample $V(s)\subset\{1,\dots,n\}$ and a training sample $T(s)=\{V(s)\}^c$. Here we note that the union of the validation samples equals the total sample: $\cup_{s=1}^{S}V(s)=\{1,\dots,n\}$, and the validations samples are disjoint: $V(s_1)\cap V(s_2)=\emptyset$ for $s_1\neq s_2$. Let $P_{T(s)}$ be the empirical distribution of the training sample s. The cross validated estimator of the risk is given by the following expression, in which the parameter is estimated on a training set and the risk is estimated in the corresponding validation set:

$$\frac{1}{S}\sum_{s=1}^{S}R\{\hat{\Psi}(P_{T(s)}),P_{V(n)}\}=\frac{1}{S}\sum_{s=1}^{S}\int L\{o,\hat{\Psi}(P_{T(s)})\}dP_{V(s)}(o). \tag{1.1}$$

Assume that we have a list of candidate estimators $\hat{\Psi}_j : j\in J$. The discrete super learner is defined as the estimator in this list for which the cross validated risk in (1.1) is the smallest. Consider now a library of candidate estimators given by all possible convex linear combinations of the candidates $\hat{\Psi}_j$. It can be shown (van der Laan, Polley, and Hubbard, 2007) that the candidate in this library with the smallest cross validated risk is be given by

$$\hat{\Psi}(P_n)(O) = \sum_{j\in J} \beta_j \hat{\Psi}_j(P_n)(O),$$

where

$$\beta = (\beta_1, \ldots, \beta_J) = \arg\min_{\beta} \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i\in V(s)} L\left\{ O_i,\ \sum_{j\in J} \beta_j \hat{\Psi}_j(P_{T(s)}) \right\}, \tag{1.2}$$

subject to $\sum_{j\in J} \beta_j = 1$ and $\beta_j \geq 0$ for all $j \in J$. Here n_s denotes the size of the validation sample s.

Candidates

Consider a sequence of values $\alpha_0, \ldots, \alpha_k$ that span the range of A and define k bins. Every candidate in our library of conditional density estimators of $g_0(A|W)$ is given by the following expression:

$$g_{n,\alpha}(P_n)(a|W) = \frac{Pr_n(A \in [\alpha_{t-1}, \alpha_t)|W)}{\alpha_t - \alpha_{t-1}}, \text{ for } \alpha_{t-1} \leq a < \alpha_t, \tag{1.3}$$

where we note that the choice of the values α_t $(t = 0, \ldots, k)$ implies defining the number and position of the bins. Here Pr_n denotes an estimator of the true probability $Pr(A \in [\alpha_{t-1}, \alpha_t)|W)$ obtained through a hazard specification and use of a model for binary variables in a pooled repeated measures dataset, as explained below. Note that we consider the estimator in (1.3) as a mapping that takes the empirical distribution P_n and maps it into an estimate of the conditional density of A given W, this notation will be helpful later in the section when we define the cross-validated estimator. Note also that

$$Pr(A \in [\alpha_{t-1}, \alpha_t)|W) = Pr(A \in [\alpha_{t-1}, \alpha_t)|A \geq \alpha_{t-1}, W) \times \prod_{j=1}^{t-1} \{1 - Pr(A \in [\alpha_{j-1}, \alpha_j)|A \geq \alpha_{j-1}, W)\}.$$

The likelihood for model (1.3) is now proportional to

$$\prod_{i=1}^{n} Pr(A_i \in [\alpha_{t-1}, \alpha_t)|W) = \prod_{i=1}^{n} \left[\prod_{j=1}^{t-1} \left\{ 1 - Pr(A_i \in [\alpha_{j-1}, \alpha_j)|A_i \geq \alpha_{j-1}, W_i) \right\} \right] \times Pr(A_i \in [\alpha_{t-1}, \alpha_t)|A_i \geq \alpha_{t-1}, W_i),$$

which corresponds to the likelihood of a binary variable in a repeated measures data set in which the observation of subject i is repeated as many times as intervals $[\alpha_{t-1}, \alpha_t)$ are before the interval to which A_i belongs, and the binary variables indicating $A_i \in [\alpha_{t-1}, \alpha_t)$ are recorded. Possible estimators for the probabilities

$$Pr(A \in [\alpha_{t-1}, \alpha_t)|A \geq \alpha_{t-1}, W) \tag{1.4}$$

include the following logistic model with only main terms:

$$\operatorname{logit}\{Pr(A \in [\alpha_{t-1}, \alpha_t) | A \geq \alpha_{t-1}, W)\} = \sum_{j=1}^{k} \gamma_j I_{[\alpha_{j-1},\infty)}(A) + \sum_{l=1}^{p} \theta_l W_l, \tag{1.5}$$

where we assume the dimension of W is p, and $I_{[\alpha_{j-1},\infty)}(A)$ denotes an indicator of $A \in [\alpha_{j-1}, \infty)$. Another candidate might be given by a logistic model including double interaction terms. In general, any estimator that has the potential of providing an accurate representation of the underlying true data generating mechanism can be postulated as a candidate for estimation of (1.4), including a super learner algorithm that takes all available candidate estimators and finds an optimal convex combination of them. Each candidate estimator in (1.3) is now indexed by choice of the values α_t and choice of an algorithm for estimating (1.4).

The only detail missing in order to completely define a library of estimators is a clever way to choose the most convenient locations for the bins (for fixed k), which will be determined by a parameter c defined below.

Denby and Mallows (2009) describe the histogram as a graphical descriptive tool in which the location of the bins can be characterized by considering a set of parallel lines cutting the graph of the empirical distribution function (ecdf). Specifically, given a number of bins k, the equal-area histogram can be regarded as a tool in which the ecdf graph is cut by $k+1$ equally spaced lines parallel to the x axis, whereas the usual equal-bin-width histogram corresponds to drawing the same lines parallel to the y axis. In both cases, the location of the cutoff points for the bins is defined by the x values of the points in which the lines cut the ecdf. As pointed out by the authors, the equal-area histogram is able to discover spikes in the density, but it oversmooths in the tails and is not able to show individual outliers. On the other hand, the equal-bin-width histogram oversmooths in regions of high density and does not respond well to spikes in the data, but is a very useful tool for identifying outliers and describing the tails of the density.

As an alternative to find a compromise between these two approaches, the authors propose a new histogram in which the ecdf is cut by lines $x + cy = bh$, $b = 1, \ldots, k+1$; where c and h are parameters defining the slope and the distance between lines, respectively. The parameter h identifies the number of bins k. The authors note that $c = 0$ gives the usual histogram, whereas $c \to \infty$ corresponds to the equal-area histogram.

We now define our library of candidate estimators for the conditional density as a collection of estimators in (1.1) by defining values of the vector α through different choices of c and k, and defining an estimator for the probabilities in (1.4). The use of this approach will result in estimators that are able to identify regions of high density as well as provide a good description of the tails and outliers of the density. For the sake of simplicity, we will only consider one candidate for estimation of (1.4): the super learner itself with candidates that may include, for example, the logistic model in (1.5). Since the choice of each α only depends on c and k, the candidate estimators $g_{n,\alpha}$ in (1.3) will now be denoted by $g_{n,j}$, where $j \in J$ is an index identifying a combination of c and k.

Cross validation

Consider the cross validation scheme presented in section 1.1. We define our estimator of the conditional density of A given W as

$$g_n(A|W) = \sum_{j\in J} \beta_j g_{n,j}(A|W),$$

where

$$\beta = (\beta_1, \dots, \beta_J) = \arg\max_{\beta} \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i\in V(s)} \log \sum_{j\in J} \beta_j \, g_{n,j}(P_{T(s)})(A_i|W_i), \tag{1.6}$$

subject to $\sum_{j\in J} \beta_j = 1$ and $\beta_j \geq 0$ for all $j \in J$.

van der Laan, Dudoit, and Keles (2004) proof that this likelihood based cross-validated estimator is asymptotically optimal in the sense that it performs as well as the oracle selector as the sample size increases. Our library of estimators includes all the estimators given by convex combinations of $g_{n,j}(A|W)$ for $j \in J$, and the oracle selector is given by the candidate estimator in the library that minimizes the Kullback-Leibler divergence with respect to the true data-generating distribution

The minimization in (1.6) is carried out by using the augmented Lagrange multiplier method as implemented in the R function `solnp()` (Ghalanos and Theussl, 2010). Technical details about the implementation of this method can be found in Yinyu (1987).

1.2 Marginal structural model estimation

In this section we provide a brief review of the MSM methodology and describe three of the MSM estimators available in the literature. A complete review of MSM methodology can be found in the works of Robins, Hernan, and Brumback (2000), Bryan, Yu, and van der Laan (2003) and Neugebauer and van der Laan (2007).

The consistency of the MSM estimators presented in this section will be used in section 4.4 to assess the performance of the density estimator proposed in section 1.1 when it is used to estimate the exposure mechanism g_0.

Consider an experiment in which an exposure variable A, a continuous or binary outcome Y and a set of covariates W are measured for n randomly sampled subjects. Let $O = (W, A, Y)$ represent a random variable with distribution P_0, and $O_1, \dots, O_n$ represent n i.i.d. observations of O. Assume that the following structural causal model (Pearl, 2000) holds:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(A, W, U_Y), \end{aligned} \tag{1.7}$$

where U_W, U_A and U_Y are exogenous random variables such that $U_A \perp U_Y$ holds, and either $U_W \perp U_Y$ or $U_W \perp U_A$ holds (randomization assumption). The true distribution P_0 of O can be factorized as

$$P_0(O) = P_0(Y|A,W)P_0(A|W)P_0(W), \tag{1.8}$$

where we denote $g_0(A|W) \equiv P_0(A|W)$, $\bar{Q}_0(A,W) \equiv E_0(Y|A,W)$, $Q_{W,0}(W) \equiv P_0(W)$, and $Pf = \int f dP$ for a function f of O. Causal inference parameters are usually defined in terms of the distribution of the counterfactual outcome Y_a that one would obtain in a controlled experiment in which the equation corresponding to A in (3.1) is removed from the SCM and the treatment A is set to be equal to some pre-specified value a deterministically.

Denote $m(a) = EY_a$, the parameter of interest is:

$$\gamma_0 = \arg \min_{\gamma \in B \subset \mathbb{R}^d} \int_{\mathscr{A}} L\{m(a), m_\gamma(a)\} h(a)\, d\mu(a), \tag{1.9}$$

where $\mathscr{A}$ is the support of A, $h(a)$ is a stabilizing weight function, L is a loss function that describes the loss obtained by approximating the true causal curve $m(a)$ with $m_\gamma(a)$, and μ is an appropriate measure (i.e., the Lebesgue or the counting measure). If $L\{m(a), m_\gamma(a)\}$ is a convex function of γ, the parameter can also be defined as the value $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0d})'$ that solves the system of equations

$$\int_{\mathscr{A}} \frac{\partial}{\partial \gamma_j} L\{m(a), m_\gamma(a)\} h(a)\, d\mu(a) = 0;\ j = 1, \ldots, d.$$

The most intuitive loss function to use is

$$\begin{aligned} L\{m(a), m_\gamma(a)\} &= \{m(a) - m_\gamma(a)\}^2 \\ \frac{\partial}{\partial \gamma_j} L\{m(a), m_\gamma(a)\} &= -2\{m(a) - m_\gamma(a)\} \frac{\partial m_\gamma(a)}{\partial \gamma_j}, \end{aligned} \tag{1.10}$$

since it defines the function m_γ as the closest to m in an L_2 sense. Another option for binary outcomes, or outcomes bounded between zero and one is

$$\begin{aligned} L\{m(a), m_\gamma(a)\} &= -m(a) \log\{m_\gamma(a)\} - \{1 - m(a)\} \log\{1 - m_\gamma(a)\} \\ \frac{\partial}{\partial \gamma_j} L\{m(a), m_\gamma(a)\} &= -\frac{m(a) - m_\gamma(a)}{m_\gamma(a)\{1 - m_\gamma(a)\}} \frac{\partial m_\gamma(a)}{\partial \gamma_j}. \end{aligned}$$

In this paper we focus in the estimation of parameters defined in terms of (1.10), but similar calculations can be made for other parameters defined by different loss functions.

Since $m(a)$ is identified as a function of the distribution of the observed data by $E(\bar{Q}_0(a,W))$, the parameter of interest is identified as the value γ_0 that solves

$$\int_{\mathscr{A}} \frac{\partial}{\partial \gamma_j} L\{E(\bar{Q}_0(a,W)), m_\gamma(a)\} h(a)\, d\mu(a) = 0;\ j = 1, \ldots, d. \tag{1.11}$$

Estimators

In this section we describe three possible estimators for the parameter γ_0 of a MSM defined in the previous section. The first estimator is an Inverse Probability of Treatment Weighted (IPTW) estimator which requires a consistent estimator of the exposure mechanism in order to be consistent. The second estimator is an augmented IPTW that solves the efficient influence curve equation and requires initial estimators of $\bar{Q}_0$ and g_0, it is consistent if either of them is consistent, and it is efficient if both are consistent. The third estimator is a targeted maximum likelihood estimator (TMLE) that has the same properties as the A-IPTW, plus additional advantages that include being a substitution estimator and not having multiple solutions.

IPTW

The IPTW estimating function is defined as $D_{IPTW}(O|g) = \{D_{IPTWj}(O|g)\}_{j=1}^{d}$, where

$$D_{IPTWj}(O|g,\gamma) = \{Y - m_\gamma(A)\}\frac{h(A)}{g(A|W)}\frac{\partial m_\gamma(A)}{\partial \gamma_j},$$

and the IPTW estimator is defined as the vector $\gamma_{n,1}$ that solves the IPTW estimating equations

$$\sum_{i=1}^{n} D_{IPTWj}(O|g,\gamma) = 0;\ j = 1,\ldots,d.$$

The IPTW is an asymptotically linear estimator with influence function D_{IPTWj}, therefore the variable $\sqrt{n}(\gamma_{n,1,j} - \gamma_{0j})$ converges in distribution to a random variable distributed as

$$N\{0, P_0 D^2_{IPTWj}(\,\cdot\,|g_0,\gamma_0)\},$$

whose variance can be estimated by $P_n D^2_{IPTWj}(\,\cdot\,|g_n^0,\gamma_{n,1})$, where P_n denotes the empirical measure. van der Laan and Robins (2003) prove that this variance estimator is conservative. We will use notation $D_{IPTW}(O)$ or $D_{IPTW}(O|g,\gamma)$ depending on whether it is necessary to emphasize the dependence on g and γ.

Augmented IPTW

The efficient influence curve $D(O)$ of (1.9) in the non-parametric model can be found through the IPTW estimating function $D_{IPTW}(O)$ as

$$D(O) = D_{IPTW}(O) - \Pi(D_{IPTW}(O)|T_{CAR}),$$

where $D_{IPTW}(O) = \{D_{IPTWj}(O)\}_{j=1}^{d}$, and $\Pi(D_{IPTW}(O)|T_{CAR})$ is the projection of $D_{IPTW}(O)$ into the space $T_{CAR} = \{s(A,W) : E\{s(A,W)|W\} = 0\}$, defined component-wise. Formally,

$$\begin{aligned}\Pi(D_{IPTWj}(O)|T_{CAR}) =& E\{D_{IPTWj}(O)|A,W\} - E\{D_{IPTWj}(O)|W\}\\ =& \{\bar{Q}(A,W) - m_\gamma(A)\}\frac{h(A)}{g(A|W)}\frac{\partial m_\gamma(A)}{\partial \gamma_j}\\ &- \int_{\mathscr{A}} \{\bar{Q}(a,W) - m_\gamma(a)\}\frac{\partial m_\gamma(a)}{\partial \gamma_j} h(a)\, d\mu(a).\end{aligned}$$

Thus, the efficient influence curve is given by $D(O|\bar{Q},g,\gamma) = \{D_j(O|\bar{Q},g,\gamma)\}_{j=1}^{d}$, where

$$D_j(O|\bar{Q},g,\gamma) = \{Y - \bar{Q}(A,W))\}\frac{h(A)}{g(A|W)}\frac{\partial m_\gamma(A)}{\partial \gamma_j} + \\ \int_{\mathscr{A}} \{\bar{Q}(a,W) - m_\gamma(a)\}\frac{\partial m_\gamma(a)}{\partial \gamma_j} h(a)\, d\mu(a), \quad (1.12)$$

and the augmented IPTW estimator is defined as the value $\gamma_{n,2}$ that solves the augmented IPTW estimating equations

$$\sum_{i=1}^{n} D_j(O_i|\bar{Q},g,\gamma) = 0;\ j = 1,\ldots,d.$$

The A-IPTW is also asymptotically linear with influence curve $D_j(O|\bar{Q}_0,g_0,\gamma_0)$. The variable $\sqrt{n}(\gamma_{n,2,j} - \gamma_{0j})$ converges in law to a random variable with distribution $N\{0, P_0 D_j^2(\cdot|\bar{Q}_0,g_0,\gamma_0)\}$, whose variance can be estimated as $P_n D_j^2(\,\cdot\,|\bar{Q}_n^0,g_n^0,\gamma_{n,2})$. van der Laan and Robins (2003) (sections 2.3.7 and 2.7.1) show that inference based on this variance estimator is valid only if g_n^0 is consistent, providing exact inference when $\bar{Q}_n^0$ is consistent, and conservative inference when $\bar{Q}_n^0$ is inconsistent.

Note that the efficient influence curve can be decomposed into three components corresponding to the orthogonal decomposition of the tangent space implied by the factorization (3.2) as:

$$D_j(O) = D_{j1}(O) + D_{j2}(O) + D_{j3}(O),$$

where

$$\begin{aligned} D_{j1}(O) &= D_j(O) - E\{D_j(O)|A,W\} = \{Y - \bar{Q}(A,W)\}\frac{h(A)}{g(A|W)}\frac{\partial m_\gamma(A)}{\partial \gamma_j}, \\ D_{j2}(O) &= E\{D_j(O)|A,W\} - E\{D_j(O)|W\} = 0, \qquad (1.13) \\ D_{j3}(O) &= E\{D_j(O)|W\} - E\{D_j(O)\} = \int_{\mathscr{A}} \{\bar{Q}(a,W) - m_\gamma(a)\}\frac{\partial m_\gamma(a)}{\partial \gamma_j} h(a)\, d\mu(a). \end{aligned}$$

Targeted maximum likelihood estimator

In order to define a targeted maximum likelihood estimator for γ_0, we need first to define three elements: (1) A loss function $L(Q)$ for the relevant part of the likelihood required to evaluate γ_0, which in this case is $Q = (\bar{Q}, Q_W)$. This function must satisfy $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$, where Q_0 denotes the true value of Q; (2) An initial estimator Q_n^0 of Q_0; (3) A parametric fluctuation $Q(\varepsilon)$ through Q_n^0 such that the linear span of $\frac{d}{d\varepsilon}L\{Q(\varepsilon)\}|_{\varepsilon=0}$ contains all the components of the efficient influence curve $D(O)$ defined in (2.5). These three elements are defined below:

Loss Function

As loss function for Q, we will consider $L(Q) = L_Y(\bar{Q}) + L_W(Q_W)$, where $L_Y(\bar{Q}) = Y\log\{\bar{Q}(A,W)\} +$

$(1-Y)\log\{1-\bar{Q}(A,W)\}$ and $L_W(Q_W) = -\log Q_W(W)$. It can be easily verified that this function satisfies $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$.

Parametric Fluctuation

Given an initial estimator Q_n^0 of Q_0, with components $(\bar{Q}_n^0, Q_{W,n}^0)$. We define the fluctuation of Q_n^0 as follows:

$$Q_{W,n}^1(\delta)(W) = \left\{1 + \sum_{j=1}^d \delta_j Z_j(W)\right\} Q_{W,n}^{,}0$$

$$\operatorname{logit} \bar{Q}_n^1(\varepsilon)(A,W) = \operatorname{logit} \bar{Q}_n^0(A,W) + \sum_{j=1}^d \varepsilon_j H_j^0(A,W),$$

where $Z_j(W) = D_{j3}(O)$, and

$$H_j(A,W) = \frac{h(A)}{g(A|W)} \frac{\partial m_\gamma(A)}{\partial \gamma_j}.$$

First of all, note that the MLE of δ is zero. Standard logistic regression software can be used to find the MLE ε_n of ε, and the TMLE as defined by van der Laan and Rubin (2006) is found in the first iteration. From these definitions it follows that $D_j(O) \in < \frac{\partial}{\partial \varepsilon} L\{Q(\varepsilon,\delta)\}|_{\varepsilon=0} + \frac{\partial}{\partial \delta} L\{Q(\varepsilon,\delta)\}|_{\delta=0} >$ $j = 1,\ldots,d$, where $< \cdot >$ denotes linear span.

Initial Estimators

The empirical distribution of W is used as initial estimator of $Q_{W,0}$.

Targeted Maximum Likelihood Estimator

The TMLE of γ_0 is now defined as the value $\gamma_{n,3}$ that solves the equations

$$\int_{\mathscr{A}} \frac{\partial}{\partial \gamma_j} L\{E_{Q_{W,n}} \bar{Q}_n^1(\varepsilon_n)(a,W), m_\gamma(a)\} h(a)\, d\mu(a) = 0;\ j = 1,\ldots,d. \tag{1.14}$$

The variance of $\gamma_{n,3,j}$ can be estimated by $P_n D_j^2(\,\cdot\,|\bar{Q}_n^1, g_n^0, \gamma_{n,3})$, which like the augmented IPTW variance estimator is consistent only if both g_n^0 and Q_n^1 are consistent, is conservative if g_n^0 is consistent but Q_n^1 is not, and is inconsistent in any other case.

1.3 Simulation

Consider the following data generating process

$$\begin{aligned}
W_1 &\sim Unif\{0,1\}.\\
W_2 &\sim Ber\{0.7\}.\\
A &\sim Gamma\{(.3 + 3\log(W_1+1) + 2.2\exp(W_1)W_2)^{-1}, 1\}.\\
Y &\sim Ber\{expit(-1 + .05A - .02AW_2 + .2A\tan(W_1^2) - .02W_1W_2 + .1AW_1W_2)\}.
\end{aligned}$$

We are interested in estimating the parameter defined in (1.9) with

$$m_\gamma(a) = \frac{1}{1+\exp(-\gamma^0 - \gamma^1 a)}, \tag{1.15}$$

and $h(a)$ equal to the marginal density of A. Note that the efficient influence curve calculations made in the previous sections remain valid in this case, and that estimators of g_0 and $Q_{W,0}$ define an estimator of h. The true value of the parameter for this data generating distribution is $\gamma_0^0 = -1.0067$ and $\gamma_0^1 = 0.1520$. Figure 2.1 presents the true counterfactual expectation $m(a)$ as well as the true MSM curve $m_{\gamma_0}(a)$. Since the definition of the MSM parameter involves weighting by the marginal density of A, the approximation of m_{γ_0} to m is almost perfect in areas of high density, at the cost of a poor approximation in the areas in which A has low density.

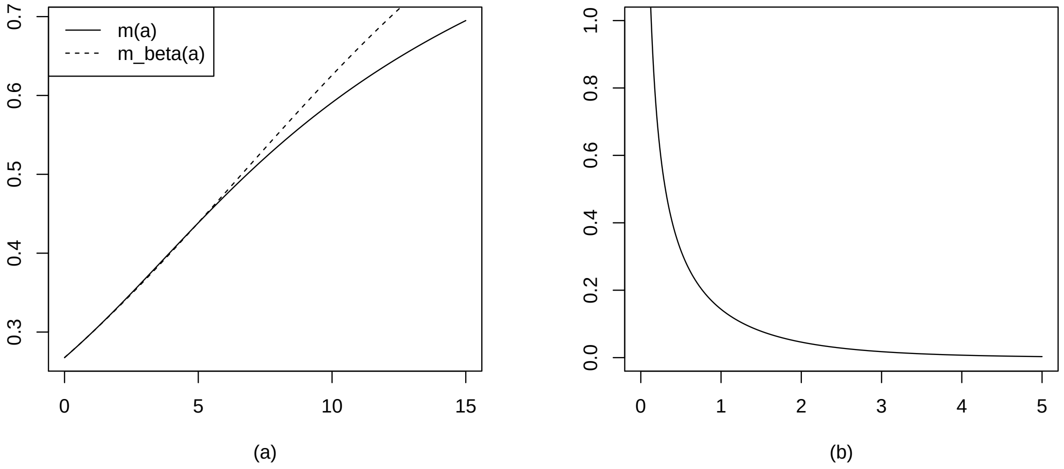


Figure 1.1: (a) True counterfactual expectation and true MSM curve. (b) Marginal density of A.

In order to explore the stability of the estimators described in the previous section when the conditional density estimator of section 1.1 is used as initial estimator for the treatment mechanism, a simulation study was performed. Three different initial estimators were used for the treatment mechanism: (a) correctly specified parametric model, (b) normal linear model with just linear terms, and (c) histogram-like cross-validated estimator of section 1.1; and two different initial estimators were considered for the expectation of Y given A and W: (1) correctly specified parametric model, and (2) logistic regression with only linear terms. The choice of the misspecification of the models performed in (b) and (2) comes from usual practice in parametric modeling in epidemiology, in which for the sake of ease of interpretation and calculation, linear models without interactions are usually assumed.

The prohibitive computational cost of the cross-validation procedure resulting in the proposed conditional density estimator restrained us from using Monte Carlo simulation to asses the properties of the MSM estimator. Instead, we drew a sample of size 10.000 from the true data generating mechanism, and computed the exposure mechanism estimate as well as the three estimates. Figure 1.2 shows the estimates and true value of the conditional densities for two given profiles, obtained by using a list of candidates estimators in (1.3) defined by all combination of values $k = 5,7,9,11,13$ and $c = 800,500,300,100,50,10,0.5,0.01$. From this graph we can see that this estimator is very close to the true data exposure mechanism, which is a surprising fact given that it does not use any knowledge about the true density or the true parametric model.

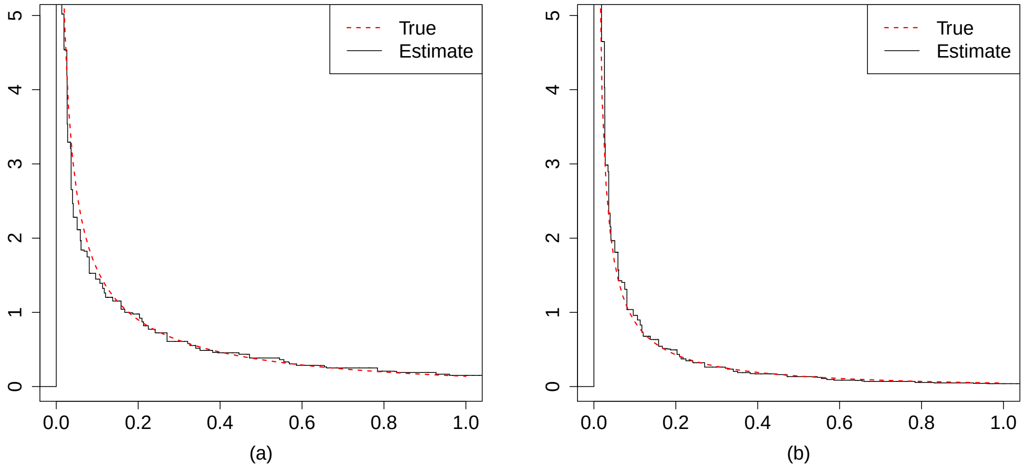


Figure 1.2: (a) Estimated and true density for profile $W_1 = 0.09$, $W2 = 1$. (b) Estimated and true density for profile $W_1 = 0.99$, $W2 = 1$.

Table 1.1 shows the three MSM estimates for model (1.15) along with their standard errors. Given the large sample size, a direct comparison of the estimates with the true value of the parameters provides an approximation to their bias. It is known that (up to positivity assumptions) the TMLE and the A-IPTW are double robust in the sense that they are unbiased if at least one of the initial estimators is consistent. The IPTW requires consistency of the estimator for the treatment mechanism in order to be unbiased.

Misspecification of the parametric model for the treatment mechanism caused a large amount of finite sample bias in the IPTW and A-IPTW estimates, both when the model for $\bar{Q}_0$ is correctly and incorrectly specified. The TMLE, although also biased, remains closer to the true value of the parameter in both cases. The estimates obtained using the histogram-like cross-validated density estimator are as close to the true value of the parameter as the estimates obtained by using a correctly specified model for g_0, showing that this estimator is preferable to parametric models, unless the true model is known to the researcher.

		(a)		(b)		(c)	
		γ^0	γ^1	γ^0	γ^1	γ^0	γ^1
(1)	IPTW	-1.0342 (0.0011)	0.1171 (0.0016)	-1.5406 (0.0015)	0.3634 (0.0049)	-1.0076 (0.0010)	0.1055 (0.0014)
	A-IPTW	-1.0194 (0.0012)	0.1159 (0.0017)	-0.7556 (0.0012)	-0.2522 (0.0024)	-1.0127 (0.0010)	0.1210 (0.0014)
	TMLE	-1.0912 (0.0011)	0.1471 (0.0017)	-0.9814 (0.0015)	0.0935 (0.0046)	-1.0073 (0.0010)	0.1076 (0.0014)
(2)	IPTW	-1.0342 (0.0011)	0.1171 (0.0016)	-1.5406 (0.0015)	0.3634 (0.0049)	-1.0076 (0.0010)	0.1055 (0.0014)
	A-IPTW	-1.0165 (0.0012)	0.1040 (0.0016)	-0.7945 (0.0012)	-0.2118 (0.0025)	-1.0064 (0.0010)	0.0979 (0.0014)
	TMLE	-1.0915 (0.0011)	0.1434 (0.0016)	-0.9764 (0.0014)	0.0656 (0.0044)	-1.0141 (0.0010)	0.1142 (0.0014)

Table 1.1: Parameter estimates for different initial estimators. (a) correctly specified parametric model for g_0, (b) normal linear model for g_0 with only linear terms, (c) histogram-like cross-validated estimator of g_0; (1) correctly specified parametric model for $\bar{Q}_0$, (2) logistic regression with just linear terms for $\bar{Q}_0$. Standard errors in parentheses.

1.4 Application

With the objective of demonstrating the use of the exposure mechanism estimator provided in section 1.1, we revisit the problem analyzed by Bembom and van der Laan (2007) and Díaz and van der Laan (2011a) of assessing the extent to which physical activity in the elderly is associated with reductions in cardiovascular morbidity and mortality, and improvement in, or prevention of metabolic abnormalities. Tager, Hollenberg, and Satariano (1998) followed a group of people over 55 years of age living around Sonoma, CA, over a time period of about ten years as part of a longitudinal study of physical activity and fitness (Study of Physical Performance and Age Related Changes in Sonomans - SPPARCS). The goal in analyzing the data that were collected as part of this study is to examine the effect of baseline vigorous LTPA (Leisure Time Physical Activity) on subsequent five-year all-cause mortality.

We use the same measure of LTPA used by Bembom and van der Laan (2007), which is a continuous score based on the number of hours that the participants were engaged in vigorous physical activities such as jogging, swimming, bicycling on hills, or racquetball in the last seven days, and the standard intensity values in metabolic equivalents (MET: Metabolic Equivalent of Task) of such activities, where one MET is approximately equal to the oxygen consumption required for sitting quietly. The primary confounding factors that we adjust for are described in Table 2.4. Age and gender are natural confounders, and the rest of the variables intend to account for the subject's underlying level of general health. Of the 2092 subjects enrolled in the SPPARCS study, 40 were missing information in at least one of this variables; our analysis is based on the remaining 2052 subjects.

Variable	Description
Gender	Female Male
Age	Age in years
Health	Self-rated health status: Excellent Fair Poor
NRB	Score of self-reported physical functioning rescaled between 0 and 1
Card	Previous occurrence of any of the following cardiac events: Angina, myocardial infarction, congestive heart failure, coronary by-pass surgery, and coronary angioplasty
Chron	Presence of any of the following chronic health conditions: stroke, cancer, liver disease, kidney disease, Parkinson's disease, and diabetes mellitus
Smoking	Never smoked Current smoker Ex-smoker
Decline	Activity decline compared to 5 or 10 years earlier

Table 1.2: Confounders.

In the sequel of this section, the vector containing the confounders will be denoted by W, the continuous MET score by A, and the indicator of five-year all-cause mortality by Y. We are interested in summarizing the causal relationship between LTPA and all cause mortality based on the MSM provided in (1.15) through estimation of the parameters involved.

Figure 2.1 shows two contrasting estimated densities $g_n(A|W)$ for different profiles W, in which a subject with better general health status is more likely to have higher levels of physical activity. As pointed out before, this methodology allows the detection of high density areas in the exposure mechanism, like the one detected at zero in Figure 1.3 (a). This spike appears because this is a "zero-inflated" exposure, in which a large proportion of the population do not practice any amount of physical activity.

As initial estimator of $\bar{Q}_0$ we also used the super learner (van der Laan, Polley, and Hubbard, 2007). Table 2.5 shows the candidates used, their cross-validated risks, and their coefficients in the final super learner predictor. In order to get a consistent estimator of $\bar{Q}_0$ the library of candidate estimators should be as large as possible. Since this is an illustrating example, we allow ourselves to use this small library. Table 1.4 shows the three estimated values for each of the two param-

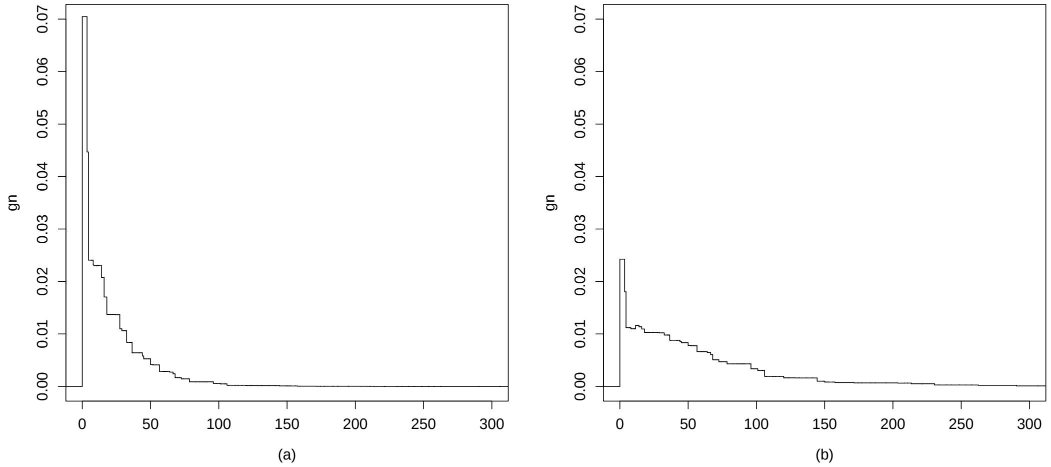


Figure 1.3: Estimated conditional density of A given the profiles: (a) age = 77, gender = female, health = fair, nrb = 0.9, card = no, smoke = ex-smoker, decline = yes, chron = yes; and (b) age = 71, gender = male, health = good, nrb = 0.88, card = no, smoke = never smoked, decline = no, chron = no

	Cross-validated Risk	Coef.
GLM main effects	0.1079	0.0000
GLM main eff. and two way interactions	0.1143	0.0835
GAM degree 2	0.1073	0.0000
GAM degree 3	0.1071	0.9165
Bayes GLM main effects	0.1078	0.0000

Table 1.3: Super learner output for estimation of $\bar{Q}_0$.

eters defined by the MSM in (1.15). Computation of these estimates required (as explained in section 1.2) an initial estimator of the exposure mechanism. The simulation in section 4.4 showed that misspecification of a parametric model for the exposure mechanism can lead to a substantial amount of bias in the MSM estimates, and that the use of the density estimator presented in section 1.1 is preferred when the true exposure mechanism is unknown. Although the three estimators for γ_1 differ in magnitude, all of them agree that there is a small protective effect of physical activity on all cause mortality, although that effect is not statistically significant, which does not mean that it is not relevant.

	IPTW	A-IPTW	TMLE
γ_0	-1.7484 (0.4274)	-1.6365 (0.2650)	-1.5893 (0.1311)
γ_1	0.0042 (0.0022)	0.0026 (0.0015)	0.0013 (0.0011)

Table 1.4: Estimates of the MSM parameters (1.2) for the phisycal activity data.

1.5 Conclusion

In this chapter we develop a conditional density estimator based on a convex linear combination of candidate histogram estimators indexed by the position and location of the bins, in which the histogram probabilities are estimated by using the super learner. We develop and use this estimator in the context of estimation of causal MSM parameters. An application in the context of stochastic intervention parameters can be found in Díaz and van der Laan (2011a). Even though these applications are both related to estimation of causal effects, the conditional density estimation technique here described is of general applicability, and can also be used as a general machine learning technique for estimation of conditional densities.

Since the estimator proposed is computationally very intense, exhaustive simulations studying its statistical properties as a density estimator imply prohibitive simulation times. However, a small simulation study was performed to show that MSM parameter estimation based on our estimator is preferable to MSM estimation based on a misspecified parametric model. This implies that unless the true exposure generation mechanism is known, the use of our estimator as initial estimator of the exposure mechanism should be preferred.

Finally, the simulation study also showed that for a very large sample size the estimated exposure mechanism is very close to the true exposure mechanism. Such feature of this particular simulation suggests an interesting line of future research in which the analytic conditions under which our estimator is consistent or equipped with an oracle inequality can be established.

Chapter 2

Stochastic interventions: shifting the exposure mechanism

Most causal inference problems are addressed by defining parameters of the distribution of the counterfactual outcome that one would obtain in a controlled experiment in which an exposure variable A is set to some pre-specified value a deterministically. A widely used example of this framework is the causal effect for a binary treatment, in which the expectation of the outcome in a hypothetical world in which everybody receives treatment is compared with its counterpart in a world in which nobody does. Other common way of addressing causal problems consists in considering parameters that reflect the difference between the distribution of a counterfactual outcome in such hypothetical intervened world and the distribution of the actual outcome; these parameters are often referred to as population intervention parameters (Hubbard and van der Laan, 2005).

In order to estimate such exposure-specific counterfactual parameters from observational data, one has to assume that all subjects in the population have a positive probability of receiving the exposure level a under consideration. This assumption is often referred to as experimental treatment assignment (ETA), or positivity assumption and can be highly unrealistic in most cases. Additionally, when the exposure of interest is not a variable that can be directly manipulated (e.g., social or behavioral phenomena), any policy intervention targeting a change in the exposure distribution will result in a population whose exposure is stochastic rather than deterministic, and the causal effect as described in the previous paragraph loses its appeal as a measure of the gain obtained by implementing such a policy.

An example that illustrates these ideas is presented in section 2.5. These data were collected by Tager, Hollenberg, and Satariano (1998) and analyzed by Bembom and van der Laan (2007) with the main goal of assessing the effect of vigorous physical activity on mortality in the elderly. Firstly, as argued by Bembom and van der Laan (2007), ETA assumptions as needed to identify the causal effect of a static treatment are quite unrealistic since health problems are expected to prevent an important proportion of the population from high levels of physical activity. Secondly, it is clear that it is not possible to put in practice a policy in which every subject is enforced to a physical activity level dictated by a deterministic rule. Therefore, any intervention on the population that targets changes in physical activity level will induce a random post-intervention exposure. These and other reasons why deterministic interventions are not always the best approach to estimate causal effects are discussed in Korb et al. (2004) and Eberhardt and Scheines (2006). Korb et al. (2004) define an intervention on a variable A in a causal model as an action that intends to change the distribution of A. This general definition includes as special cases static and dynamic deterministic interventions (through degenerate distributions), but it also allows the definition of the causal effect in terms of a non degenerate distribution, as exploited in this article.

In our example, the question of whether higher levels of leisure-time physical activity (LTPA) cause a reduction in mortality rates in the elderly can be better addressed by considering the effect of a policy that aims to cause an increase in the mean of LTPA, possibly depending on covariates such as health status or socioeconomic level. As we will see in section 2.1, this problem corresponds to considering the effect of an intervention that shifts the location of the treatment mechanism. We focus the discussion on the definition and estimation of the effect of this specific type of interventions.

Despite the previous considerations, current developments and applications have almost exclusively focused on deterministic interventions. Among the few works using stochastic interventions figure Cain et al. (2010), who used a stochastic intervention in the context of comparing dynamic treatment regimes with a grace period; and Taubman et al. (2009a), who considered an intervention in the BMI defined by a truncation of the original exposure distribution.

Other type of stochastic interventions of interest arises in applications in which the interest relies in estimating the effect of a policy that enforces the exposure level below a certain threshold. Such policies can modify the distribution of the exposure in various ways. For example, if a policy that constrains air pollution emissions below a cutoff point is put in place, it is reasonable to think that the probability mass associated with values above that cutoff in the original exposure mechanism will be relocated around the cutoff after the intervention. This is because under such a policy, high-polluting companies will not have any incentive to go below the enforced cutoff point.

Alternative threshold-like interventions can lead to a truncated version of the original density, relocating the mass above the threshold across all values of the exposure distribution below the threshold (as opposed to relocating it in the cutoff point). In fact, as proven by Stitelman, Hubbard, and Jewell (2010b), the intervention obtained by considering a dichotomous version of a continuous treatment and defining a usual static intervention (e.g., the BMI intervention in Taubman et al. (2009a)), corresponds to a stochastic intervention on the original continuous treatment that truncates its density below the value defining the dichotomization.

Our major goal is to introduce stochastic intervention causal parameters as a way of measuring the effect that certain policies have on the outcome of interest. As we will see, estimation of the these parameters requires weaker assumptions than estimation of other causal parameters (e.g., MSM), relaxing assumptions about positivity and consistency of the initial estimators, and thus providing a more flexible way of estimating causal effects. We will start in section 2.1 by defining the parameter of interest, in section 2.2 we present its efficient influence curve, and discuss the double robustness of estimators that solve the efficient influence curve equation. This section also provides the tools for defining the targeted maximum likelihood estimators in section 3.3. In section 2.4 we present a simulation study demonstrating consistency and efficiency properties of the estimators, and in section 2.5 we present an application example.

2.1 Data and parameter of interest

Consider an experiment in which an exposure variable A, a continuous or binary outcome Y and a set of covariates W are measured for n randomly sampled subjects. Let $O=(W,A,Y)$ represent a random variable with distribution P_0, and $O_1,\dots,O_n$ represent n i.i.d. observations of O. Assume that the following NPSEM holds:

$$W=f_W(U_W);\ A=f_A(W,U_A);\ Y=f_Y(A,W,U_Y), \tag{2.1}$$

where U_W, U_A and U_Y are exogenous random variables such that $U_A \perp\!\!\!\perp U_Y$ holds, and either $U_W \perp\!\!\!\perp U_Y$ or $U_W \perp\!\!\!\perp U_A$ holds (randomization assumption). The true distribution P_0 of O can be factorized as

$$P_0(O)=P_0(Y|A,W)P_0(A|W)P_0(W),$$

where we denote $g_0(A|W)\equiv P_0(A|W)$, $\bar{Q}_0(A,W)\equiv E_0(Y|A,W)$, $Q_{W,0}(W)\equiv P_0(W)$, and $Pf=\int f dP$ for a given function f.

Counterfactual outcomes under stochastic interventions are denoted by Y_{P_δ}, and are defined as the outcome of a causal model in which the equation in the NPSEM (3.1) corresponding to A is removed, and A is set equal to a with probability $P_\delta(g_0)(A=a|W)$. The latter is called the intervention distribution, which we allow to depend on the true exposure mechanism g_0. Any stochastic intervention of interest can be defined in this way, and in this chapter we focus the discussion on the intervention distribution:

$$P_\delta(g_0)(A=a|W)=g_0(a-\delta(W)|W), \tag{2.2}$$

for a known function $\delta(W)$. This is a shifted version of the current treatment mechanism, where the shifting value is allowed to vary across strata defined by the covariates. As discussed in section 2.5, one can be interested in the effect of a policy that encourages people to exercise more, leading to a population where the distribution of physical activity is shifted according to certain health and socioeconomic variables. As implicitly stated in (2.2), we will assume that the functional form of the exposure mechanism induced by the intervention differs from the original exposure mechanism only through its conditional expectation given the covariates.

Identification

Let A_{P_δ} denote the exposure variable under the intervened system (i.e., A_{P_δ} is distributed according to $P_\delta(g)$). We have that

$$P(Y_{P_\delta}=y)=\sum_{a\in\mathscr{A}}\sum_{w\in\mathscr{W}}P(Y_{P_\delta}=y|A_{P_\delta}=a,W=w)P_\delta(g)(A=a|W=w)P(W=w),$$

where $\mathscr{A}$ and $\mathscr{W}$ are the support of A and W respectively. From the NPSEM (3.1) we have that $P(Y_{P_\delta}=y|A_{P_\delta}=a,W=w)=P(Y_a=y|A_{P_\delta}=a,W=w)$, where Y_a is the counterfactual outcome when the exposure is set to level a with probability one. Note also that the usual randomization assumption $A \perp\!\!\!\perp Y_a | W$ implies $A_{P_\delta} \perp\!\!\!\perp Y_a | W$, and therefore $P(Y_a = y | A_{P_\delta} = a, W = w) = P(Y_a = y | W = w)$. Under the consistency assumption ($A = a$ implies $Y_a = Y$) the latter quantity is identified by $P(Y = y | A = a, W = w)$. Our counterfactual distribution can be written as

$$P(Y_{P_\delta} = y) = \sum_{a \in \mathscr{A}} \sum_{w \in \mathscr{W}} P(Y = y | A = a, W = w) P_\delta(g)(A = a | W = w) P(W = w).$$

We define the parameter of interest as a mapping $\Psi : \mathscr{M} \to R$ that takes an element in a statistical model $\mathscr{M}$ and maps it into a number in the reals. The true value of the parameter is given by the mapping evaluated at the true distribution $P_0 \in \mathscr{M}$, and is denoted by $\psi_0 = \Psi(P_0)$. Our causal and statistical parameter of interest is then given by

$$E(Y_{P_\delta}) = \Psi(P) = \sum_{A \in \mathscr{A}} \sum_{W \in \mathscr{W}} \bar{Q}(A, W) P_\delta(g)(A | W) Q_W(W). \tag{2.3}$$

Note that this parameter depends only on $Q = (\bar{Q}, g, Q_W)$. Therefore, in an abuse of notation, we will use the expressions $\Psi(Q)$ and $\Psi(P)$ interchangeably.

2.2 Efficient influence curve

In this section we derive the efficient influence curve for the parameter in (2.3) when $P_\delta(g_0)$ is given by (2.2), which can be writen as

$$\Psi(P) = \sum_{A \in \mathscr{A}} \sum_{W \in \mathscr{W}} \bar{Q}(A, W) g(A - \delta(W) | W) Q_W(W) = E_P\{\bar{Q}(A + \delta(W), W)\}. \tag{2.4}$$

The last equality can be checked by changing the index in the summation to $A - \delta(W)$. Equation (2.4) corresponds exactly with computing the marginal mean of Y from the joint distribution of (W, A, Y) with A replaced by $A + \delta(W)$. Note also that if $\delta(W) = 0$, equation (2.4) is equal to the expectation of Y under P.

The efficient influence curve is a key element in semi-parametric efficient estimation, since it defines the linear approximation of any efficient and regular asymptotically linear estimator, and therefore provides an asymptotic bound for the variance of all regular asymptotically linear estimators (Bickel et al., 1997).

Result 1. *The efficient influence curve of (2.4) is*

$$D(P)(O) = \frac{g(A - \delta(W) | W)}{g(A | W)} \{Y - \bar{Q}(A, W)\} + \bar{Q}(A + \delta(W), W) - \Psi(P). \tag{2.5}$$

Since this influence curve as well as the parameter of interest depend only on Q, we will also use the notations $D(P)(O)$ and $D(Q)(O)$ interchangeably.

Proof First of all, notice that the nonparametric estimator of ψ_0 is given by

$$\begin{aligned}\hat{\Psi}(P_n) &= \sum_{y\in\mathscr{Y}}\sum_{a\in\mathscr{A}}\sum_{w\in\mathscr{W}} yP_n(y|a,w)P_n(a-\delta(w)|w)P_n(w)\\ &= \sum_{y\in\mathscr{Y}}\sum_{a\in\mathscr{A}}\sum_{w\in\mathscr{W}} y\frac{P_nf_{y,a,w}}{P_nf_{a,w}}P_nf_{a-\delta(w),w}, \end{aligned} \tag{2.6}$$

where $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{o_i}$ is the empirical measure, $f_{y,a,w} = I(Y = y, A = a, W = w)$, $f_{a,w} = I(A = a, W = w)$, $f_{a-\delta(w),w} = I(A = a-\delta(w), W = w)$, and $I(\cdot)$ denotes the indicator function. Here Pf denotes $\int f dP$.

Recall that the efficient influence curve in a non-parametric model corresponds with the influence curve of the non-parametric estimator. This is true because the influence curve of any regular estimator is also a gradient, and a non-parametric model has only one gradient. Rose and van der Laan (2011) show that if $\hat{\Psi}(P_n)$ is a substitution estimator such that $\psi_0 = \hat{\Psi}(P_0)$, and $\hat{\Psi}(P_n)$ can be written as $\hat{\Psi}^*(P_nf : f \in \mathscr{F})$ for some class of functions $\mathscr{F}$ and some mapping Ψ^*, the influence curve of $\hat{\Psi}(P_n)$ is equal to

$$IC(P_0)(O) = \sum_{f\in\mathscr{F}} \frac{d\hat{\Psi}^*(P_0)}{dP_0f}\{f(O) - P_0f\}.$$

Applying this result to (2.6) with $\mathscr{F} = \{f_{y,a,w}, f_{a,w}, f_{a-\delta(w),w}\}$ gives the desired result. □

This efficient influence curve can be decomposed in three parts corresponding to the orthogonal decomposition of the tangent space implied by the factorization of the likelihood:

$$\begin{aligned} D_1(P)(O) &= \frac{g(A-\delta(W)|W)}{g(A|W)}\{Y - \bar{Q}(A,W)\}\\ D_2(P)(O) &= \bar{Q}(A+\delta(W),W) - E_P\{\bar{Q}(A+\delta(W),W)|W\}\\ D_3(P)(O) &= E_P\{\bar{Q}(A+\delta(W),W)|W\} - \Psi(P). \end{aligned} \tag{2.7}$$

This decomposition of the score is going to be useful later on during the construction of a targeted maximum likelihood estimator of ψ_0. The following result provides the conditions under which an estimator that solves the efficient influence curve equation is consistent.

Result 2. *Let $D(O|\psi_0, \bar{Q}, g)$ be the estimating function implied by the efficient influence curve $D(P)(O)$:*

$$D(O|\psi_0, \bar{Q}, g) = \frac{g(A-\delta(W)|W)}{g(A|W)}\{Y - \bar{Q}(A,W)\} + \bar{Q}(A+\delta(W),W) - \psi_0,$$

let $w(g)(a,w) = g(a-\delta(w)|w)/g(a|w)$, and let $\sup_{a\in\mathscr{A}} w(g_0)(a,W) < \infty, -$ *a.e. We have that $E_{P_0}D(O|\psi_0, \bar{Q}, g) = 0$ if either g is such that $w(g) = w(g_0)$, or $\bar{Q} = \bar{Q}_0$*

Proof Conditioning first on (A,W) and then on W we get

$$E_{P_0}D(O|\psi_0,\bar{Q},g) = E_{P_0}\left[\sum_{a\in\mathscr{A}}\frac{g_0(a|W)}{g(a|W)}g(a-\delta(W)|W)\{\bar{Q}_0(a,W)-\bar{Q}(a,W)\}\right]$$
$$+E_{P_0}\left[\sum_{a\in\mathscr{A}}g_0(a-\delta(W)|W)\bar{Q}(a,W)\right]-E_{P_0}\left[\sum_{a\in\mathscr{A}}g_0(a-\delta(W)|W)\bar{Q}_0(a,W)\right],$$

which completes the proof. □

As a consequence of result 2, under regularity conditions stated in theorem 1 of van der Laan and Rubin (2006), a substitution estimator of $\Psi(P_0)$ that solves the efficient influence curve equation $P_nD(\,\cdot\,|\psi_0,\bar{Q},g)$ will be consistent if either one of $w(g_0)$ and Q_0 is estimated consistently, and it will be efficient if and only if both $w(g_0)$ and Q_0 are estimated consistently. We only rely on consistent estimation of the weight function $w(g_0)$. This consistency can be easier to obtain than consistent estimation of the density g_0, which is required for double robustness of parameters in marginal structural models (Neugebauer and van der Laan, 2007). Since $\Psi(P)$ depends on both $\bar{Q}$ and g, double robustness is a very unexpected result. Some intuition about it is provided by the definition of the parameter in (2.4): if $\bar{Q}_0$ is known, a consistent estimator can always be obtained by computing the empirical mean of $\bar{Q}_0(A+\delta(W),W)$; if the weight function $w(g_0)$ is known, a consistent estimate of ψ_0 would be given by a weighted average of Y with weights $w(g_0)(A,W)$.

Positivity assumption

Alternatives to definition and estimation of causal effects in the context of continuous or categorical multilevel treatments are given by marginal structural models (MSM) and parameters like the ones presented in Petersen et al. (2010). One of the assumptions required to estimate those parameters (the positivity assumption) is given by

$$\sup_{a\in\mathscr{A}}\frac{h(a)}{g_0(a|W)}<\infty,-\ a.e.,$$

for a user-specified weight function h. The function $h(a)=1$ is commonly used, since it implies giving equal weights to all the possible treatment values.

From the formula of the efficient influence curve, the positivity assumption needed to identify and estimate our parameter of interest is given by

$$\sup_{a\in\mathscr{A}}\frac{g_0(a-\delta(W)|W)}{g_0(a|W)}<\infty,-\ a.e. \qquad (2.8)$$

Suppose $\inf_{a\in\mathscr{A}}g_0(a|W)>\varepsilon$ for some small ε. Since the function δ is user-given, we can try to define it in a way so that it is useful to answer the causal question of interest, and yet it does not produce unstable weights. As a result, the positivity assumption as needed to estimate our parameter of interest is more easily achievable than the positivity assumption as required to estimate other causal parameters for continuous exposures.

2.3 Estimators

In this section we present three possible estimators for the parameter of interest. The TMLE and the A-IPTW estimators solve the efficient influence curve equation, and therefore, from result 2, are consistent estimators if either one of $Q_0(A,W)$ and $g_0(A|W)$ is estimated consistently. They are efficient if and only if both of these quantities are estimated consistently. The IPTW is inefficient, and will be consistent only if the estimator of $g_0(A|W)$ is consistent. The TMLE is expected to perform better than the A-IPTW if the positivity assumption is violated, which will be the case if the causal question of interest requires the use of a function δ that produces unstable weights in (2.8). The TMLE is also a better alternative than the A-IPTW when the efficient estimating equation has multiple solutions, or its solution goes out of the natural bounds for the parameter of interest.

The estimators presented in this section require initial estimates of $\bar{Q}_0(A,W)$ and $g_0(A|W)$, which can be obtained through machine learning techniques, parametric or semi-parametric models. The consistency of these initial estimators will determine the consistency and efficiency of the estimators of ψ_0, as discussed previously. Parametric models are commonly used for the sole sake of their nice analytical properties, but they encode assumptions about the distribution of the data that are not legitimate knowledge about the phenomenon under study and usually cause a large amount of bias in the estimated parameter. As an alternative, we recommend the use of machine learning techniques such as the super learner (van der Laan, Polley, and Hubbard, 2007). Super learner is a methodology that uses cross-validated risks to find an optimal estimator among a library defined by the convex hull of a user-supplied list of candidate estimators. One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution). Proofs and simulations regarding these and other asymptotic properties of the super learner can be found in van der Laan, Dudoit, and Keles (2004) and van der Laan and Dudoit (2003).

Influence curve based variance estimators are provided for these three estimators. Consistency of the variance estimators also depends on the consistency of the initial estimates of $\bar{Q}_0$ and g_0. These dependency can be avoided at the cost of computational time and effort by using bootstrapped estimates of the variance.

IPTW

Given an estimator g_n^0 of the exposure density, the IPTW estimator of ψ_0 is defined as

$$\psi_{n,1} = \frac{1}{n}\sum_{i=1}^{n} \frac{g_n^0(A_i - \delta(W_i)|W_i)}{g_n^0(A_i|W_i)} Y_i.$$

The IPTW is an asymptotically linear estimator with influence curve

$$D_{IPTW}(O|\psi_0, g_0) = \frac{g_0(A-\delta(W)|W)}{g_0(A|W)} Y - \psi_0,$$

therefore the variable $\sqrt{n}(\psi_{n,1}-\psi_0)$ converges in distribution to $N\{0,P_0D^2_{IPTW}(g_0)\}$, whose variance can be estimated by $P_nD^2_{IPTW}(\,\cdot\,|\psi_{n,1},g^0_n)$. This variance estimator is conservative, as proved in van der Laan and Robins (2003) and corroborated in the simulation section.

Augmented IPTW

The augmented IPTW is the value $\psi_{n,2}$ that solves the equation $\sum_{i=1}^n D(O_i|\psi_0,\bar{Q}^0_n,g^0_n)=0$, for initial estimates $\bar{Q}^0_n$ and g^0_n of $\bar{Q}_0$ and g_0.

$$\psi_{n,2}=\frac{1}{n}\sum_{i=1}^n\frac{g^0_n(A_i-\delta|W_i)}{g^0_n(A_i|W_i)}\{Y_i-\bar{Q}^0_n(A_i,W_i)\}+\bar{Q}^0_n(A_i+\delta(W_i),W_i).$$

If the estimators $\bar{Q}^0_n$ and g^0_n are consistent, the A-IPTW is an asymptotically linear estimator with influence curve $D(O|\psi_0,\bar{Q}_0,g_0)$. As in the case of the IPTW, the variable $\sqrt{n}(\psi_{n,2}-\psi_0)$ converges in law to a random variable with distribution $N\{0,P_0D^2(\cdot|\psi_0,\bar{Q}_0,g_0)\}$, whose variance can be estimated as $P_nD^2(\,\cdot\,|\psi_{n,2},\bar{Q}^0_n,g^0_n)$. van der Laan and Robins (2003) (sections 2.3.7 and 2.7.1) show that inference based on this variance estimator is valid only if g^0_n is consistent, providing exact inference when $\bar{Q}^0_n$ is consistent, and conservative inference when $\bar{Q}^0_n$ is inconsistent.

Targeted maximum likelihood estimator

Targeted maximum likelihood estimation (van der Laan and Rubin, 2006) is a loss-based semiparametric estimation method that yields a substitution estimator of a target parameter of the probability distribution of the data that solves the efficient influence curve estimating equation, and thereby yields a double robust locally efficient estimator of the parameter of interest, under regularity conditions.

In order to define a targeted maximum likelihood estimator for ψ_0, we need first to define three elements: (1) A loss function $L(Q)$ for the relevant part of the likelihood required to evaluate $\Psi(P)$, which in this case is $Q=(\bar{Q},g,Q_W)$. This function must satisfy $Q_0=\arg\min_Q E_{P_0}L(Q)(O)$, where Q_0 denotes the true value of Q; (2) An initial estimator Q^0_n of Q_0; (3) A parametric fluctuation $Q(\varepsilon)$ through Q^0_n such that the linear span of $\frac{d}{d\varepsilon}L\{Q(\varepsilon)\}|_{\varepsilon=0}$ contains the efficient influence curve $D(P)$ defined in (2.5). These elements are defined below:

Loss Function

As loss function for Q, we will use $L(Q)=L_Y(\bar{Q})+L_A(g)+L_W(Q_W)$, where for continuous Y we set $L_Y(\bar{Q})=\{Y-\bar{Q}(A,W)\}^2$, for binary Y we set $L_Y(\bar{Q})=Y\log\{\bar{Q}(A,W)\}+(1-Y)\log\{1-\bar{Q}(A,W)\}$, $L_A(g)=-\log g(A|W)$, and $L_W(Q_W)=-\log Q_W(W)$. It can be easily verified that this function satisfies $Q_0=\arg\min_Q E_{P_0}L(Q)(O)$.

Parametric Fluctuation

Given an estimator Q^k_n of Q_0, with components $(\bar{Q}^k_n,g^k_n,Q^k_{W,n})$, we define the $(k+1)$th fluctuation of Q_n^k as follows:

$$\begin{aligned}\bar{Q}_n^{k+1}(\varepsilon_1)(A,W) &= \bar{Q}_n^k(A,W) + \varepsilon_1 H_1^k(A,W)\\ g_n^{k+1}(\varepsilon_1)(A|W) &= \frac{\exp\{\varepsilon_1 H_2^k(A,W)\} g_n^k(A|W)}{\int_{\mathscr{A}} \exp\{\varepsilon_1 H_2^k(A,W)\} g_n^k(A|W)}\\ Q_{W,n}^{k+1}(\varepsilon_2)(W) &= \frac{\exp\{\varepsilon_2 H_3^k(W)\} Q_{W,n}^k(W)}{\int_{\mathscr{W}} \exp\{\varepsilon_2 H_3^k(W)\} Q_{W,n}^k(W)},\end{aligned}$$

where $H_1^k(A,W) = g_n^k(A-\delta(W)|W)/g_n^k(A|W)$, $H_2^k(A,W) = D_2(P^k)(O)$ and $H_3(W) = D_3(P^k)(O)$, with D_2 and D_3 defined as in (2.7). We define these fluctuations using a two-dimensional ε with two different parameters ε_1 and ε_2. It is theoretically correct to define these fluctuations using any dimension for ε, as long as the condition $D(P) \in < \frac{d}{d\varepsilon} L\{Q(\varepsilon)\}|_{\varepsilon=0} >$ is satisfied, where $< \cdot >$ denotes linear span. The convenience of the particular choice made here will be clear once the TMLE is defined.

Targeted Maximum Likelihood Estimator

The TMLE is defined by the following iterative process:

1. Initialize $k = 0$.
2. Estimate ε as $\varepsilon_n^k = \arg\min_\varepsilon P_n L\{Q_n^k(\varepsilon)\}$.
3. Compute $Q_n^{k+1} = Q_n^k(\varepsilon_n^k)$.
4. Update $k = k+1$ and iterate steps 2 through 4 until convergence (i.e., until $\varepsilon_n^k = 0$)

First of all, note that the value of ε_2 that minimizes the part of the loss function corresponding to the marginal distribution of W in the first step (i.e., $-P_n \log Q_{W,n}^1(\varepsilon_2)$) is $\varepsilon_2^1 = 0$. Therefore, the iterative estimation of ε only involves the estimation of ε_1. The kth step estimation of ε_1 is obtained by minimizing $P_n[L_Y\{\bar{Q}_n^k(\varepsilon_1)\} + L_A\{g_n^k(\varepsilon_1)\}]$, which implies solving the estimating equation

$$S^k(\varepsilon_1) = \sum_{i=1}^n \left[Y_i - \{\bar{Q}_n^k(A_i,W_i) + \varepsilon_1 H_1^k(O_i)\}\right] H_1^k(O_i) + D_2(P_n^k)(O_i) - \frac{\sum_{A\in\mathscr{A}} D_2(P_n^k)(O_i)\, \exp\{\varepsilon_1 D_2(P_n^k)(O_i)\}\, g_n^k(A_i|W_i)}{\sum_{A\in\mathscr{A}} \exp\{\varepsilon_1 D_2(P_n^k)(O_i)\}\, g_n^k(A_i|W_i)} \qquad (2.9)$$

where

$$D_2(P_n^k)(O) = Q_n^k(A+\delta(W),W) - \sum_{A\in\mathscr{A}} Q_n^k(A+\delta(W_i),W_i) g_n^k(A|W_i).$$

The TMLE of ψ_0 is defined as $\psi_{n,3} \equiv \lim_{k\to\infty} \Psi(P_n^k)$, assuming this limit exists. In practice, the iteration process is carried out until convergence in the values of ε_n^k is achieved, and an estimator

Q_n^* is obtained. The variance of $\psi_{n,3}$ can be estimated by $P_n D^2(\,\cdot\,|\psi_{n,3}, \bar{Q}_n^*, g_n^*)$, which like the augmented IPTW variance estimator is consistent only if both g_n^* and Q_n^* are consistent, is conservative if g_n^* is consistent but Q_n^* is not, and is inconsistent in any other case.

2.4 Simulation study

In order to provide an example of the finite sample properties of the estimators discussed in section 2.3, a simulation study was performed. We focus on just one data generating distribution, which provides a limited but useful situation to demonstrate our claims about consistency and efficiency.

$$\begin{aligned}
W_1 &\sim U\{0,1\}\\
W_2 &\sim Ber\{0.7\}\\
A|W_1,W_2 &\sim Poisson\{\exp(3+.3\log(W_1)-.2\exp(W_1)W_2)\}\\
Y|A,W_1,W_2 &\sim N\{1+.5A-.2AW_2+2A\tan(W_1^2)-2W_1W_2+AW_1W_2,\ 1\}.
\end{aligned}$$

Assuming that we are interested in estimating the effect of a constant shift of $\delta(W_1,W_2)=2$, the true parameter value for this data generating distribution is $\psi_0 = 22.95$, and the efficiency bound equals $\{Var_{P_0} D(P_0)(O)\}^{1/2} = 17.81$.

For sample sizes $n = 50, 100, 200$ and 500, we simulated 2000 samples from the previous data generating distribution, and estimated ψ_0 using the three estimators proposed in the previous section. As initial estimators of $\bar{Q}_0(A,W)$ and $g_0(A|W)$ we considered four cases: 1) correctly specified model for both $\bar{Q}_0(A,W)$ and $g_0(A|W)$, 2) incorrectly specified model for $\bar{Q}_0(A,W)$ but correctly specified for $g_0(A|W)$, 3) correctly specified model for $\bar{Q}_0(A,W)$ but incorrectly specified for $g_0(A|W)$, and 4) incorrectly specified model for both $\bar{Q}_0(A,W)$ and $g_0(A|W)$; where misspecification of the models was performed by considering the correct distribution and link function but only main terms in the linear predictor.

TML estimation of ψ_0 was performed using the R `tmle.shift()` function presented in appendix A.1. The average and variance of the estimates across the 2000 samples was computed as an approximation to the expectation and variance of the estimator (Table 2.1), respectively.

The results in Table 2.1 confirm the double robustness of the TMLE and A-IPTW, which had been proven analytically in result 2. The TMLE and A-IPTW are unbiased even for small sample sizes, whereas the IPTW needs larger sample sizes to achieve unbiasedness.

Regarding the variance of the estimators, Table 2.2 shows that the IPTW estimator is inefficient, and its influence-curve-based variance estimator is very conservative. The variances of the TMLE and A-IPTW are approximately equal to the efficiency bound if the models for $\bar{Q}_0$ and g_0 are correctly specified, although the same equality is observed if only $\bar{Q}_n^0$ is misspecified. This is because, as stated in result 2, we only need consistent estimation of the weights $w(g_0)(A,W)$, which can be obtained through a possibly misspecified estimator of g_0. On the other hand, the variance of these estimators is considerably affected by misspecification of the model for $\bar{Q}_0$ (models 3 and 4), even if g_n^0 is correctly specified.

n	Model	TMLE	IPTW	A-IPTW
50	1	22.99	22.66	22.99
	2	22.99	22.49	22.99
	3	22.88	22.66	22.91
	4	22.01	22.49	22.04
100	1	22.95	22.81	22.95
	2	22.96	22.61	22.95
	3	22.89	22.81	22.92
	4	21.97	22.61	22.00
200	1	22.99	22.89	22.99
	2	22.99	22.68	22.99
	3	22.94	22.89	22.96
	4	21.99	22.68	22.02
500	1	22.97	22.93	22.97
	2	22.97	22.71	22.97
	3	22.93	22.93	22.96
	4	21.97	22.71	22.00

Table 2.1: Expectation of the estimators for different sample sizes and model specifications. True value is 22.95.

The fact that influence curve based variance estimators of the TMLE and A-IPTW are consistent even for misspecified g_n^0 can be taken to be a coincidence associated with this particular data simulating scheme. As explained in section 2.3, this type of consistency does not hold in general.

Since all estimators considered are asymptotically linear, 95% normal-based confidence intervals can be computed. Their coverage probabilities are presented in Table (2.3). The conservativeness of the IPTW can also be appreciated here. The consistent TMLE and A-IPTW based confidence intervals have perfect asymptotic coverage probability. Intervals associated to inconsistent estimators (model 4) have, as expected, confidence levels below the nominal value. In this simulation we do not observe significant differences between the TMLE and the A-IPTW.

2.5 Application

With the objective of illustrating the procedure described in the previous sections, we revisit the problem analyzed by Bembom and van der Laan (2007) of assessing the extent to which physical activity in the elderly is associated with reductions in cardiovascular morbidity and mortality, and improvement in, or prevention of metabolic abnormalities. Tager, Hollenberg, and Satariano (1998) followed a group of people over 55 years of age living around Sonoma, CA, over a time period of about ten years as part of a longitudinal study of physical activity and fitness (Study of Physical Performance and Age Related Changes in Sonomans - SPPARCS). The goal in analyzing

n	Model	TMLE	IPTW	A-IPTW
50	1	17.94 (17.66)	20.33 (26.80)	17.94 (17.66)
	2	17.94 (17.67)	19.16 (25.03)	17.94 (17.66)
	3	18.92 (17.81)	20.33 (26.80)	18.94 (18.08)
	4	18.21 (18.07)	19.16 (25.03)	18.25 (17.77)
100	1	17.93 (17.74)	20.36 (27.63)	17.93 (17.74)
	2	17.93 (17.75)	19.04 (25.72)	17.93 (17.75)
	3	18.96 (18.14)	20.36 (27.63)	18.98 (18.45)
	4	18.34 (18.37)	19.04 (25.72)	18.35 (18.06)
200	1	17.77 (17.77)	20.17 (28.00)	17.77 (17.77)
	2	17.77 (17.78)	18.93 (25.97)	17.77 (17.77)
	3	18.62 (18.35)	20.17 (28.00)	18.64 (18.68)
	4	17.98 (18.57)	18.93 (25.97)	18.00 (18.24)
500	1	17.38 (17.79)	20.40 (28.37)	17.39 (17.79)
	2	17.38 (17.80)	18.94 (26.24)	17.39 (17.80)
	3	18.50 (18.49)	20.40 (28.37)	18.52 (18.84)
	4	17.74 (18.71)	18.94 (26.24)	17.76 (18.36)

Table 2.2: Standard error of the estimator (times $\sqrt{n}$). Expectation of the influence curve based estimator of the variance (times $\sqrt{n}$) in parentheses. Efficiency bound is 17.81

the data that were collected as part of this study is to examine the effect of baseline vigorous LTPA (Leisure Time Physical Activity) on subsequent five-year all-cause mortality.

In this chapter, we use the same measure of LTPA used by Bembom and van der Laan (2007), which is a continuous score based on the number of hours that the participants were engaged in vigorous physical activities such as jogging, swimming, bicycling on hills, or racquetball in the last seven days, and the standard intensity values in metabolic equivalents (MET: Metabolic Equivalent of Task) of such activities, where one MET is approximately equal to the oxygen consumption required for sitting quietly.

The primary confounding factors that we adjust for are described in Table 2.4. Age and gender are natural confounders, and the rest of the variables intend to account for the subject's underlying level of general health. Of the 2092 subjects enrolled in the SPPARCS study, 40 were missing information in at least one of this variables; our analysis is based on the remaining 2052 subjects.

In the sequel of this section, the vector containing the confounders will be denoted by W, the continuous MET score by A, and the indicator of five-year all-cause mortality by Y. We are interested in estimating the effect of a policy that will produce an increase of 12 METs (corresponding, for instance, to bicycling during three hours at less than 10mph per week) in the average of the conditional distribution physical activity, given the covariates. Note that our intervention could also be defined by using different values of MET in each strata defined by the covariates W.

Initial estimators of the conditional density $g_0(A|W)$ and the conditional expectation $\bar{Q}_0(A,W)$

n	Model	TMLE	IPTW	A-IPTW
50	1	0.93	0.97	0.93
	2	0.93	0.96	0.93
	3	0.92	0.97	0.92
	4	0.90	0.96	0.89
100	1	0.94	0.98	0.94
	2	0.94	0.98	0.94
	3	0.93	0.98	0.94
	4	0.89	0.98	0.89
200	1	0.95	0.98	0.95
	2	0.95	0.97	0.95
	3	0.94	0.98	0.95
	4	0.87	0.97	0.87
500	1	0.95	0.99	0.95
	2	0.95	0.98	0.95
	3	0.94	0.99	0.95
	4	0.78	0.98	0.78

Table 2.3: Coverage probability of normal based confidence intervals.

are presented below.

Initial estimator of g_0

For the estimation of the density $g_0(A|W)$, we consider the estimator presented in chapter 1. We now provide a summary of the rationale behind this estimator. Consider $k+1$ values $\alpha_0, \alpha_1, \ldots, \alpha_k$ spanning the range of the data and defining k bins. Now, consider the following class of histogram-like candidate estimators of the conditional density $g_0(A|W)$

$$g_{n,\alpha}(A=a|W) = \frac{Pr_n\{A \in [\alpha_{m-1}, \alpha_m)|W\}}{\alpha_m - \alpha_{m-1}}, \text{ for } \alpha_{m-1} \leq a < \alpha_{m-1},$$

where the choice of the α values and the number of bins index the candidates in the class. The probabilities in the numerator are estimated through the super learner. The final estimator of the density consists of a convex combination of these estimators that minimizes the cross-validated empirical risk.

As an example, Figure 2.1 shows two contrasting estimated densities $g_n(A|W)$ for different profiles W, in which a subject with better general health status is more likely to have higher levels of physical activity. As pointed out in chapter 1, this methodology allows the detection of high density areas in the exposure mechanism, like the one detected at zero in Figure 2.1 (a). This spike appears because this is a “zero-inflated” exposure, in which a large proportion of the population do not practice any amount of physical activity.

Variable	Description
Gender	Female Male
Age	Age in years
Health	Self-rated health status: Excellent Fair Poor
NRB	Score of self-reported physical functioning rescaled between 0 and 1
Card	Previous occurrence of any of the following cardiac events: Angina, myocardial infarction, congestive heart failure, coronary by-pass surgery, and coronary angioplasty
Chron	Presence of any of the following chronic health conditions: stroke, cancer, liver disease, kidney disease, Parkinson's disease, and diabetes mellitus
Smoking	Never smoked Current smoker Ex-smoker
Decline	Activity decline compared to 5 or 10 years earlier

Table 2.4: Confounders.

Initial estimator of $\bar{Q}_0$

For the initial estimator of $\bar{Q}_0$ we used the super learner (van der Laan, Polley, and Hubbard, 2007). Table 2.5 shows the candidates used, their cross-validated risks, and their coefficients in the final super learner predictor. In order to get a consistent estimator of $\bar{Q}_0$ the library of candidate estimators should be as large as possible. Since this is an illustrating example, we allow ourselves to use this small library.

	Cross-validated Risk	Coef.
GLM main effects	0.1079	0.0000
GLM main eff. and two way interactions	0.1143	0.0835
GAM degree 2	0.1073	0.0000
GAM degree 3	0.1071	0.9165
Bayes GLM main effects	0.1078	0.0000

Table 2.5: Super learner output for estimation of $\bar{Q}_0$.

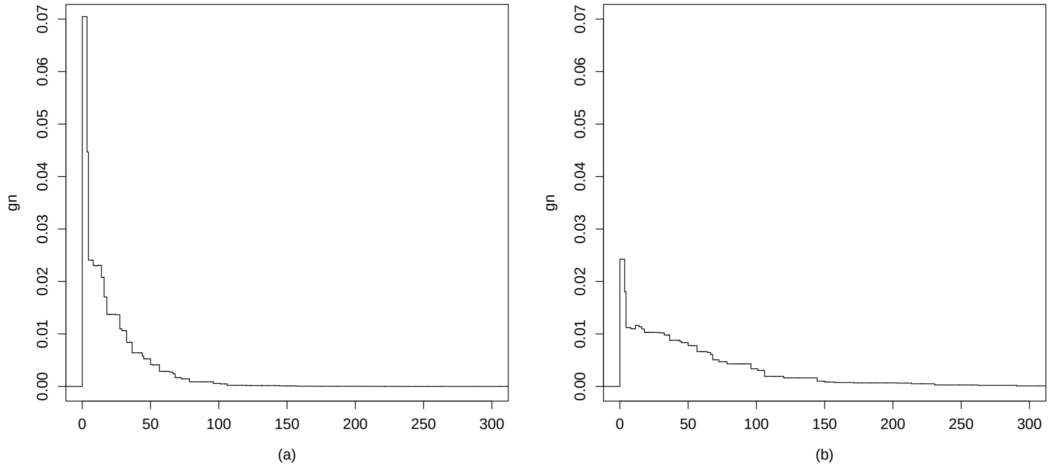


Figure 2.1: Estimated conditional density of A given the profiles: (a) age = 77, gender = female, health = fair, nrb = 0.9, card = no, smoke = ex-smoker, decline = yes, chron = yes; and (b) age = 71, gender = male, health = good, nrb = 0.88, card = no, smoke = never smoked, decline = no, chron = no

Estimators of ψ_0

Table 2.6 shows the three estimates of ψ_0 with their standard errors, as described in section 2.3. As an example, the TML estimated value of $\psi_{n,3} = 0.16$ indicates that if a policy that increases the average leisure time physical activity by the equivalent of 12 METs is implemented, the estimated risk of death in the intervened population will be 16%.

If the objective is to perform a comparison with the current risk of death, we can define a population intervention parameter ψ_0^1 as

$$\psi_0^1 = \psi_0 - E_{P_0}(Y).$$

This is a parameter that compares the expected risk of death in the intervened population with the current risk of death, and therefore describes the gain obtained by carrying out the intervention of interest. For a given estimator ψ_n of ψ_0, an asymptotically linear estimator of ψ_0^1 is given by $\psi_n^1 = \psi_n - \bar{Y}$. Its influence curve can be computed as $D^1(P)(O) = D(P)(O) - \{Y - E_P(Y)\}$, and its variance is estimated through the sample variance of $D^1(P)(O)$. Here $D(P)(O)$ is the influence curve of each of the estimators defined in section 2.3. The estimates of ψ_0^1 and their standard errors are presented in table 2.6. Confidence intervals and p-values for hypothesis testing can be computed based on the normal approximations for asymptotically linear estimators described in section 2.3. In light of the results from the simulation section and the theoretical properties of the estimators, we rely on the TMLE and A-IPTW to measure the effect of the intervention of interest. The estimated value of ψ_n^1 means that if a policy increasing the average time of physical activity by

	TMLE	A-IPTW	IPTW
ψ_0	0.1600(0.0104)	0.1599(0.0105)	0.1454(0.0135)
ψ_0^1	$-0.0179(0.0071)$	$-0.0179(0.0071)$	$-0.0324(0.0117)$

Table 2.6: Estimates of ψ_0.

the equivalent of 12 METs (corresponding, for instance, to bicycling during three hours per week at less than 10mph) is put in place, the risk of all-cause mortality in the elderly would be reduced by 1.79%. These results are consistent with the findings of Bembom and van der Laan (2007).

2.6 Discussion

In this chapter we define a new parameter measuring the causal effect of a population intervention that (as opposed to most of the parameters presented in the literature) accounts for the fact that in most cases the post-intervention exposure continues to be a random variable. We argue that this parameter makes more intuitive sense when the objective is to assess the causal effect of policies intending to modify an exposure variable that cannot be directly intervened upon. For example, as argued in Bembom and van der Laan (2007), it makes little sense to assess the effect of a realistic policy in terms of a static intervention in which every subject in a population of elderly people is required to increase his level of physical activity to the maximum, or even to a level defined by a deterministic function of the covariates. Such interventions are never possible due to particular health conditions, physical functioning constraints, or simple inability to enforce every subject to comply with the treatment level dictated by the intervention. Hence, deterministic interventions do not provide an accurate tool to measure the causal effect of a realistic policy that renders a stochastic exposure.

Another appealing feature of the framework presented in this chapter is that it provides a natural way of defining and estimating causal effects for continuous variables, or discrete variables with more than two levels, which are currently defined through the specification of a working MSM (Neugebauer and van der Laan, 2007). The positivity assumption required to estimate our proposed causal parameter can be made weaker than the positivity assumption required to estimate MSM parameters.

Three estimators of the parameter were proposed, two of which are double robust to misspecification of the models for the treatment mechanism g_0 and the conditional expectation $\bar{Q}_0$, even when the parameter depends on these two quantities. This double robustness property is proven analytically, and corroborated in a simulation study.

Chapter 3

Stochastic interventions: truncating the exposure mechanism

Current approaches to causal inference (Rubin, 1974; Rubin, 1978; Pearl, 2000; Pearl, 2009) define causal parameters as functions of the distribution of random variables generated by a system in which the stochastic nature of a set of variables is intervened on, leading to changes in the stochastic nature of the variables that depend causally on them. Such interventions may be defined in various ways: static, dynamic or stochastic. A static intervention is one in which the treatment is set to a given fixed value deterministically, while a dynamic intervention allows such value to depend on variables that precede it causally. Static interventions have also been called deterministic (Korb et al., 2004) or atomic (Pearl, 2000).

In spite of their wide use, deterministic interventions (whether static or dynamic) do not provide an appropriate framework to answer causal questions about phenomena that are not subject to direct intervention. Feasible interventions often interact with other factors (e.g., a medication has impact in several organs), fail to put the exposure of interest into a deterministic state (e.g., it is unrealistic to set an individuals' exercising regime according to a deterministic function), or are the result of implementing policies that target stochastic changes in the behavior of a population (e.g., the use of mass media messages advertising condom use as a means of prevention of HIV infection is a deterministic treatment at the community level that renders a stochastic one at the individual level, because each individual will react stochastically to the intervention depending upon exogenous observed or non observed factors (McAlister, 1991)).

In general (Korb et al., 2004), an intervention can be simply defined as an external manipulation of a causal system, whether that manipulation is deterministic or stochastic. A static intervention corresponds to an alteration of the causal system in which the density of the exposure is changed to a degenerate one. One can also intervene in the exposure by changing its density in any arbitrary way, which leads to a natural generalization of the counterfactual framework of Rubin (1978). This general approach is perhaps of more interest from a policy making standpoint: if the counterfactual distribution of the exposure reflects the expected changes induced by a hypothetical intervention policy, the intervened model contains all the information about the causal effect of the intervention in the distribution of the outcome.

Stochastic interventions also provide a new, natural way of non-parametrically defining causal parameters for any type of exposure (e.g., continuous ones), regardless of its support and dominating measure. Thus far this was only possible through the use of misspecified parametric models or the use of marginal structural models (Neugebauer and van der Laan, 2007). Some advantages of stochastic interventions with respect to marginal structural models include weakening the positivity assumption and robustness with respect to misspecification of the model for the treatment mechanism.

Because stochastic interventions generalize static and dynamic interventions, and since several intervention policies are not representable in terms of either static or dynamic interventions, the development of methods for identification and estimation of parameters defined in terms of stochastic interventions is of main interest to the causal inference research community.

Among the few works dealing with the mathematical formalization of stochastic interventions figure Didelez, Dawid, and Geneletti (2006) and Dawid and Didelez (2010), who provide a systematic and comprehensive discussion of identification of parameters of stochastic, dynamic and static interventions, studying them from a decision-theoretic viewpoint, exploiting representations of causal systems in terms of regime indicators and influence diagrams, and presenting a parallel between their theory and existing theory for dynamic, non-stochastic regimes. Tian (2008) shows that the identification of sequential intervention, whether stochastic or not, can be reduced to identification of a specific set of sequential static interventions, for which there are complete identifications algorithms available in the literature. It is therefore no surprise that identification of our parameter in section 3.1 requires no further assumptions than those required for identification of a static intervention.

Stochastic interventions arise in applications either inspired by a deterministic intervention, or because they are of interest in themselves. The most popular example of the former situation is given by the definition of natural direct effects, in which the effect of A on Y is confounded by W and mediated by a variable Z. If A and Z are binary, one can define the counterfactual Y_{1,Z_0} (Robins and Greenland, 1992; Pearl, 2001; Zheng and van der Laan, 2011b; Hafeman and VanderWeele, 2011) as the outcome under a model in which A has been set to $a = 1$ with probability one, and the distribution of Z has been changed to that of Z_0, the latter being the counterfactual of Z obtained when A is set to $a = 0$ with probability one. This setting provides an example in which the intervention of interest is performed in two nodes, using a static intervention for A, and a stochastic intervention for Z. Didelez, Dawid, and Geneletti (2006) and Robins and Richardson (2010) discuss in detail the case in which several direct and indirect effects are defined and studied in the context of stochastic interventions. Taubman et al. (2009b) considered an intervention in the BMI defined by a truncation of the original exposure distribution, which, contrary to the truncation that we will use in this chapter, relocates the mass originally located above the threshold across all the values below the threshold. As explained by Stitelman, Hubbard, and Jewell (2010b), such intervention is usually the result of dichotomizing a continous variable and considering a static intervention in the dichotomous version of the treatment variable. This dichotomization represents current common practice, in section 3.2 we will discuss the differences with the approach presented in this chapter. Cain et al. (2010) briefly discuss a stochastic intervention in the context of comparing dynamic treatment regimes for HIV infected patients. The regimes they discuss are of the type “initiate treatment within m months after the recorded CD4 cell count first falls below x”, and they are interested in an atomic intervention in the CD4 cell count X, and a discrete uniform $\{0,m\}$ post-intervention distribution for the number of months before treatment M. Such intervention is discussed in more detail by Young et al. (2011).

Among the applications in which stochastic interventions arise as an interest in themselves, in chapter 2 we considered the effect of an intervention in a population of people over 55 years of age that aimed to change the distribution of the amount of energy spent in leisure time physical activity on all cause mortality. In the present chapter we will analyze the effect of an intervention that intends to reduce air pollution levels below a certain threshold, but allows a stochastic distribution of air pollutants below such threshold. The claims about identifiability and properties of the estimators presented in this chapter are valid only for this stochastic intervention, although they can be generalized to a broader class of interventions.

Consistent and efficient estimation of statistical parameters in semi parametric models has been studied by Bickel et al. (1997), van der Laan and Robins (2003), Rose and van der Laan (2011), and Tsiatis (2006), among others. In particular, Rose and van der Laan (2011) provide a very valuable link between efficient estimation theory in semiparametric models and causal inference, empowering researchers with tools to define a causal parameter of interest, truthfully propose a model for the distribution of the data, and compute an efficient, targeted estimate of the parameter of interest under that model. By a truthful definition of the statistical model we mean that the start point is a completely non parametric model, that can only be reduced in size if real knowledge about the distribution of the data is obtained. Parametric and other assumptions often made for the sake of computational convenience are not allowed: they do not represent knowledge about the phenomena under study and therefore result in biased estimates.

In this chapter we will demonstrate the use of stochastic interventions to assess the effect of a (hypothetical) law that enforces pollution levels below a certain cutoff point. For estimation of causal effects we use efficiency theory in semiparametric models, and in particular the targeted minimum loss based estimation road map as described by Rose and van der Laan (2011)

The chapter is organized as follows. In section 3.1 we define the observed and counterfactual data, as well as the causal and statistical parameter and its efficient influence function. In section 3.2 we discuss how this problem would be tackled with existing methods, and argue that the conclussions of such methods are misleading. In section 3.3 we present three estimators of the statistical parameter of interested: an inverse probability of treatment weighted estimator (IPTW), an augmented IPTW that solves the efficient influence curve equation, and a targeted minimum loss based estimator (TMLE). section 3.4 provides an extension to longitudinal data settings and illustrates its use to measure the effect of NO_2 concentrations in the air on asthma symptoms in children between 6 and 11 years of age. Finally, section 3.5 provides some concluding remarks and directions of future research.

3.1 Observed data, counterfactuals and parameter of interest

Causal and statistical models

Consider an experiment in which an exposure variable A, a continuous or binary outcome Y and a set of covariates W are measured for n randomly sampled subjects, and the outcome is measured subject to an indicator of missingness denoted by C. Let $O = (W, A, C, CY)$ represent a random variable with distribution P_0, and $O_1, \ldots, O_n$ represent n i.i.d. observations of O. Assume that the following non parametric structural equation model Pearl, 2000, NPSEM holds:

$$W = f_W(U_W); \quad A = f_A(W, U_A); \quad C = f_C(A, W, U_C); \quad CY = Cf_Y(A, W, U_Y), \tag{3.1}$$

where U_W, U_A, U_C and U_Y are exogenous random variables assumed to satisfy the randomization assumption $(U_C, U_A) \perp\!\!\!\perp U_Y | W$. The true distribution of O can be factorized as

$$P_0(O) = P_0(W)P_0(A|W)P_0(C|A,W)\{P_0(Y|A,W,C)\}^C\{I(CY=0)\}^{1-C}, \tag{3.2}$$

and we denote $g_0(A|W) \equiv P_0(A|W)$, $\phi_0(A,W) \equiv P_0(C=1|A,W)$, and $\bar{Q}_0(A,W,C) \equiv E_0(Y|A,W,C)$.

In the next subsections we will use this data structure to define a causal and statistical parameter of interest, find its efficient influence curve (Bickel et al., 1997; van der Laan and Robins, 2003), and establish the asymptotic properties of estimators that solve the efficient curve equation.

Causal and statistical parameters

Assume that the interest of the researcher relies in estimating the effect of a policy that will cause a truncation on the exposure, relocating the probability mass originally located above certain threshold δ_2 in an interval (δ_1, δ_2), where $\delta_1 = \delta_2 - \varepsilon$ for some small ε. Formally put within the causal framework of Pearl (2000), such policy can be described by considering the modified system

$$W = f_W(U_W); \quad A_{P_\delta} = T(g_I)\{f_A(W, U_A), W\}; \quad C_{P_\delta} = 1; \quad Y_{P_\delta,1} = f_Y(A_{P_\delta}, W, U_Y), \tag{3.3}$$

where g_I denotes a user-given (but possibly unknown, e.g. one could set $g_I = g_0$) conditional distribution of A given W,

$$T(g_I)(A,W) = \begin{cases} G_I^{-1}\{G_0(A)\} & \text{if } A < \delta_1 \\ G_I^{-1}\left\{\dfrac{G_I(A|W) - G_I(\delta_1|W)}{K(g_I)(W)} + G_I(\delta_1|W)\middle| W\right\} & \text{if } A \geq \delta_1, \end{cases} \tag{3.4}$$

and G_I denotes the distribution function corresponding to g_I. The distribution of A_{P_δ} is given by

$$P_\delta(g_I)(A_{P_\delta} = a|W) = \begin{cases} g_I(a|W) & \text{if } a < \delta_1 \\ g_I(a|W)K(g_I)(W) & \text{if } \delta_1 \leq a \leq \delta_2 \\ 0 & \text{otherwise ,} \end{cases} \tag{3.5}$$

where

$$K(g)(W) = \frac{1 - G\{\delta_1|W\}}{G\{\delta_1, \delta_2|W\}},$$

and in an abuse of notation $G\{\delta_1, \delta_2|W\} \equiv \int_{\delta_1 \leq a \leq \delta_2} g(a|W)d\mu(a)$. This intervention has two consequences on the distribution of the exposure: (1) it changes the distribution of values of A below δ_1 from g_0 to g_I; and (2) it relocates the values of A above δ_1 between δ_1 and δ_2 according to distribution (3.5). As special case we will consider the case $g_I = g_0$, which is of particular interest when we weant to assess the effect of policies that enforce the value of certain exposure below a pre-specified level. In such cases the distribution of the set of individuals that already comply with the enforced cut-off is expected to remain unchanged, making consequence (1) void.

Under the randomization assumption, the expectation of the outcome $Y_{P_\delta,1}$ is identified as a function of the observed data generating mechanism P_0 as

$$\Psi(P_0) = E(Y_{P_\delta,1}) = E_{g_I,Q_W}\left\{\bar{Q}_0(A,W,1) \times M(g_I)(A,W)\right\}, \tag{3.6}$$

where $M(g)(A,W) = I_{\delta_1}(A) + I_{\delta_1,\delta_2}(A) \times K(g)(W)$, $I_{\delta_1}(A) = I(A < \delta_1)$ and $I_{\delta_1,\delta_2}(A) = I(\delta_1 \leq A \leq \delta_2)$. This identification result follows from the following argument. The usual consistency assumption $(A = a, C = 1) \Rightarrow Y_{a,1} = Y$ implies $(A_{P_\delta} = a, C = 1) \Rightarrow Y_{P_\delta,1} = Y_{a,1}$, therefore $P(Y_{P_\delta,1} = y|A_{P_\delta} = a, C = 1, W = w) = P(Y_{a,1} = y|A_{P_\delta} = a, C = 1, W = w)$. It is easy to verify that $(U_A, U_C) \perp\!\!\!\perp U_Y|W$ implies $(A_{P_\delta}, C) \perp\!\!\!\perp Y_{a,c}|W$ for all a, c. Thus $P(Y_{a,1} = y|A_{P_\delta} = a, C = 1, W = w) = P(Y_{a,1} = y|W = w)$, which from standard arguments for identification of static interventions (see for example Pearl (2000)) can be shown to be identified by $P(Y = y|A = a, C = 1, W = w)$. This result can also be derived using the "G-recursion" formula, presented by Dawid and Didelez (2010), which generalizes the G-computation formula for dynamic regimes of Robins (1986). It can also be shown that the assumptions stated here are equivalent to the assumption of "simple stability" as defined by Dawid and Didelez (2010), which generalizes the (sequential) randomization assumption to the case of stochastic interventions.

The parameter in (3.6) is a weighted mean of $\bar{Q}_0(A,W,1)$ (with respect to the joint distribution of A and W), in which values of $\bar{Q}_0(A,W,1)$ for which $A < \delta_1$ receive weight one, values for which $\delta_1 \leq A \leq \delta_2$ receive weight $K(g_I)(W)$, and values for which $A > \delta_2$ receive weight 0. This makes intuitive sense; if the portion of the population whose exposure is originally above δ_2 is relocated in exposure levels in $[\delta_1, \delta_2]$, the expected outcome of individuals in $[\delta_1, \delta_2]$ should be reweighed by $K(g_I)(W)$, and the portion above δ_2 should be reweighed by zero, given that no portion of the population will fall in that region after the intervention.

As a consequence of the formal equivalence between the counterfactual and the non-parametric structural equation model frameworks (Pearl, 2000, section 7.4.4.), all the results presented in this chapter can be derived under either paradigm. Furthermore, parameter (3.6) is a purely statistical parameter defined as the expectation of the outcome under a different distribution of A given W, and can therefore be of interest in itself, without any underlying causal assumption or interpretation. In the following subsections we deal with estimation of (3.6) under a non-parametric model.

Efficient influence curve

The efficient influence curve is a key element in semi-parametric efficient estimation, since it defines the linear approximation of any efficient and regular asymptotically linear estimator, and therefore provides an asymptotic bound for the variance of all regular asymptotically linear estimators (Bickel et al., 1997). We limit the discussion to efficient estimation of parameter (3.6) when $g_I = g_0$; the case of a user given function g_I is easier and can be studied using similar arguments.

Result 3. *The efficient influence curve of parameter (3.6) when $g_I = g_0$ is given by*

$$\begin{aligned}
D(P_0)(O) =& \frac{C}{\phi_0(A,W)} M(g_0)(A,W)\{Y - \bar{Q}_0(A,W,1)\} && (3.7)\\
&+0 && (3.8)\\
&+M(g_0)(A,W)\left\{Q_0(A,W,1) - \frac{E_{P_0}\{\bar{Q}_0(A,W,1)I_{\delta_1,\delta_2}(A)|W\}}{G_0\{\delta_1,\delta_2|W\}}\right\} &&\\
&\quad + \frac{E_{P_0}\{\bar{Q}_0(A,W,1)I_{\delta_1,\delta_2}(A)|W\}}{G_0\{\delta_1,\delta_2|W\}} - E_{P_0}\{\bar{Q}_0(A,W,1)M(g_0)(A,W)|W\} && (3.9)\\
&+E_{P_0}\{\bar{Q}_0(A,W,1)M(g_0)(A,W)|W\} - \Psi(P_0), && (3.10)
\end{aligned}$$

where the terms (3.7)-(3.10) are denoted by $D_1(P_0)$, $D_2(P_0)$, $D_3(P_0)$, and $D_4(P_0)$; respectively, and correspond to the orthogonal decomposition of the efficient influence curve implied by the factorization of the likelihood in (3.2).

This decomposition of the score is going to be useful later on during the construction of a targeted maximum likelihood estimator of ψ_0, to define the correct parametric fluctuations. The following result provides the conditions under which an estimator that solves the efficient influence curve equation is consistent.

Result 4. *Let $D(O|\bar{Q},g,\phi,\psi_0)$ be the estimating equation implied by the efficient influence function of result 3:*

$$\begin{aligned}
D(O|\bar{Q},g,\phi,\psi_0) =& \frac{C}{\phi(A,W)} M(g)(A,W)\left\{Y - \bar{Q}(A,W,1)\right\} + M(g)(A,W)\times\\
&\left\{\bar{Q}(A,W,1) - \frac{E_P\{\bar{Q}(A,W,1)I_{\delta_1,\delta_2}(A)|W\}}{G\{\delta_1,\delta_2|W\}}\right\} + \frac{E_P\{\bar{Q}(A,W,1)I_{\delta_1,\delta_2}(A)|W\}}{G\{\delta_1,\delta_2|W\}} - \psi_0. \quad (3.11)
\end{aligned}$$

We have that $E_{P_0}D(O|\bar{Q},g,\phi,\psi_0) = 0$ if and only if $K(g) = K(g_0)$ and either $\bar{Q} = \bar{Q}_0$ or $\phi = \phi_0$.

As a consequence of result 4, a substitution estimator of $\Psi(P_0)$ that solves the efficient influence curve equation will be consistent if and only if $K(g_0)$ and either $\bar{Q}_0$ or ϕ_0 are estimated consistently, and it will be efficient if and only if all of the estimators for $K(g_0)$, $\bar{Q}_0$ and ϕ_0 are consistent. The robustness of this estimating equation is then tied to robustness of the estimator for $K(g_0)$. This consistency condition on the initial estimator g_n is weaker than the conditions needed for other methods for continuous exposures (e.g., the marginal structural models of Neugebauer and van der

Laan (2007)); we only need an estimator g_n that is consistent in the sense that $K(g_n) \to K(g_0)$, which is much weaker than the condition of $g_n \to g_0$ required for marginal structural models. This is because $K(g_0)$ only depends on the conditional probabilities $G\{\delta_1|W\}$ and $G\{\delta_1, \delta_2|W\}$, which can be consistently estimated by a misspecified estimator of the density.

An additional advantage with respect to marginal structural models and other methods for continuous variables (Petersen et al., 2010) is given by the positivity assumption needed to identify and estimate the parameter of interest. The positivity assumption required to estimate marginal structural models is

$$\sup_{a \in \mathscr{A}} \frac{h(a)}{g_0(a|W)} < \infty, -\ a.e.,$$

for a user-specified weight function h. The function $h(a) = 1$ is commonly used, since it implies giving equal weights to all the possible treatment values. The positivity assumption needed to identify and estimate our parameter of interest is given by

$$G_0\{\delta_1, \delta_2|W\} > 0, -\ a.e.,$$

which is a condition that depends on the choice of the interval (δ_1, δ_2) and its probability under G_0, and is thus more likely to be true than positivity of the density g_0 for all the values $a \in \mathscr{A}$.

3.2 Common practice

An alternative formulation of the causal problem of assessing the effect of a truncation in the exposure, which is the current standard in applications of causal inference methods (e.g., Brotman et al., 2008; Bryan, Yu, and van der Laan, 2004; Joffe et al., 2004; Tager et al., 2004), is given by the use of a dichotomous version $A^* = I(A < \delta_2)$ of the continuous treatment variable. The effect of a truncation of A is evaluated in terms of the static intervention $A^* = 1$, and the parameter is defined as $E\{E(Y|A^* = 1, W)\}$, which corresponds (as proven by Stitelman, Hubbard, and Jewell, 2010b) with a stochastic intervention on A in which g_0 is changed to

$$P_\delta(g_0)(A_{P_\delta} = a|W) = \begin{cases} g_0(a|W)/G\{\delta_2|W\} & \text{if } a < \delta_2 \\ 0 & \text{otherwise ,} \end{cases}, \tag{3.12}$$

which is equal to (3.5) only if $G\{\delta_1|W\} = 0$. This means that $E\{E(Y|A^* = 1, W)\}$ measures the effect of a policy that will cause a truncation in the exposure, but will relocate the mass of the non-compliers (i.e., $G\{\delta_2|W\}$) across all the values below δ_2. As a consequence, the two parameters assess policies with different hypothetical effects on the density of the exposure; it is the researcher's responsibility to judge which option is a more likely post-intervention distribution for the policy that is being evaluated.

For instance, in section 3.4 we estimate the effect of a policy that enforces pollutant levels below a predefined threshold. Under such a policy, individuals polluting above the threshold will only have an incentive to reduce their pollution levels to a value that is in accordance with the policy, having no further incentive to go below the enforced cut-off point once they have reached it.

Therefore, the most likely post intervention distribution for this policy is one that locates the probability mass associated to the non-compliers around the cut-off point, i.e., intervention (3.5). The use of intervention (3.12) in this example could lead to misleading conclusions. As an example, consider the following data generating mechanism

$$\begin{aligned} W_1 \sim U\{0,1\}; \quad W_2 \sim Ber\{0.7\}; \quad W_3 \sim N\{W_1, .25\exp(2W_1)\} \\ A \sim Beta\{S_1(W), S_2(W)\} \\ \bar{Q}(A,W) = \text{expit}\{1 + W_1 + 1.5A + 2AW_1 + .5AW_2 - 2W_1W_2 + .2W_1W_3\}, \end{aligned}$$

where we consider four different values for S_1 and S_2: (1) $S_1(W) = S_2(W) = S(W)$, (2) $S_1(W) = S(W)$ and $S_2(W) = \text{expit}\{S(W)\}$, (3) $S_1(W) = \text{expit}\{S(W)\}$ and $S_2(W) = S(W)$; and (4) $S_1(W) = \text{expit}\{S(W)\}$ and $S_2(W) = \text{expit}\{S(W)\}$; for $S(W) = 2.5 + .6W_1 + .3W_2W_3 - .2W_1W_3 - .1(1 - W_2)W_3$. This four scenarios provide four different shapes of the beta distribution: (1) symmetric bell-shaped, (2) skewed to the left, (3) skewed to the right; and (4) symmetric U-shaped. For these four scenarios, table 4.4 shows the parameter $E(Y_{P_\delta} - Y)$ under interventions (3.5) and (3.12) for $(\delta_1, \delta_2) = (0.8, 0.9)$, providing a situation in which the conclusions obtained from the two analysis are very different. In this example the two effects are fairly similar when $G\{\delta_2|W\} \approx 1$, i.e., models (1) and (2). The use of the standard practice of dichotomizing the exposure would lead to misleading results, particularly for models (3) and (4).

Intervention	(1)	(2)	(3)	(4)
(3.5)	-0.0001	-0.0001	-0.0019	-0.0009
(3.12)	-0.0009	-0.0002	-0.0112	-0.0111

Table 3.1: Parameter values under different scenarios.

3.3 Estimators

Initial estimators

In this section we present three estimators for the parameter defined in (3.6). The TMLE and the A-IPTW estimators solve the efficient influence curve equation and inherit the properties derived from result 4. The IPTW is inefficient, and will be consistent only if the estimator of ϕ_0 is consistent. The TMLE is expected to perform better than the A-IPTW if the positivity assumption $\sup_{a\in\mathscr{A}} \phi_0(A,W) > 0, -\ a.e.$ is violated.The finite sample properties of these estimators have been studied elsewhere (e.g., Porter et al., 2011; Rose and van der Laan, 2011).

The estimators presented in this section require initial estimates of $\bar{Q}_0$, g_0 and ϕ_0, which can be obtained through machine learning techniques, parametric or semi-parametric models. The consistency of these initial estimators will determine the consistency and efficiency of the estimators of ψ_0, as discussed previously. Parametric models are commonly used for the sole sake of their convenient analytical properties, but they encode assumptions on the distribution of the data that are not legitimate knowledge about the phenomenon under study and usually cause a large amount of bias in the estimated parameter. As an alternative, we recommend the use of machine learning techniques such as the super learner (van der Laan, Polley, and Hubbard, 2007). Super learner is a methodology that uses cross-validated risks to find an optimal estimator among a library defined by the convex hull of a user-supplied list of candidate estimators. One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution). Proofs and simulations regarding these and other asymptotic properties of the super learner can be found in van der Laan, Dudoit, and Keles (2004) and van der Laan and Dudoit (2003). We will assume that g_0 is estimated consistently in the sense that $K(g_n) \to K(g_0)$.

Influence curve based variance estimators are provided for these three estimators. Consistency of the variance estimators also depends on the consistency of the initial estimators of $\bar{Q}_0$, and ϕ_0. These dependency can be avoided at the cost of computational time and effort by using bootstrapped estimates of the variance.

IPTW

Given an estimator g_n^0 of the exposure density g_0, and an estimator ϕ_n^0 of the missing mechanism, the IPTW estimator of ψ_0 is defined as

$$\psi_{n,1} = \frac{1}{n}\sum_{i=1}^{n} \frac{C_i}{\phi_n^0(A_i, W_i)} M(g_n^0)(A_i, W_i)Y_i.$$

The IPTW is an asymptotically linear estimator with influence curve

$$D_{IPTW}(O|g_0, \phi_0, \psi_0) = \frac{C}{\phi_0(A, W)} M(g_0)(A, W)Y - \psi_0,$$

therefore the variable $\sqrt{n}(\psi_{n,1} - \psi_0)$ converges in distribution to $N(0, P_0 D^2_{IPTW}(g_0))$, whose variance can be estimated as the empirical variance of $D^2_{IPTW}(O|g_n^0, \phi_n^0, \psi_{n,1})$. This is a conservative estimator of the variance of the IPTW, as proven in van der Laan and Robins (2003).

Augmented IPTW

The augmented IPTW is the value $\psi_{n,2}$ that solves the equation $\sum_{i=1}^{n} D(O_i|\bar{Q}_n^0, g_n^0, \phi_n^0, \psi_0) = 0$, for initial estimates $\bar{Q}_n^0$, g_n^0 and ϕ_n^0 of $\bar{Q}_0$, g_0 and ϕ_0.

$$\psi_{n,2} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{C_i}{\phi_n^0(A_i, W_i)} M(g_n^0)(A_i, W_i)\left\{Y_i - \bar{Q}_n^0(A_i, W_i, 1)\right\} + M(g_n^0)(A_i, W_i) \times \right.$$
$$\left.\left\{\bar{Q}_n^0(A_i, W_i, 1) - \frac{E_{g_n^0}\{\bar{Q}_n^0(A, W, 1)I_{\delta_1,\delta_2}(A)|W_i\}}{G_n^0\{\delta_1, \delta_2|W_i\}}\right\} + \frac{E_{g_n^0}\{\bar{Q}(A, W, 1)I_{\delta_1,\delta_2}(A)|W_i\}}{G_n^0\{\delta_1, \delta_2|W_i\}}\right]. \quad (3.13)$$

If the initial estimators are consistent, the A-IPTW is an asymptotically linear estimator with influence curve $D(O|\bar{Q}_0, g_0, \phi_0, \psi_0)$. As in the case of the IPTW, the variable $\sqrt{n}(\psi_{n,2} - \psi_0)$ converges in law to a random variable with distribution $N\{0, P_0 D^2(\cdot|\bar{Q}_0, g_0, \phi_0, \psi_0,)\}$, whose variance can be estimated as the empirical variance of $D^2(O|\bar{Q}_n^0, g_n^0, \phi_n^0, \psi_{n,2})$. Rose and van der Laan (2011, Appendix 18) show that inference based on this variance estimator is valid only if ϕ_n^0 is consistent, providing exact inference when $\bar{Q}_n^0$ is consistent, and conservative inference when $\bar{Q}_n^0$ is inconsistent.

Targeted maximum likelihood estimator

Targeted maximum likelihood estimation van der Laan and Rubin (2006) is a loss-based semiparametric estimation method that yields a substitution estimator of a target parameter of the probability distribution of the data that solves the efficient influence curve estimating equation, and thereby yields a double robust locally efficient estimator of the parameter of interest, under regularity conditions.

In order to define a targeted maximum likelihood estimator for ψ_0, we need first to define three elements: (1) A loss function $L(Q)$ for the relevant part of the likelihood required to evaluate $\Psi(P)$, which in this case is $Q = (\bar{Q}, g, Q_W)$. This function must satisfy $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$, where Q_0 denotes the true value of Q; (2) An initial estimator Q_n^0 of Q_0; (3) A parametric fluctuation $Q(\varepsilon)$ through Q_n^0 such that the linear span of $\frac{d}{d\varepsilon} L\{Q(\varepsilon)\}|_{\varepsilon=0}$ contains the efficient influence curve $D(P)$ defined in result (3). These elements are defined below:

Loss Function

As loss function for Q, we will consider $L(Q) = L_Y(\bar{Q}) + L_A(g) + L_W(Q_W)$, where for continuous Y we set $L_Y(\bar{Q}) = \{Y - \bar{Q}(A,W,C)\}^2$, for binary Y we set $L_Y(\bar{Q}) = Y \log\{\bar{Q}(A,W,C)\} + (1 - Y)\log\{1 - \bar{Q}(A,W,C)\}$, $L_A(g) = -\log g(A|W)$, and $L_W(Q_W) = -\log Q_W(W)$. It can be easily verified that this function satisfies $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$.

Parametric Fluctuation

Given an initial estimator Q_n^k of Q_0, with components $(\bar{Q}_n^k, g_n^k, Q_{W,n}^k)$, and an initial estimator ϕ_n^0 of ϕ_0, we define the $(k+1)$th fluctuation of Q_n^k as follows:

$$\begin{aligned} m\{\bar{Q}_n^{k+1}(\varepsilon_1)(A,W)\} &= m\{\bar{Q}_n^k(A,W)\} + \varepsilon_1 H_1^k(A,W) \\ g_n^{k+1}(\varepsilon_1)(A|W) &\propto \exp\{\varepsilon_1 H_3^k(A,W)\} g_n^k(A|W) \\ Q_{W,n}^{k+1}(\varepsilon_2)(W) &\propto \exp\{\varepsilon_2 H_4^k(W)\} Q_{W,n}^k(W), \end{aligned}$$

where

$$H_1^k(A,W) = \frac{C}{\phi_n^0(A,W)} M(g_n^k)(A,W),\ H_3^k(A,W) = D_3(P^k)(O), \text{ and } H_4(W) = D_4(P^k)(O),$$

with D_3 and D_4 defined as in result 3, and m is the identity or logit function depending on whether the outcome is continuous or binary. Note that this fluctuation satisfies the condition $D(P) \in < \frac{d}{d\varepsilon} L\{Q(\varepsilon)\}|_{\varepsilon=0} >$, which is a key element of targeted minimum loss based estimation.

Targeted Maximum Likelihood Estimator
The TMLE is defined by the following iterative process:

1. Initialize $k = 0$.
2. Estimate ε as $\varepsilon_n^k = \arg\min_\varepsilon P_n L\{Q_n^k(\varepsilon)\}$.
3. Compute $Q_n^{k+1} = Q_n^k(\varepsilon_n^k)$.
4. Update $k = k+1$ and iterate steps 2 through 4 until convergence (i.e., until $\varepsilon_n^k = 0$)

First of all, note that the value of ε_2 that minimizes the part of the loss function corresponding to the marginal distribution of W in the first step (i.e., $-P_n \log Q_{W,n}^1(\varepsilon_2)$) is $\varepsilon_2^1 = 0$. Therefore, the iterative estimation of ε only involves the estimation of ε_1. The kth step estimation of ε_1 is obtained by numerically minimizing $P_n(L_Y(\bar{Q}_n^k(\varepsilon_1)) + L_A(g_n^k(\varepsilon_1)))$.

The TMLE of ψ_0 is defined as $\psi_{n,3} \equiv \lim_{k\to\infty} \Psi(P_n^k)$, assuming this limit exists. In practice, the iteration process is carried out until convergence in the values of ε_k is achieved, and an estimator Q_n^* is obtained. The variance of $\psi_{n,3}$ can be estimated by the empirical variance of $D^2(O|\bar{Q}_n^*, g_n^*, \phi_n^0, \psi_{n,3})$, which is a consistent estimator only if both ϕ_n^0 and $\bar{Q}_n^*$ are consistent, is conservative if ϕ_n^0 is consistent but $\bar{Q}_n^*$ is not, and is inconsistent in any other case.

3.4 Extension to longitudinal data and application

Longitudinal interventions

Assume now that the observed data structure is the same presented in section 3.1, but now we have repeated measures in the sense that for each subject the observed variables were recorded at time points $t = 1, \ldots, T$. That is, the observed data in this case can be described as a vector $O = (W_t, A_t, C_t, C_tY_t : t = 1, \ldots, T) = (O_t : t = 1, \ldots, T)$. We can now define a time specific counterfactual outcome given by $Y_{t,P_{t,\delta}}$, where the stochastic intervention of interest is performed by changing each time-specific exposure mechanism $g_{t,0}$ to $P_{t,\delta}$, with $P_{t,\delta}$ analogous to P_δ in (3.5). The parameter of interest can be defined now as a causal effect based on a marginal structural model (MSM, Neugebauer and van der Laan, 2007) with only intercept:

$$\beta_0 = \arg\min_\beta \sum_{t=1}^{T} \{E_0(Y_{t,P_{t,\delta}}) - m_\beta(t)\}^2 w(t),$$

where we set $m_\beta(t) = \beta$, and $w(t)$ is a weight function initially set to $1/T$. For this case (usually called intercept only model), our parameter of interest reduces to

$$\beta_0 = \sum_{t=1}^{T} w(t) E_0(Y_{t,P_\delta}), \tag{3.14}$$

which is the weighted average of the time specific causal effects. This parameter provides a measure of the overall effect of a policy when applied repetitively at every time point. Parameters given by more complex marginal structural models motivated by different research questions can also be defined by an appropriate MSM, for example the expectation of the counterfactual outcome at the last time point in the study, which will provide a measure of the final effect of implementing a given policy during T units of time, or an MSM that takes measures the possible trend in the expectation of the counterfactual outcome.

The efficient influence curve of parameter (3.14) is given by the weighted average of the time point specific influence curves:

$$D_\beta(O|\bar{Q},g,\phi,\beta_0)=\sum_{t=1}^{T} w(t)D(O_t|\bar{Q}_t,g_t,\phi_t,\psi_{t,0}), \tag{3.15}$$

where D is defined in (3.11) and $\bar{Q}_t, g_t, \phi_t$ and $\psi_{t,0}$ denote the conditional expectation of the outcome, exposure mechanism, missingness mechanism and expectation of the counterfactual outcome for each time specific data structure. Estimators that solve the efficient influence curve equation

$$\sum_{i=1}^{n}\sum_{t=1}^{T} w(t)D(O_{it}|\bar{Q}_t,g_t,\phi_t,\psi_{t,0}), \tag{3.16}$$

inherit the consistency and efficiency properties of estimators mentioned in result 4, where the consistency conditions are now replaced by consistency in the estimation of all the time specific mechanisms $\bar{Q}_{t,0}, g_{t,0}$ and $\phi_{t,0}$. To estimate each of these initial parameters we can choose to fit different estimators for each time point, or we can also choose to do smoothing over t, by including it as a covariate in each of the conditional expectations and probabilities involved.

Estimation of the parameter in (3.14) can now be performed by applying the estimators presented in section 3.3 to a pooled dataset in which time has been added as a covariate and each row corresponds to a specific subject time point combination. The IPTW estimator, for example, would now be given by

$$\psi_{n,1}=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{C_{it}}{\phi^0_{n,t}(A_{it},W_{it})}M(g^0_{n,t})(A_{it},W_{it})Y_{it},$$

and the augmented IPTW by

$$\begin{aligned}\psi_{n,2}=\frac{1}{nT}\sum_{i=1}^{n}\sum_{i=1}^{T}\Bigg[&\frac{C_{it}}{\phi^0_n(A_{it},W_{it})}M(g^0_{n,t})(A_{it},W_{it})\left\{Y_{it}-\bar{Q}^0_{n,t}(A_{it},W_{it},1)\right\}+M(g^0_{n,t})(A_{it},W_{it})\times\\ &\left\{\bar{Q}^0_{n,t}(A_{it},W_{it},1)-\frac{E_{g^0_{n,t}}\{\bar{Q}^0_{n,t}(A,W,1)I_{\delta_1,\delta_2}(A)|W_{it}\}}{G^0_{n,t}\{I_{\delta_1,\delta_2}(A)|W_{it}\}}\right\}+\\ &\frac{E_{g^0_{n,t}}\{\bar{Q}(A,W,1)I_{\delta_1,\delta_2}(A)|W_{it}\}}{G^0_{n,t}\{I_{\delta_1,\delta_2}(A)|W_{it}\}}\Bigg],\end{aligned} \tag{3.17}$$

which can be seen to solve equation (3.16). The TML estimator is defined analogous to the definition given in the previous section, with $\bar{Q}_n^k$, H_1^k, ϕ_n^k, H_3^k, Q_W^k, and H_4^k replaced by their t-specific counterparts. However, the same parameters ε_1 and ε_2 are used to fluctuate all these t-specific estimates. Estimation of ε in step 2 of the iterative process that defines the TMLE is performed now with respect to the empirical distribution P_{nT} given by the pooled dataset, and the estimating equation in result (4) is replaced by its counterpart summing also over t and with t-specific estimated values of $\bar{Q}_n^k$, H_1^k, ϕ_n^k, H_3^k, Q_W^k, and H_4^k. The estimators of the variance of these estimators presented in the previous section can also be adapted to these longitudinal estimators. Remarks about consistency of the variance estimators of section 3.3 carry on to these variance estimators.

Application

In this section we present the results of applying the method for longitudinal data described in the previous section to assess the effect of a program that constrains air pollution levels on wheezing in children with asthma. These data were originally analyzed by Mann et al. (2010) as part of the Fresno Asthmatic Childrens Environment Study (FACES). In the original chapter whose objective was to evaluate whether exposure to ambient pollution is associated with increased respiratory symptoms, wheeze was found to be associated with short-terms exposures to NO_2 with an odds ratio of 1.10 (C.I. (1.02, 1.20)) for a 8.7 parts per billion increase. The data consisted of a sample of 315 children between 6 and 11 years of age who have active asthma. Reports of morning wheeze were collected for 14 days, up to three times a year, from December 2000 through March 2005, which lead to approximately 12 data panels for each child. For a comprehensive description of the study, the interested reader is referred to the original chapter.

We are interested in investigating the effect of NO_2 concentrations measured 24 hours before each visit on the current presence of wheezing. The confounders we considered (i.e., W variables) are: gender, age, race, height, low birth weight, born prematurely, atopy, presence of eczema, rhinitis, mother smoked during pregnancy, whether child was ever breastfed, presence of asthma in father and mother, no smoking policy in the house, anyone smokes in the house, relative humidity, temperature, season of the year, whether the house is rented or owned and income.

We estimated the effect of a policy that enforces NO_2 levels below 28.15 ppb. We assume that such intervention will produce a change in the population distribution of the exposure corresponding to a relocation of the probability mass originally above 28.15 ppb between 26.05 and 28.15 ppb in the intervened population. The values 28.15 and 26.05 ppb correspond with the 85th and 80th percentile of the distribution of NO_2, respectively.

If the objective is to perform a comparison of the prevalence of wheezing in the hypothetical intervened population with the prevalence in the current population, we can define a population intervention parameter ψ_0^1 as $\psi_0^1 = \psi_0 - \mu_0$, where $\mu_0 = E_{P_0}(Y)$. This parameter compares the expectation of the outcome under the policy of interest with its current expectation, and therefore provides a measure of the gain obtained by implementing the policy.

Since we observed a coarsened version of Y, we cannot use the empirical mean as an estimator of μ_0. Because estimation of this expectation is equivalent to estimation of the expectation of the outcome under the intervention $C = 1$, we suggest the use of the TMLE for static interventions as described in Rose and van der Laan (2011, chapter 4). Such estimator also utilizes initial estimators of $\phi_0(A,W)$ and $\bar{Q}_0(A,W,C)$, and is double robust under misspecification of either model. For further details about the properties and implementation the TMLE for μ_0, the reader is referred to the original sources.

For a given estimator ψ_n of ψ_0, and an asymptotically linear estimator μ_n of $E_{P_0}(Y)$ with influence curve $D_\mu(P)$, an asymptotically linear estimator of ψ_0^1 is given by $\psi_n^1 = \psi_n - \mu_n$. Its influence curve can be computed as $D_{\psi^1}(P)(O) = D_\psi(P)(O) - D_\mu(P)(O)$, and its variance can be estimated through the sample variance of $D_{\psi^1}(P)(O)$. Here $D_\psi(P)(O)$ represents the influence curve of each of the estimators defined in section 3.3. The estimates of ψ_0^1 and their standard errors are presented in Table 3.2. Confidence intervals and p-values for hypothesis testing can be computed based on the normal approximations for asymptotically linear estimators described in section 3.3. In light of the theoretical properties of these estimators, we rely on the TMLE and A-IPTW to measure the effect of the intervention of interest. The estimated value of ψ_n^1 means that under a policy that enforces places with NO_2 levels above 28.15 to decrease their levels to some value in the interval $(26.05, 28.15)$, the prevalence of wheezing in children with asthma between 6 and 11 years of age would be reduced by 0.50%. However, our estimated effect is not significant at a 95% confidence level, which does not mean that the effect is inexistent or epidemiologically irrelevant.

Table 3.2: Estimates of ψ_0^1 and ψ_0 (in %).

	TMLE	A-IPTW	IPTW
ψ_0^1	0.50 (0.40)	0.15 (0.89)	0.99 (1.04)
ψ_0	13.53 (0.56)	13.17 (0.99)	14.63 (1.14)

3.5 Conclusion

In this chapter we propose a specific type of causal parameter defined by a stochastic intervention in terms of a truncation of the original distribution of the exposure. We present an application example in which the effect of a potential policy enforcing pollution levels under certain threshold is measured. Our approach allows the estimation of the effect of potential policies that result in stochastic interventions (for example because they fail to put every subject in a predefined exposure level). We argue that our parameter makes more sense from a policy– and decision–making point of view as compared to current practice.

The stochastic interventions framework allowed us to naturally define an effect for a continuous exposure, which is a topic that has received little attention in the causal inference literature. Assumptions like the positivity assumption and the consistency of an initial estimator for the exposure mechanism are weakened as compared to those required for estimating other causal parameters for continuous or categorical exposures. Two consistent and efficient estimators for the parameter of interest were proposed, and their use was illustrated with an example.

Chapter 4

Data adaptive estimation of the causal dose response curve

Estimating the causal effect of an exposure A on an outcome Y when the relation between them is confounded by a set of covariates is a very common problem in causal inference, of high relevance for applications in epidemiology, medical, and social research, among other fields.

Causal effects in this setting are defined as parameters of the distribution of the counterfactual outcome (see, for example Rubin, 1974; Pearl, 2000) Y_a that would have been observed if, possibly contrary to the fact, the subject would have received level a of the exposure. Computation of causal parameters involves expectations with respect to the distribution of the stochastic process that one would have observed if, for each subject, all the counterfactual outcomes were observed. Since the observed data contains only one of the counterfactuals, namely $Y = Y_A$, additional untestable assumptions are needed in order to identify parameters of the counterfactual process distribution as parameters of the observed data distribution. These assumptions are usually described in terms of the so-called no unmeasured confounders assumption, a particular case of the coarsening at random assumption, which roughly states that the censoring or exposure processes cannot depend on unobserved covariates that are also related to the outcome.

In spite of the large number of causal inference problems that are inherently defined in terms of exposures of continuous nature, most of the attention in the field of causal inference has focused in the definition and estimation of parameters for binary treatments, in which it is natural to compare the counterfactual outcome under two possible exposure levels. Estimation of causal parameters for binary exposures has been widely studied (e.g., Rubin, 1978; Rosenbaum and Rubin, 1983; Robins, 1986; van der Laan and Robins, 2003; Rubin, 2006; R. Mansson, 2007; Rose and van der Laan, 2011). The main reason why consistent and efficient estimators of the causal dose response curve (CDRC) for continuous treatments in the nonparametric model have not yet been developed is that it is not a pathwise differentiable parameter (see Bickel et al., 1997, chapter 3, 5), and therefore cannot be estimated at a consistency rate of $n^{-1/2}$. Examples of pathwise differentiable parameters that measure the causal effect of a continuous exposure on an outcome of interest are given by the parameters defined in chapter 2 and 3. These approaches make use of stochastic interventions (Korb et al., 2004; Didelez, Dawid, and Geneletti, 2006; Dawid and Didelez, 2010) as a means to define a counterfactual outcome in a post-intervened world, which compared to the expectation of the actual outcome defines the causal effect of an intervention.

The most widely known method for estimation of the CDRC for continuous exposures is the so called marginal structural model (MSM) framework, which was first proposed by Robins, Hernan, and Brumback (2000), and whose validity relies on the correct specification of a parametric model for the CDRC. Neugebauer and van der Laan (2007) generalize this setting to avoid dependence on the correct specification of a parametric model by defining the parameter of interest as the projection of the true CDRC on the space of functions defined by the parameterization implied by the MSM, providing robustness against misspecification of the parametric MSM. Their work also includes identification results for this projection parameter, as well as IPTW, G-Comp and augmented IPTW double robust estimators. Marginal structural models represent only a provisional solution to the problem, because in many instances the interest relies on estimating the actual CDRC and not its projection on some parametric space of functions.

An alternative and widely used method for estimating non pathwise differentiable parameters is the selection of the best performing candidate among a list of algorithm estimators, where performance is defined in terms of the cross-validated risk. Formal analytical asymptotic arguments backing the use of cross-validation as an estimator selection tool were first given by van der Laan and Dudoit (2003); van der Vaart (2003); van der Laan, Dudoit, and Keles (2004), among others. The main result of these works is a finite sample size inequality that bounds the risk of the cross-validation selector by the risk of the oracle selector (the selector based on the true distribution), which is in turn used to establish, under certain conditions, the asymptotic equivalence between the cross-validation and the oracle selectors. These results are later explored in specific contexts by Dudoit and van der Laan (2005); van der Laan, Dudoit, and van der Vaart (2006), among others. Of special interest is the work of van der Vaart, Dudoit, and van der Laan (2006), in which the cross-validation oracle inequalities are extended to candidate libraries with a continuous index set and unbounded loss functions. van der Laan, Dudoit, and van der Vaart (2006) demonstrates that this oracle property for cross-validation combined with the right library of estimators results in a minimal adaptive optimal estimator. van der Laan, Polley, and Hubbard (2007) use these optimality results in order to define the super learner prediction algorithm, implemented in the `SuperLearner` R library. van der Laan and Petersen (2012) describe a general methodology in which the CV-A-IPTW estimators of the risk are replaced by CV-TMLE estimators.

For the particular case of the CDRC, van der Laan and Dudoit (2003, pag. 52) proof that under convergence of the initial estimators, the candidate selector based on the cross validated A-IPTW risk is asymptotically equivalent to the oracle selector. Since A-IPTW estimators are not substitution estimators, they can fall outside the parameter space, and are very sensitive to violations of the positivity assumption. Violations to the positivity assumption are very likely to occur when working with continuous exposures, since the exposure mechanism is now a conditional density.

The main contribution of this chapter is to present a cross validated targeted minimum loss based estimator of the risk of a CDRC candidate estimator that is endowed with an oracle inequality analogue to that of the A-IPTW. The CV-TMLE we propose is more robust to empirical violations of the positivity assumption, and it is a substitution estimator, which guarantees estimates that are within the bounds of the parameter space. These two estimators have also been proven to be asymptotically linear with influence function equal to the efficient influence function, under certain conditions, which implies that they are consistent and efficient estimators of the risk.

The chapter is organized as follows. In section 4.1 we formally describe the inference problem, define the loss and risk functions, and present the efficient influence function of the parameter of interest. In order to first introduce relevant concepts, in section 4.1 we present four estimators (G-comp, IPTW, A-IPTW, TMLE) of the risk when the candidate estimators of the CDRC are assumed fixed functions. In section 4.1 we generalize these estimators to the case when the candidates are estimated from the sample, and present the corresponding cross-validated versions of the A-IPTW and TMLE. In section 4.2 we present a theorem describing the conditions under which the CV-TML estimator of the risk is an asymptotically linear estimator, the conditions under which it is consistent and efficient, as well as a discussion on the estimation of its variance. section 4.3 presents the main contribution of this chapter; an oracle inequality for the selector based on the CV-TML estimator of the risk, and the conditions under which it is asymptotically equivalent to the oracle selector. Finally, in section 4.4 we use Monte Carlo simulation to compare the performance of CV-TMLE and CV-A-IPTW selectors and estimators of the risk in finite sample sizes.

4.1 Definition and estimation of the risk of an estimator of the CDRC

Consider an experiment in which an exposure variable A, a continuous or binary outcome Y and a set of covariates W are measured for n randomly sampled subjects. Let $O = (W,A,Y)$ represent a random variable with distribution P_0, and $O_1,\dots,O_n$ represent n i.i.d. observations of O. The range of W, A and Y will be denoted by $\mathscr{W}$, $\mathscr{A}$ and $\mathscr{Y}$, respectively. Assume that the following non-parametric structural equation model (NPSEM) holds:

$$W = f_W(U_W);\ A = f_A(W,U_A);\ Y = f_Y(A,W,U_Y), \tag{4.1}$$

where U_W, U_A and U_Y are exogenous random variables such that $U_A \perp\!\!\!\perp U_Y$ holds, and either $U_W \perp\!\!\!\perp U_Y$ or $U_W \perp\!\!\!\perp U_A$ holds (randomization assumption). The true distribution P_0 of O can be factorized as

$$P_0(O) = P_0(Y|A,W)P_0(A|W)P_0(W),$$

where we denote $g_0(A|W) \equiv P_0(A|W)$, $\bar{Q}_{1,0}(A,W) \equiv E_0(Y|A,W)$, $\bar{Q}_{2,0}(A,W) \equiv E_0(Y^2|A,W)$, $Q_{W,0}(W) \equiv P_0(W)$, and $Pf = \int f dP$ for a given function f. For a given value $a \in \mathscr{A}$, the counterfactual of Y is defined as the value $Y_a = f_Y(a,W,U_Y)$, the counterfactual process of Y is given by $(Y_a : a \in \mathscr{A})$, and the full data is denoted by $X = \{W,(Y_a : a \in \mathscr{A})\} \sim F_0$.

In this chapter we will discuss the estimation of the causal dose-response curve within strata of the covariates $Z \subset W$, given by the expression

$$\Psi^f(F_0)(a,Z) = E_{F_0}(Y_a|Z) = \arg\min_{\psi} R^f(\psi,F_0), \tag{4.2}$$

where $R^f(\psi,F_0) = F_0L^f(\psi)$, $L^f(\psi)(X) = \int_{\mathscr{A}}\{Y_a - \psi(a,Z)\}^2 h(a,Z)d\mu(a)$, the superscript f stands for full data, and h is a non-negative function such that $\int hd\mu = 1$. The second equality in (4.2) is true because $E_{F_0}(Y_a|Z)$ is the projection of Y_a into the space of functions of Z, and $F_0L^f(\psi)$ is the integral over $\mathscr{A}$ of the squared norm of $Y_a - \psi(a,Z)$. The randomization assumption implies that $Y_a \perp\!\!\!\perp A|W$, which allows identification of the full data parameter (4.2) in terms of a function of the observed data distribution as the mapping

$$\Psi(P)(a,Z) = E_P\{\bar{Q}(a,W)|Z\}, \tag{4.3}$$

where we denote $\psi_0 = \Psi(P_0)$. If A is continuous, $\Psi(P)$ is not a pathwise differentiable parameter in the non parametric model, and $\sqrt{n}-$consistent estimation is not possible (Bickel et al., 1997, chapter 3, 5). However, the risk of a given candidate value ψ_k, is a pathwise differentiable parameter for which it is possible to find regular asymptotically linear estimators.

Following the ideas of Wang, Bembom, and van der Laan (2006), consider a list of candidates values $\psi_k : k = 1,\ldots,K_n$ for ψ_0^f. Throughout the chapter we will make a distinction between candidate values (denoted ψ_k) and candidate estimators (denoted $\hat{\Psi}_k$), where the difference is that the former are given functions, whereas the latter are functions of (a,Z) estimated from the sample.

If the full data X were observed, a general selection procedure would involve computing $R^f(\psi_k,F_0) : k = 1,\ldots,K_n$, and estimating ψ_0^f based on ψ_{k_0}, where $k_0 = \arg\min_k R^f(\psi_k,F_0)$. Of course this optimization procedure cannot be carried out as described previously, because: 1) only a coarsened version of X denoted by O is observed, 2) the distribution P_0 of O is unknown, and 3) in most cases we have a list of candidate estimators $\hat{\Psi}_k$, as opposed to a list of candidate values ψ_k, which arises the issue of over-fitting.

In order to overcome these obstacles one needs to:

1. Find a mapping $R(\psi,\cdot) : \mathscr{M} \to \mathbb{R}$ that identifies R^f, i.e., a mapping such that $R(\psi,P_0)$ equals $R^f(\psi,F_0)$, under certain assumptions. It is common that $R(\psi,P) = PL_{\Gamma(P)}(\cdot,\psi)$ for a loss function $L_{\Gamma(P)}$ that is now indexed by a nuisance parameter $\Gamma : \mathscr{M} \to \mathscr{F}_\gamma$.

2. If P_0 is known, the value $R(\psi,P_0)$ suffices to find a selector among the K_n candidate values. However, since P_0 is unknown, we now need to estimate $R(\psi,P_0)$. At this point it is worth to note that even though $\Psi(P)$ is not a pathwise differentiable parameter, the mapping $R(\psi,\cdot)$ is pathwise differentiable, and can therefore be $\sqrt{n}$-consistently estimated under regularity conditions.

3. If candidate values are not available it is necessary to estimate the risk of candidate estimators $\hat{\Psi}_k$ that are trained in the sample, which makes necessary the use of cross-validated versions of these estimators.

In the remaining of this section we will discuss the identification of R^f. The risk of a candidate ψ is given by $R^f(\psi,F) = FL^f(\psi)$, and is identified as a function of the observed data distribution by

$$R(\psi,P) = E_PL_{\bar{Q}(P)}(O,\psi), \tag{4.4}$$

where

$$\begin{aligned} L_{\bar{Q}}(O,\psi) =& \int_{\mathscr{A}} E_P\{(Y-\psi(a,Z))^2|A=a,W\}h(a,Z)d\mu(a). \\ &= \int_{\mathscr{A}} \{\bar{Q}_2(a,W)-2\bar{Q}_1(a,W)\psi(a,Z)+[\psi(a,Z)]^2\}h(a,Z)d\mu(a), \end{aligned} \tag{4.5}$$

given the randomization assumption and the positivity assumption

$$\sup_{a\in\mathscr{A}} \frac{h(a,Z)}{g_0(a,W)} < \infty,\ Q_{W,0}-a.e. \tag{4.6}$$

Note that the loss function that defines the risk is not unique, since the loss functions

$$L_g(O,\psi) = \frac{(Y-\psi(A,Z))^2}{g(A,W)}h(A,Z), \tag{4.7}$$

$$\begin{aligned} L_{\bar{Q},g}(O,\psi) =& \frac{h(A,Z)}{g(A,W)}\left[\{Y^2-\bar{Q}_2(A,W)\}-2\psi(A,Z)\{Y-\bar{Q}_1(A,W)\}\right]+ \\ & \int_{\mathscr{A}} \{\bar{Q}_2(a,W)-2\psi(a,Z)\bar{Q}_1(a,W)+\psi^2(a,Z)\}h(a,Z)d\mu(a) \end{aligned} \tag{4.8}$$

lead to the same definition of the risk. Loss functions (4.5) and (4.7) come from more intuitive definitions of the risk, whereas the loss function (4.8) comes from efficient estimation theory, and is closely related to the efficient influence function of $R(\psi,P)$. This fact is exploited by Wang, Bembom, and van der Laan (2006) in order to define estimators of the risk as a cross-validated average of estimators of these loss functions. We will work towards the definition of a CV-TMLE analogue of those estimators, and present similar results to those obtained by van der Laan and Dudoit (2003) in terms of an oracle inequality, as well as the conditions under which the estimator of the risk is asymptotically linear.

The loss function (4.8), referred to as the double robust loss function, defines the efficient influence function of parameter $R(\psi)$ and plays a very important role in double robust and efficient estimation of $R(\psi)$, as explained in the next section.

Parameter (4.4) is a pathwise differentiable parameter, for which consistent asymptotically linear estimators can be found. Note that $R(\psi,P)$, as defined in (4.4), depends on P only through $Q=(\bar{Q},Q_W)$, where $\bar{Q}=(\bar{Q}_1,\bar{Q}_2)$. In an abuse of notation, we will use $R(\psi,P)$ and $R(\psi,Q)$ indistinctly, and the true value $R(\psi,Q_0)$ will be denoted by $R_0(\psi)$. We will also use the notations $R(\psi,Q)$ and $R(\psi)(Q)$ indistinctly. In section 4.1 we will focus on the estimation of the risk when the candidates are given values. Given candidate values constitute a situation that is not very common in research problems, but provides an easy way to introduce the estimators that are going to be developed in section 4.1, in which we will generalize these estimators to the case of a candidate estimated from the sample. Cross validation will be used as a tool to avoid over-fitting, and will lead to an oracle inequality presented in section 4.3.

The efficient influence function of the risk $R(\psi,Q)$ is given by the expression

$$D(Q,g,\psi)(O) = L_{\bar{Q},g}(O,\psi)-R(\psi,Q), \tag{4.9}$$

with $L_{\bar{Q},g}$ defined in (4.8).

Estimators of the risk of a candidate parameter value

In this section we exploit the definitions of the risk in terms of loss functions given in the previous section in order to define various estimators of the risk. As we will see, the definitions of the risk through the different loss functions previously described lead to the definition of G-comp, IPTW and A-IPTW estimators. We will also use the efficient influence function of $R(\psi,P)$ in order to define a targeted maximum likelihood estimator of $R_0(\psi)$. The A-IPTW loss function is closely related to the efficient influence curve of $R(\psi,P)$, which results in the consistency and efficiency of the A-IPTW and TMLE. Analytical properties of these estimators has been discussed elsewhere (van der Laan and Robins, 2003; van der Laan and Rubin, 2006; Rose and van der Laan, 2011).

We will assume that ψ is a given function of a and Z in the sense that it is not estimated from the sample. Such scenario is attainable, for example, in situations in which a pilot study is conducted in order to postulate candidate estimators with the objective of assessing their performance with data from a posterior study.

Let $\hat{\bar{Q}} = (\hat{\bar{Q}}_1, \hat{\bar{Q}}_2)$ and $\hat{g}$ be initial estimators of $\bar{Q}_0 = (\bar{Q}_{1,0}, \bar{Q}_{2,0})$ and g_0, respectively. These estimators will be denoted $\hat{\bar{Q}}$ or $\hat{\bar{Q}}(\mathbb{P})$, depending on whether it is necessary to emphazise their dependence on the empirical distribution

$$\mathbb{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{O_i}$$

with δ_x denoting a Dirac delta with a point mass at x.

G-comp, IPTW and A-IPTW estimators

The equivalent definitions of the risk through G-comp, IPTW and A-IPTW loss functions allow the straightforward definition of three estimators of the risk of a candidate value, given by:

$$\hat{R}^G(\psi) = \frac{1}{n}\sum_{i=1}^{n} L_{\hat{\bar{Q}}}(O_i,\psi),\;\; \hat{R}^I(\psi) = \frac{1}{n}\sum_{i=1}^{n} L_{\hat{g}}(O_i,\psi),\; \text{and } \hat{R}^{DR}(\psi) = \frac{1}{n}\sum_{i=1}^{n} L_{\hat{\bar{Q}},\hat{g}}(O_i,\psi),$$

which can be seen as solutions in R of the corresponding estimating equations $\mathbb{P}D^I(\cdot|\hat{g},\psi,R) = 0$, $\mathbb{P}D^G(\cdot|\hat{\bar{Q}},\psi,R) = 0$, and $\mathbb{P}D^{DR}(\cdot|\hat{\bar{Q}},\hat{g},\psi,R) = 0$, where

$$\begin{aligned} D^I(O|g,\psi,R) =& L_g(O,\psi) - R \\ D^G(O|\bar{Q},\psi,R) =& L_{\bar{Q}}(O,\psi) - R \\ D^{DR}(O|\bar{Q},g,\psi,R) =& L_{\bar{Q}.g}(O,\psi) - R. \end{aligned}$$

According to theorem 5.11 of van der Vaart (2002), if $L_{\hat{g}}$ falls in a Glivenko-Cantelli class $\{L_g : g \in \mathscr{G}\}$ with probability tending to one, and $P_0(L_{\hat{g}} - L_{g_0})^2 \to 0$, then the IPTW estimator is consistent for $R_0(\psi)$. Under an appropriate Donsker condition and consistency of $\hat{g}$, the IPTW estimator is also asymptotically linear with influence function $D^I(O|g_0,\psi,R_0)$, as explained in theorem 6.18 of van der Vaart (2002) and the theorems of chapter 2 of van der Laan and Robins (2003). As a consequence, it is an inefficient estimator of the risk $R_0(\psi)$, and its variance can be estimated with the empirical variance of $D^I(O|\hat{g},\psi,R_0)$. Equivalent statements are also true for the G-comp estimator.

Following similar arguments, the A-IPTW estimator is double robust in the sense that it is consistent if either of $\hat{\bar{Q}}$ or $\hat{g}$ is consistent. It is also efficient if both $\hat{\bar{Q}}$ and $\hat{g}$ are consistent. Even though the A-IPTW represents an important improvement with respect to the G-comp or the IPTW, it suffers from some of the drawbacks inherited from the estimating equation methodology. One of the most important problems of such methodology is the possibility of solutions out of the parameter space, or very unstable estimators if the positivity assumption is practically violated. For this reason we prefer estimators that are substitution estimators, i.e., estimators that are the result of applying the map $R(\psi)$ to a certain estimated distribution $P^* \in \mathscr{M}$. As we will see, the TMLE is such a substitution estimator.

Targeted minimum loss based estimator

For a review on TMLE and its properties we refer the interested reader to Rose and van der Laan (2011). TML estimation requires the specification of three components: a valid loss function for the relevant part of the likelihood, a parametric submodel whose generalized score equals the efficient influence function, and initial estimators of the relevant parts of the likelihood.

We will assume that Y is binary, or that $P(Y \in [a,b]) = 1$ for known values a and b, in which case we can work with $Y^* = (Y-a)/(b-a)$ and interpret the results accordingly. Consider the loss functions $-L_j\{(\bar{Q}_j)(O)\} = Y^j \log \bar{Q}_j(A,W) + \{1-Y^j\}\log\{1-\bar{Q}_j(A,W)\};\ j=1,2$, for $\bar{Q}_j$, and the parametric fluctuations given by $\operatorname{logit}\bar{Q}_j(\varepsilon_j) = \operatorname{logit}\bar{Q}_j + \varepsilon_j H_j(\psi,g)$, where

$$\begin{aligned} H_1(\psi,g)(A,W) =& -2\psi(A,Z)\frac{h(a,Z)}{\hat{g}(A,W)}, \\ H_2(\psi,g)(A,W) =& \frac{h(a,Z)}{\hat{g}(A,W)}. \end{aligned}$$

Note that these loss functions are not related to those in (4.5), (4.7) or (4.8). The generalized scores are equal to

$$\begin{aligned} \frac{d}{d\varepsilon_1}L_1\{\bar{Q}_1(\varepsilon_1),O\}|_{\varepsilon_1=0} &= -2\frac{\psi(A,Z)h(A,Z)}{\hat{g}(A,W)}\{Y-\bar{Q}_1(A,W)\} \\ \frac{d}{d\varepsilon_2}L_2\{\bar{Q}_2(\varepsilon_2),O\}|_{\varepsilon_2=0} &= \frac{h(A,Z)}{\hat{g}(A,W)}\{Y^2-\bar{Q}_2(A,W)\}, \end{aligned}$$

corresponding with the first two parts of the efficient influence curve presented in (4.9). The marginal distribution of W is estimated with the empirical distribution $Q_W(\mathbb{P})$ of $W_1,\ldots,W_n$. It can be shown that $Q_W(\mathbb{P})$ solves $L_{\bar{Q}}(\psi) - E_{Q_W}L_{\bar{Q}}(\psi)$ (the third part of the efficient influence curve equation) at any $\bar{Q}$.

For initial estimators $\hat{\bar{Q}}$ and $\hat{g}$, the first step TMLE of $\bar{Q}_0$ is given by $\hat{\bar{Q}}_j^* = \hat{\bar{Q}}_j(\hat{\varepsilon}_j)$, where

$$\hat{\varepsilon}_j = \arg\min_{\varepsilon} \mathbb{P}L_j\{\hat{\bar{Q}}_j(\varepsilon)\}. \tag{4.10}$$

The TMLE of $R_0(\psi)$ is now defined as the plug-in estimator $\hat{R}(\hat{\Psi}_k) \equiv R(\psi)(\hat{Q}^*)$, where $\hat{Q}^* = (\hat{\bar{Q}}_1^*, \hat{\bar{Q}}_2^*, Q_W(\mathbb{P}))$.

Under certain conditions explained in detail in Rose and van der Laan, 2011, Appendix A.18, if $\bar{Q}_0$ and g_0 are consistently estimated, this TMLE of $R_0(\psi)$ is asymptotically linear with influence curve $D(O|\bar{Q}_0, g_0, R_0(\psi))$, which means that it is consistent and efficient. If $\hat{g}$ is consistent but $\hat{\bar{Q}}^*$ is not, the TMLE is consistent but inefficient, and its variance can be conservatively estimated by

$$\hat{\sigma}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\{D^{DR}(O_i|\bar{Q}^*, \hat{g}, \hat{R}(\hat{\Psi}_k))\}^2.$$

If one uses data-adaptive estimators in $\hat{\bar{Q}}$ and $\hat{g}$, it is often appropriate to replace the estimate of the variance by a cross-validated estimator.

The conditions needed for asymptotic linearity of the TMLE (see Rose and van der Laan, 2011, Appendix 18) include a Donsker condition on the class of functions that contains the estimated efficient influence function D. Such Donsker conditions impose certain restrictions on the type of algorithms that can be used for estimation of $\bar{Q}_0$ and g_0, forcing the user to find a trade off between obtaining the best possible prediction algorithms and not using algorithms that are too data-adaptive, because data-adaptive algorithms might lead to estimators that do not belong to a Donsker class (e.g., random forest).

The cross-validated TMLE, whose theoretical properties are discussed in Zheng and van der Laan (2011a), provides a template for the joint use of cross-validation and TMLE methodology that avoids Donsker conditions and therefore allows the use of very data-adaptive techniques in order to find consistent estimators of $\bar{Q}_0$ and g_0. An additional advantage of CV-TMLE in this setting is that it allows us to have a valid estimator of the risk of an estimated CDRC, solving the issue of over-fitting through the use of cross-validation.

Estimators of the risk of a candidate estimator

The previous section provided an algorithm to estimate the risk of a candidate value for the causal dose-response curve, when the value is given and not estimated from the sample. That scenario is very rare in real data applications, and it is very common that the CDRC candidates have to be estimated from the sample as well. In such situations, if the algorithms $\hat{\Psi}_k$ are trained in the whole sample, the use of the estimators of the risk presented in the previous sections would lead to the selection of the candidates that overfit the data.

van der Laan and Dudoit (2003), van der Vaart (2003); van der Laan, Dudoit, and Keles (2004), among others, show that cross-validation is a powerful tool for estimating the risk of a candidate estimator of a non pathwise differentiable parameter, and show that such cross-validation based selection endows the selector with an oracle inequality that translates into asymptotic optimality.

Assume now that $\hat{\Psi}_k$ is a mapping that maps elements in a non parametric statistical model into a space of functions of a and Z ($\hat{\Psi}_k : \mathscr{M} \to \mathscr{H}$). An estimate of $\psi_0 = \Psi(P_0)$ is now seen as such map evaluated in the empirical distribution of $O_1, \ldots, O_n$, i.e., $\hat{\Psi}_k(\mathbb{P})$.

Consider the following cross-validation scheme. Let a random variable S taking values in $\{0,1\}^n$ index a random sample split into a validation sample $V_S = \{i \in \{1, \ldots, n\} : S_i = 1\}$ and a training sample $T_S = \{V_S\}^c$, where S has a uniform distribution over a given set $\{s_1, \ldots, s_m\}$ such that $\sum_j s_{i,j} > 1$ for all $i = 1, \ldots, m$. Here we note that the union of the validation samples equals the total sample: $\cup_S V_S = \{1, \ldots, n\}$, and the validations samples are disjoint: $V_{s_1} \cap V_{s_2} = \emptyset$ for $s_1 \neq s_2$. Denote $\mathbb{P}_{T_S}$ and $\mathbb{P}_{V_S}$ the empirical distributions of a training and validation sample, respectively. For a function $g\{T_S, V_S\}$, we denote $E_S g(T,V) = \frac{1}{m}\sum_{j=1}^{m} g\{T_{s_j}, V_{s_j}\}$.

Since $\hat{\Psi}_k(\mathbb{P})$ is now a value that depends on the sample, it does not make sense to talk about a parameter $R\{\hat{\Psi}_k(\mathbb{P}), Q_0\}$, because it does not agree with the formal definition of a parameter. Nonetheless, in an abuse of language we will talk about "estimation" of the "parameter" $E_S R\{\hat{\Psi}_k(\mathbb{P}_T), Q_0\}$, which we call the conditional (on the sample) risk of $\hat{\Psi}_k$.

Cross validated augmented IPTW

This estimator is also discussed by Wang, Bembom, and van der Laan (2006), and is given by the solution of the cross-validated version of the A-IPTW estimating equation, given by

$$E_S \mathbb{P}_V L_{\hat{Q}(\mathbb{P}_T), \hat{g}(\mathbb{P}_T)}\{\hat{\Psi}_k(\mathbb{P}_T)\}.$$

This estimator is asymptotically linear under the conditions presented in van der Laan and Dudoit (2003). An oracle inequality for the selector based on the A-IPTW risk estimator is also proved in the original paper.

Cross validated TMLE

The cross-validated targeted maximum likelihood estimator was introduced by Zheng and van der Laan (2010) as an alternative to the TMLE that avoids the Donsker conditions on the efficient influence curve (discussed in section 4.1). Donsker conditions on the class of functions generated by the estimated efficient function D represent an important limitation to the kind of algorithms that can be used in the initial estimators of $\bar{Q}_0$ and g_0: very data adaptive techniques will give as a result functions that do not belong to a Donsker class. As discussed in section 4.1, the consistency and efficiency of the risk estimator depend on the consistency of the initial estimator of $\bar{Q}_0$ and g_0. It is common practice in statistics to assume parametric models in order to estimate these quantities. Such parametric models are often chosen ad-hoc, based on arbitrary preferences of the researcher, and do not encode legitimate knowledge about the data generating process. Thus, we avoid such parametric assumptions, and prefer to use data-adaptive techniques to find the algorithm that best approximates $\bar{Q}_0$ and g_0.

As we will see in the next section, the use of cross-validation also equips the CV-TML selector with an oracle inequality, meaning that such selector performs asymptotically as well as a selector in which the risk is computed based on the true (unknown) probability distribution.

Zheng and van der Laan (2010) present two types of CV-TML estimators: one for a general parameter, and a specific CV-TMLE for the case in which Q can be partitioned as (Q_1, Q_2) and the mapping that defines the parameter is linear in Q_1. As discussed in section 4.1, the risk $R(\psi, P)$ of a given candidate depends on P only through $Q(P) = \{\bar{Q}(P), Q_W(P)\}$, and it can be easily verified that $R(\psi, Q)$ is linear in Q_W.

The construction of a CV-TML estimator requires the specification of the same three components discussed in section 4.1: a logistic loss function, a logistic parametric fluctuation, and an initial estimator of Q. For each S, let

$$\operatorname{logit} \bar{Q}_{j,k}(\mathbb{P}_{T_S})(\varepsilon_{j,k}) = \operatorname{logit} \bar{Q}_j(\mathbb{P}_{T_S}) + \varepsilon_{j,k} H_j\{\hat{\Psi}_k(\mathbb{P}_{T_S}), \hat{g}(\mathbb{P}_{T_S})\},$$

where $H_j;\ j = 1,2$ were defined in section 4.1. This is the same fluctuation considered before, but defined only based on the training sample. With this modification, the CV-TMLE is defined analogous to the regular TMLE. Let

$$\hat{\varepsilon}_{j,k} = \arg\min_{\varepsilon} E_S \mathbb{P}_V L_j\{\hat{\bar{Q}}_{j,k}(\mathbb{P}_T)(\varepsilon)\}; \quad j = 1,2, \tag{4.11}$$

and for each S define the updates

$$\hat{\bar{Q}}^*_{j,k}(\mathbb{P}_{T_S}) = \hat{\bar{Q}}_j(\mathbb{P}_{T_S})(\hat{\varepsilon}_{j,k});\ j = 1,2, \tag{4.12}$$

which results in the plug-in estimator of the oracle risk

$$\hat{R}(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), \hat{\bar{Q}}^*_k(\mathbb{P}_T), Q_W(\mathbb{P}_V)\} =$$
$$\frac{1}{m} \sum_{s \in \{s_1, \ldots, s_m\}} \frac{1}{n_s} \sum_{i \in V_s} \int_{\mathscr{A}} \Big\{ \hat{\bar{Q}}^*_{2,k}(\mathbb{P}_{T_s})(a, W_i) - 2\hat{\bar{Q}}^*_{1,k}(\mathbb{P}_{T_s})(a, W_i) \hat{\Psi}_k(\mathbb{P}_{T_s})(a, Z_i) +$$
$$[\hat{\Psi}_k(\mathbb{P}_{T_s})(a, Z_i)]^2 \Big\} h(a, Z_i) d\mu(a), \tag{4.13}$$

where $\hat{\bar{Q}}^*_k(\mathbb{P}_{T_S}) = \{\hat{\bar{Q}}^*_{1,k}(\mathbb{P}_{T_S}), \hat{\bar{Q}}^*_{2,k}(\mathbb{P}_{T_S})\}$, $Q_W(\mathbb{P}_{V_S})$ denotes the empirical distribution of W in the validation sample S, and n_s denotes the size of V_S.

For a definition of the CV-TMLE for general parameters the interested reader is referred to the original article. In the next sections we will present the asymptotic linearity of the previous estimator, as well as an oracle inequality for the selector based on it.

4.2 Asymptotic linearity of CV-TML estimator of the risk

In this section we present a theorem establishing asymptotic linearity of the CV-TML estimator of the risk. This theorem is analogue to the theorems presented in Zheng and van der Laan (2010), and its proof uses the same ideas presented in that paper.

An analogue version of this theorem for the CV-A-IPTW is presented in van der Laan and Dudoit (2003). The CV-TMLE is expected to perform better than the CV-A-IPTW in finite sample sizes, in which practical positivity violations are often present and lead to CV-A-IPTW estimators that are either very unstable or provide solutions out of the range of the parameter of interest.

Theorem 1 (Asymptotic linearity)**.** *Define*

$$\hat{R}_0(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), Q_0\} \text{ and } \hat{R}(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), \hat{\bar{Q}}_k^*(\mathbb{P}_T), Q_W(\mathbb{P}_V)\}$$

with $R\{\psi, Q\} = Q_W L_{\bar{Q}}(\psi)$. *For a function* $f(\mathbb{P}_{T_S})$ *of O, define the norm* $||f(\mathbb{P}_T)||_{0,S} = \sqrt{E_S P_0 f(\mathbb{P}_T)^2 h}$. *Assume:*

1. *There exist constants* $\delta_1 > 0$ *and* $\delta_2 > 0$ *such that* $P(\hat{g}(\mathbb{P})(A|W) > \delta_1) = 1$ *and* $g_0(a|w) > \delta_2 \; \forall \; a, w$.
2. $||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}^2 = o_P(1/\sqrt{n})$
3. $\hat{\bar{Q}}_1^*(\mathbb{P}_{T_S}), \hat{\bar{Q}}_2^*(\mathbb{P}_{T_S})$ *and* $\hat{\Psi}_k(\mathbb{P}_{T_S})$ *converge to some fixed* $\hat{\bar{Q}}_1^*(P_0), \hat{\bar{Q}}_2^*(P_0)$ *and* $\hat{\Psi}_k(P_0)$ *in the sense that*
$$\begin{aligned} ||\hat{g}(\mathbb{P}_T) - g_0||_{0,S} ||\hat{\bar{Q}}_2^*(\mathbb{P}_T) - \hat{\bar{Q}}_2^*(P_0)||_{0,S} &= o_P(1/\sqrt{n}) \\ ||\hat{g}(\mathbb{P}_T) - g_0||_{0,S} ||\hat{\bar{Q}}_1^*(\mathbb{P}_T) - \hat{\bar{Q}}_1^*(P_0)||_{0,S} &= o_P(1/\sqrt{n}) \\ ||\hat{g}(\mathbb{P}_T) - g_0||_{0,S} ||\hat{\Psi}_k(\mathbb{P}_T) - \hat{\Psi}_k(P_0)||_{0,S} &= o_P(1/\sqrt{n}) \end{aligned}$$
4. *For some mean zero function* $IC_g(P_0) \in L_0^2(P_0)$, *we have*
$$P_0 \frac{g_0 - \hat{g}(\mathbb{P})}{g_0^2} h \left[\{\bar{Q}_{2,0} - \hat{\bar{Q}}_2^*(P_0)\} - 2h\psi_0 \{\bar{Q}_{1,0} - \hat{\bar{Q}}_1^*(P_0)\} \right] = \\ (\mathbb{P} - P_0) IC_g(P_0) + o_P(1/\sqrt{n}),$$

Then we have that

$$\hat{R}(\hat{\Psi}_k) - \hat{R}_0(\hat{\Psi}_k) = (\mathbb{P} - P_0) \left[D\{\hat{\bar{Q}}_{\bar{k}}(P_0), Q_{W,0}, g_0, \psi_0\} + IC_g(P_0) \right] + o_P(1/\sqrt{n}),$$

for $D\{\bar{Q}, Q_W, g, \psi\} = L_{\bar{Q},g}(\psi) - Q_W L_{\bar{Q}}$ *the efficient influence function of* $R\{\psi, Q(P)\}$.

The proof of this theorem is presented in appendix B.1. Next we will discuss the plausibility and implications of the assumptions of theorem 1.

Discussion on the assumptions of theorem 1

1. This assumption is a natural assumption, equivalent to the positivity assumption for binary treatments, and needed to identify and also needed to estimate the risk using IPTW or A-IPTW estimators.

2. This is a very important assumption stating that $\hat{g}$ is a consistent estimator of g_0. It is required that the rate of convergence is $n^{-1/4}$ or faster. This condition is automatically true in randomized control trials (RCT), in which the treatment mechanism is known. It is also true if g is known to belong to a parametric model, and in semi parametric models that assume enough smoothness of g_0. If g_0 is completely unknown, it is important to use aggressive data adaptive estimation techniques such as the super learner (van der Laan, Polley, and Hubbard, 2007) to find an estimator $\hat{g}$ that is more likely to satisfy this assumption.

3. This assumption states that the updated estimator $\hat{\bar{Q}}_1^*$ converges to some unspecified limit at a certain rate. It is worth to note that such limit is not assumed to be $\bar{Q}_{1,0}$, the only requirement is convergence to some value at a certain rate that depends on the rate of convergence of $\hat{g}$ to g_0. The desired rate of convergence can be achieved if, for example, $\hat{g}$ is $\sqrt{n}$-consistent (i.e., $\sqrt{n}||\hat{g}(\mathbb{P}_T) - g_0||_{0,S} = O_P(1)$) and $\hat{\bar{Q}}_1^*(\mathbb{P}_{T_S})$ converges to $\hat{\bar{Q}}_1^*(P_0)$ at any rate (i.e., $||\hat{\bar{Q}}_1^*(\mathbb{P}_T) - \hat{\bar{Q}}_1^*(P_0)||_{0,S} = o_P(1)$). The same is true for $\hat{\bar{Q}}_2^*$ and $\hat{\Psi}_k$.

4. In an RCT, in which g_0 is known, one could set $\hat{g}(\mathbb{P}) = g_0$ and this condition would be trivially satisfied. On the other hand, since cross-validation allows for the use of very aggressive techniques for estimation of $\bar{Q}_0$, we could have that $\hat{\bar{Q}}^*(P_0) = \bar{Q}_0$, and the condition would also be satisfied.

   In other cases, this assumption seems to be conflicting with assumption 2. If the treatment mechanism is completely unknown, it is necessary to use very aggressive data adaptive techniques to find estimators that satisfy assumption 2. The use of such estimators will usually lead to estimates of g_0 that do not provide the asymptotic linearity needed in 4. Likewise, the use of an inconsistent estimator that satisfies this condition (e.g., a parametric model) will violate assumption 2. In that case, it is necessary to rely on the consistency of $\hat{\bar{Q}}^*(\mathbb{P})$ in the sense that $\hat{\bar{Q}}^*(P_0) = \bar{Q}_0$, in which case assumption 4 will be trivially satisfied. This condition seems to suggest that the initial estimator $\hat{g}$ must also be fluctuated to target a smooth functional of g_0. This is a direction of future research, beyond the scope of this article.

   As opposed to the regular TMLE or A-IPTW, in which the Donsker conditions on D limit the use of very aggressive techniques for estimation of $\bar{Q}_0$, the use of cross-validation allows us to implement any type of algorithm, which in turn makes consistency of $\hat{\bar{Q}}^*(\mathbb{P})$ a very sensible assumption. We encourage the use of super learning for estimation of both $\bar{Q}_0$ and g_0. Super learner is a methodology that uses cross-validated risks to find an optimal estimator among a library defined by the convex hull of a user-supplied list of candidate estimators. One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution). Proofs and simulations regarding these and other asymptotic properties of the super learner can be found in van der Laan, Dudoit, and Keles (2004) and van der Laan and Dudoit (2003).

4.3 Asymptotic optimality of the CDRC estimate selector based on CV-TMLE risk

If the objective is to choose the best candidate among a list of candidate estimators $\hat{\Psi}_k : k = 1,\ldots,K_n$, it suffices to construct a ranking based on the pseudo-risk

$$R^{\dagger}(\psi)(\bar{Q}_1, Q_W) = E_{Q_W}\int_{\mathscr{A}} \psi(a,Z)\{\psi(a,Z) - 2\bar{Q}_1(a,W)\}h(a,Z)d\mu(a).$$

which has the advantage that $\bar{Q}_{2,0}$ does not need to be estimated, providing additional robustness of the candidate selector. In an abuse of notation $R^{\dagger}$ and $\bar{Q}_1$ will also be denoted by R and $\bar{Q}$ whenever the difference is clear from the context. Estimation of this pseudo-risk can be carried out in a similar fashion to estimation of the full risk presented in the previous section, with efficient influence function given by

$$\begin{aligned} D^{\dagger}(Q,g,\psi)(O) = -2\frac{h(A,Z)\psi(A,Z)}{g(A,W)}\{Y - \bar{Q}_1(A,W)\}+ \\ \int_{\mathscr{A}} \psi(a,Z)\{\psi(a,Z) - 2\bar{Q}_1(a,W)\}h(a,Z)d\mu(a) - R(\psi)(\bar{Q}_1, Q_W), \end{aligned} \quad (4.14)$$

which results in a CV-TMLE defined as

$$\hat{R}(\hat{\Psi}_k) \equiv E_S R\{\hat{\Psi}_k(\mathbb{P}_T), \hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V)\}$$

with $\hat{\bar{Q}}^*(\mathbb{P}_{T_S})$ exactly as in (4.12). We will discuss now asymptotic optimality of the selector based on the CV-TMLE. Assume that we have a list of candidate estimators for the CDRC given by $\hat{\Psi}_k$; $k = 1,\ldots K_n$. Each of these algorithms is viewed as a map $\hat{\Psi}_k : \mathscr{M} \to \mathscr{F}$, where $\mathscr{F}$ is the space of functions of a and Z. Define the CV-TMLE selector as

$$\hat{k} = \arg \min_{k=1,\ldots,K_n} \hat{R}(\hat{\Psi}_k),$$

and the oracle selector as

$$\tilde{k} = \arg \min_{k=1,\ldots,K_n} \hat{R}_0(\hat{\Psi}_k),$$

with $\hat{R}_0(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), Q_0\}$. The following theorem proves that these two selectors are asymptotically equivalent under certain consistency conditions of the initial estimator of g_0.

Theorem 2 (Oracle inequality)**.** *For each k, define*

$$\hat{\varepsilon}_k = \arg \min_{\varepsilon \in B \subset \mathbb{R}} E_S \mathbb{P}_V L\{\hat{\bar{Q}}_k(\mathbb{P}_T)(\varepsilon)\}$$

where $|B| = n^c$ *for finite c,*

$$-L(\bar{Q})(O) = Y\log\{\bar{Q}(A,W)\} + (1-Y)\log\{1 - \bar{Q}(A,W)\},$$

and

$$\operatorname{logit}\bar{Q}_k(\mathbb{P}_{T_S})(\varepsilon) = \operatorname{logit}\bar{Q}(\mathbb{P}_{T_S}) - 2\varepsilon\frac{h\hat{\Psi}_k(\mathbb{P}_{T_S})}{\hat{g}(\mathbb{P}_{T_S})}.$$

Let $\hat{\bar{Q}}^*_k(\mathbb{P}_{T_S}) = \hat{\bar{Q}}(\mathbb{P}_{T_S})(\hat{\varepsilon}_k)$ *be the CV-TMLE targeted towards estimation of the true conditional risk*

$$\hat{R}_0(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), Q_0\}.$$

Assume that $h/\hat{g}$*,* h/g_0 $\hat{\Psi}_k$*,* ψ_0*,* $\bar{Q}_0$*, and* $\hat{\bar{Q}}$ *have supremum norm smaller than a constant* $C < \infty$ *with probability 1. Let* M_n *be the total number of possible points for* (k, ε_k) *across* $k = 1, \ldots, K_n$*, so that* $M_n \leq n^c K_n$*. Define* $\hat{R}(\hat{\Psi}_k, \psi_0) \equiv \hat{R}(\hat{\Psi}_k) - \hat{R}(\psi_0)$ *and* $\hat{R}_0(\hat{\Psi}_k, \psi_0) \equiv \hat{R}_0(\hat{\Psi}_k) - R_0(\psi_0)$*, where*

$$\hat{R}(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), \hat{\bar{Q}}^*_k(\mathbb{P}_T), Q_W(\mathbb{P}_V)\}$$

is the TMLE of $\hat{R}_0(\hat{\Psi}_k)$*. The expression* $a_n \lesssim b_n$ *means that* $a_n \leq c b_n$ *for a constant c. For a function* $f(\mathbb{P}_{T_S})$ *of O, define the norm* $||f(\mathbb{P}_T)||_{0,S} = \sqrt{E_S P_0 f(\mathbb{P}_T)^2 h}$*. We have for each* $\delta > 0$*, there exists a* $c(M, \delta) < \infty$ *so that*

$$\begin{aligned}\sqrt{E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0)} - \sqrt{(1+2\delta)E\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0)} \lesssim\ & \sqrt{c(M,\delta)\frac{1+\log M_n}{n}} \\ & + \sqrt{(1+\delta)E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}\frac{1+\log M_n}{\sqrt{n}}} \\ & + (1+\delta)E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}E||\hat{\bar{Q}}^*_{\bar{k}}(\mathbb{P}_T) - \bar{Q}_0||_{0,S} \\ & + (1+\delta)E||(\hat{g}(\mathbb{P}_T) - g_0)(\hat{\bar{Q}}^*_0(\mathbb{P}_T) - \bar{Q}_0)||_{0,S},\end{aligned}$$

where $\hat{\bar{Q}}^*_0$ *is the CV-TMLE of* $\bar{Q}_0$ *obtained when the target parameter is* $R(\psi_0)$*, and* $\bar{k}$ *is either* $\hat{k}$ *or* $\tilde{k}$*, whichever gives the worst bound.*

A proof of this theorem is provided in appendix B.2. The use of a grid of size n^c for constant c when estimating ε_k does not represent a limitation of the result of the theorem, since the result without the grid will be similar up to a term $O_P(1/\sqrt{n})$ that does not affect the asymptotic behavior of the CV-TMLE selector. However, a grid of size n^c allows the proof presented in appendix B.2.

The following corollary provides the conditions under which the CV-TMLE selector is asymptotically equivalent to the oracle selector.

Corollary 1 (Asymptotic optimality)**.** *In addition to the conditions of theorem 2, assume that*

$$\frac{1+\log M_n}{n}\frac{1}{E\hat{R}_0(\hat{\Psi}_{\tilde{k}},\psi_0)} \to 0 \quad as \quad n\to\infty$$
$$\frac{1+\log M_n}{\sqrt{n}}\frac{E||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}}{E\hat{R}_0(\hat{\Psi}_{\tilde{k}},\psi_0)} \to 0 \quad as \quad n\to\infty$$
$$\frac{E^2||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}E^2||\hat{\bar{Q}}^*_{\tilde{k}}(\mathbb{P}_T)-\bar{Q}_0||_{0,S}}{E\hat{R}_0(\hat{\Psi}_{\tilde{k}},\psi_0)} \to 0 \quad as \quad n\to\infty$$
$$\frac{E^2||(\hat{g}(\mathbb{P}_T)-g_0)(\hat{\bar{Q}}^*_0(\mathbb{P}_T)-\bar{Q}_0)||_{0,S}}{E\hat{R}_0(\hat{\Psi}_{\tilde{k}},\psi_0)} \to 0 \quad as \quad n\to\infty.$$

then

$$\frac{E\hat{R}_0(\hat{\Psi}_{\hat{k}},\psi_0)}{E\hat{R}_0(\Psi_{\tilde{k}},\psi_0)} \to 1 \quad as \quad n\to\infty.$$

Since

$$E\hat{R}_0(\hat{\Psi}_{\tilde{k}},\psi_0) = E\int\int(\hat{\Psi}_{\tilde{k}}(\mathbb{P}_T)-\psi_0)^2 d\mu\, dQ_{W,0} = E||(\hat{\Psi}_{\tilde{k}}(\mathbb{P}_T)-\psi_0)/\sqrt{g_0}||^2_{0,S},$$

the convergences assumed in corollary 1 are expected to hold, for example, if $\hat{g}$ converges to g_0 at a rate faster than $\hat{\Psi}$ converges to ψ_0.

In the following section we will show the results of a simulation study in which the finite sample size properties of the CV-TMLE based selector of their risk are explored for a specific data generation process.

4.4 Simulation

In order to explore some of the finite sample size properties of the risk estimators and the selectors based on them, we performed a Monte Carlo simulation. We generated 500 samples of sizes 100, 500, and 1000 from the following data generating process:

$$\begin{aligned}
W_1 &\sim U\{0,1\}\\
W_2 &\sim Ber\{0.7\}\\
W_3 &\sim N\{W_1, 0.25\times\exp(2W_1)\}\\
A &\sim Beta\{\nu(W)\mu(W), \nu(W)[1-\mu(W)]\}\\
Y &\sim Ber\{Q_0(A,W)\},
\end{aligned}$$

where

$$\begin{aligned}
\nu(W) &= \exp\{1+2W_1\,\text{expit}(W_3)\}\\
\mu(W) &= \text{expit}\{.03-.8\log(1+W_2)+.9\exp(W_1)W_2-.4\arctan(W_3+2)W_2W_1\}\\
\bar{Q}_0(A,W) &= \text{expit}\{-2+1.5A+5A^3-2.5W_1+.5AW_2-\log(A)W_1W_2+.5A^{3/4}W_1W_3\}.
\end{aligned}$$

Under this parameterization $E(A|W) = \mu(W)$. We considered four candidates algorithms given by marginal structural models (MSM) of the form logit $\Psi_p(a) = m_p(a,\beta)$, where m_p is a polynomial of degree $p = 1,\ldots,4$ on a with coefficients $\beta_j : \ j = 0,\ldots,p$. The coefficients β_j were estimated with IPTW estimators as presented by Robins, Hernan, and Brumback (2000) and Neugebauer and van der Laan (2007). The true value of $\psi_0(a) = E\{\bar{Q}_0(a,W)\}$ was computed from this data generating distribution by drawing a sample of size 100.000 and, for each a, computing the empirical mean of $\bar{Q}_0(a,W)$. All the simulations were performed assuming the true parametric model for the outcome and treatment mechanism were known. Figure 4.1 presents the true dose-response curve, as well as the expectation of the candidate estimators across the 500 samples. From this graph, we can see that among the candidates chosen, a polynomial of degree 2 seems to provide the closest approximation to the true dose-response curve without over-fitting the data.

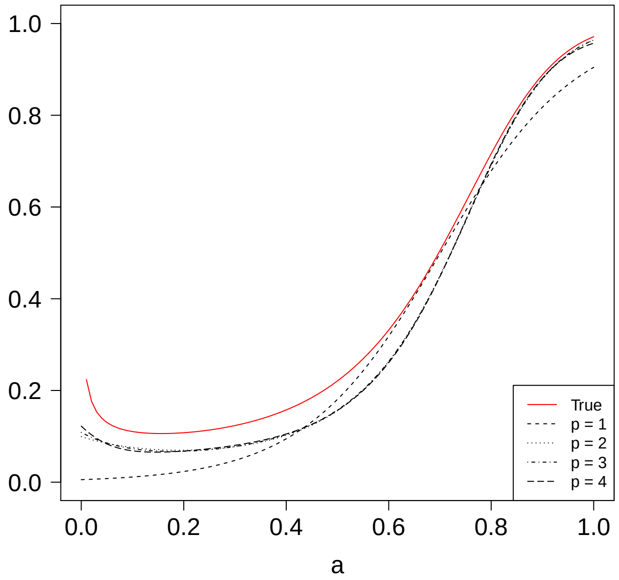


Figure 4.1: True $\psi_0(a)$ and expectations of the four candidate estimators of degree p

Table 4.1 shows the expectation of the random variable $\hat{R}(\Psi_p) - \hat{R}_0(\Psi_p)$, which from theorem 1 should approach zero as the sample size increases. As we can see, that is not the case for the CV-A-IPTW estimator with sample size 100 due to the presence of empirical violations of the positivity assumption that cause very small treatment weights and therefore very unstable, non-regular estimates. However, that problem seems to be fixed asymptotically, since for large sample sizes empirical violations of the positivity assumption are less likely to occur.

Table 4.2 shows the proportion of estimates that fell outside the interval $(-10,10)$ or fell out of the parameter space. The interval $(-10,10)$ was chosen arbitrarily, and represents inadmissi-

Risk	Candidate	*n*		
Estimator	Degree	100	500	1000
CV-A-IPTW	1	-1.1436	0.0067	0.0048
	2	-1.1716	-0.0061	0.0042
	3	-1.0788	-0.0054	0.0043
	4	-1.0194	-0.0053	0.0032
CV-TMLE	1	0.0063	0.0054	0.0046
	2	0.0085	0.0054	0.0041
	3	0.0091	0.0059	0.0043
	4	0.0094	0.0058	0.0042

Table 4.1: Expectation of $\hat{R}(\Psi_p) - \hat{R}_0(\Psi_p)$ across 500 simulated samples.

		Outliers			Out of bounds		
Risk	Candidate	*n*					
Estimator	Degree	100	500	1000	100	500	1000
CV-A-IPTW	1	0.0098	0.0000	0.0000	0.0547	0.0020	0.0020
	2	0.0078	0.0000	0.0000	0.0527	0.0020	0.0020
	3	0.0098	0.0000	0.0000	0.0488	0.0020	0.0020
	4	0.0098	0.0000	0.0000	0.0508	0.0020	0.0020

Table 4.2: Proportion of estimates outliers (< -10 or > 10) and proportion of estimates out of bounds (< 0 or > 1).

ble bounds for an estimator of a parameter that ranges in the interval $(0,1)$. Since the TMLE is a substitution estimator, it all the estimates fell within the parameter space, and are thus not presented. Due to practical violations of the positivity assumption previously mentioned, an important proportion (around 5%) of the A-IPTW estimates fell outside the parameter space for sample size 100.

Table 4.3 contains the expected values of $\hat{R}(\Psi_p) - \hat{R}_0(\Psi_p)$ across 500 simulated samples once the estimates that fell outside the interval $(0,1)$ were removed. In this case, the expectation of the A-IPTW based estimator of the risk is much closer to what is expected theoretically and had already been achieved by the TML estimator.

Finally, table 4.4 shows the proportion of times that a given candidate is chosen according to the A-IPTW, TMLE, and the oracle selector. As we can see, both the A-IPTW and the TMLE based selectors perform similar to the oracle selector, particularly as the sample size increases, thus showing no apparent advantage (at least for this particular data generating mechanism) of either method when evaluated as a candidate selector procedure.

Risk Estimator	Candidate Degree	n 100	500	1000
CV-A-IPTW	1	0.0034	0.0064	0.0048
	2	-0.0076	-0.0062	0.0042
	3	-0.0083	-0.0054	0.0043
	4	-0.0083	-0.0055	0.0032
CV-TMLE	1	0.0063	0.0054	0.0046
	2	0.0085	0.0054	0.0041
	3	0.0091	0.0059	0.0043
	4	0.0094	0.0058	0.0042

Table 4.3: Expectation of $\hat{R}(\Psi_p) - \hat{R}_0(\Psi_p)$ across 500 simulated samples after removing estimates out of bounds.

p	n 100			500			1000		
	TMLE	A-IPTW	Oracle	TMLE	A-IPTW	Oracle	TMLE	A-IPTW	Oracle
1	0.37	0.24	0.39	0.11	0.08	0.05	0.03	0.05	0.00
2	0.44	0.48	0.54	0.63	0.60	0.74	0.59	0.59	0.74
3	0.11	0.17	0.06	0.14	0.20	0.17	0.24	0.21	0.20
4	0.07	0.11	0.01	0.12	0.12	0.04	0.14	0.15	0.06

Table 4.4: Proportion of times that a given candidate is chosen according to each risk estimator.

Chapter 5

Application to prognosis and variable importance in severe trauma patients

A primary goal in evidence-based medicine is to design prognosis tools that take into account a possibly large set of measured characteristics in order to predict a patient's most likely medical outcome. An equally important goal is to establish which of those measured characteristics is decisive in the development of the predicted outcome. In the statistics literature these two goals have been called prediction and variable importance analysis, respectively. In addition to understanding the underlying biological mechanisms related to positive medical outcomes, the joint use of these tools can help doctors devise the optimal treatment plan according to the specific characteristics of the subject, simultaneously taking into account hundreds of variables collected for each patient. Despite the current ability to measure a patient's clinical history in detail, medical practice still involves care decisions based on physician's experience and rules of thumb that use only a few variables and therefore fail to take into consideration the possible intricate relations between all the measured underlying factors that determine a patient's health status. In the last years researchers in the fields of biostatistics and bioinformatics have become increasingly more interested in developing mathematical and computational tools that help make optimal care decisions based on all the collected information about a patient's health status and history. Because of the large number of variables and the complexity of the relations between them, prediction and variable importance would be impossible to achieve without the use of complex statistical algorithms accompanied by powerful computers able to carry out a large number of computations in large data sets within reasonable time frames that help doctors make the right treatment decisions in a timely fashion.

From a technical and practical point of view prediction and variable importance are different goals whose optimal achievement requires the use of different tools. The objective in prediction is to specify a well defined algorithm that is capable of doing accurate predictions, where accuracy can be defined in a variety ways. For prediction it is only relevant whether the prediction algorithm is accurate or not, it is unnecessary and sometimes inappropriate (e.g., with non-probabilistic predictors) to use the intermediate calculations of the prediction algorithm to find statistical or causal relations between the variables involved. On the other hand, variable importance (VIM) methods are aimed to measure the degree to which changes in the prediction are caused by changes in each of the predictor variables. VIM methods often provide a ranking of the most likely causes of a the predicted outcome, and are intended to supply doctors with tools for making optimal treatment decisions. This difference between prediction and VIM has two main consequences. First, VIM problems are of a *causal* nature, whereas prediction problems are merely *associational*. Second, in order to help the decision making process, VIM parameters must be as informative as possible, having an interpretation in terms of the expected change in the outcome under a pool of possible interventions. As explained below, a meaningful interpretation can only be obtained through an intelligible characterization of VIM as a statistical (or causal) parameter defined as a mapping from a honest, tenable statistical model into an Euclidean space.

Current practice in biostatistics and bioinformatics involves the use of machine learning algorithms for prediction and the posterior computation of VIM quantities based on its output and intermediate calculations (see e.g., Breiman, 2001; Olden and Jackson, 2002; Olden, Joy, and Death, 2004; Strobl et al., 2007, for discussions on random forests and neural networks variable importance). Because these measures are defined in terms of an algorithm that was targeted to perform well at prediction, they result in variable importance measures that can seldom be considered estimates of a well defined causal or statistical parameter. As an example, consider the case of regression and classification trees (e.g., random forests), where the VIM for a variable X is defined as the difference between prediction error when X is perturbed versus the prediction error otherwise (Breiman, 2001; Ishwaran, 2007). The relevance of this quantity as a measure of VIM is unclear because: 1) it does not represent a statistical or causal parameter, 2) it does not have an interpretation in terms of the mechanistic process that generates the data, and 3) its interpretation may be difficult to communicate to the public, even the public trained in statistics. As an example of the technical difficulties arising from this practice, Strobl et al. (2007) discuss the “bias” of random forest VIM measures, missing the fact that bias can only be defined in terms of a target statistical parameter, which is never specified in random forest VIM analysis. Additionally, no formal inference (p-values) methods exist for regression and classification trees based VIM.

Furthermore, an algorithm designed to perform well at prediction is not guaranteed to also do a good job at estimating VIM measures, because good performance is defined differently for each goal. Performance in prediction is typically assessed through quantities like the area under the ROC, the false positive rate, or the expected risk of a sensible loss function. Performance in estimation of Euclidean parameters is assessed in terms of statistical properties like consistency and efficiency (related to bias and variance). Prediction algorithms are designed to perform well at estimating the entire regression model, resulting in an incorrect bias-variance trade-off for each VIM measure.

However, defining VIM parameters in terms of causal relations for continuous variables poses additional technical challenges. When researchers using causal inference methods are faced with exposures of continuous nature, the most common approach is to dichotomize the continuous exposure and consider the effect of its binary version on the outcome. This approach suffers from various flaws. First, the causal parameter does not answer questions about plausible modifications to the data generating mechanism. Stitelman, Hubbard, and Jewell (2010a) show that the additive causal effect of a dichotomized exposure compares an intervention in which the density of the exposure is truncated below the dichotomization threshold with an intervention in which the density is truncated above it. Such interventions are seldom realistic, and might not be of great interest for specific applications. Second, even if truncation interventions are realistic, the data analyst still has to choose a cutoff point for the dichotomization. Most of the times the decision about such cutoff point is data-driven (i.e., comparing quantiles), or made completely arbitrarily. This practice renders a parameter that is dependent on the data, making its interpretation in terms of the original, continuous exposure even more difficult. For instance, if VIM measures for continuous outcomes are defined in terms of a dichotomization, it is often possible to define the right cutoff point that makes the continuous variable more important than a given binary variable of reference. It is thus necessary to argue why the chosen cutoff point makes these VIM measures comparable.

In this paper we explore a VIM problem in which it is necessary to rank a list of both continuous and binary variables in terms of their importance for developing a medical outcome, which is a very common problem in variable importance analysis. We use state of the art methods for causal inference to solve prediction and VIM problems and illustrate the use of our methods using a medical application, but the methods we develop and the arguments we present are completely general and can be applied to any prediction or VIM problem (e.g., the analysis of ecological data, genomics, educational and social research, economics). For prediction, we use a machine learning technique called super learning which uses cross validation to choose an optimal convex combination of a list of prediction algorithms provided by the user. The properties of this method have been extensively studied through analytical calculations as well as simulations by van der Laan and Dudoit (2003) and van der Laan, Dudoit, and Keles (2004); van der Laan, Polley, and Hubbard (2007), among others. We define VIM measures in terms of appropriate interventions in a causal model, which results in parameters that have a clear interpretation in terms of the expected outcome under a clinical intervention. VIM measures with causal interpretation are more relevant than their machine learning/modelling counterpart because they attempt to discover the factors that must be intervened upon in order to obtain a significant improvement in the outcome, and not just the factors that are associated to the outcome in question. We define VIM measures that respect the continuous or binary nature of the variable, and are comparable in the sense that their mathematical definition is equal up to first order, providing a valid ranking of the variables in terms of their causal importance. In order to find VIM estimators with the best possible statistical properties we use the tools for efficient inference in semi-parametric models described by Bickel et al. (1997); van der Laan and Robins (2003), and Rose and van der Laan (2011) among others, which allow us to use asymptotically linear estimators of the VIM parameters that are consistent and efficient in the non-parametric model (under regularity conditions).

We demonstrate the use of these techniques in an example predicting clinical outcomes and evaluating the VIM of a set of competing variables in severe trauma patients. Trauma is the leading cause of death between the ages of 1 and 44, according to the World Health Organization. The vast majority of these deaths take place quickly and much of the initial resuscitative and decision-making action takes place in the first minutes to hours after injury (Hess, Holcomb, and Hoyt, 2006; Holcomb et al., 2007). In addition, it is clear that as patients live through their initial resuscitation, operative conduct and early ICU care are the principal drivers of their current physiologic state and future outcome are dynamic. Different variables are important in the first 30 minutes after injury than at 24 hours after a patient has survived long enough to receive large volume resuscitation, operative intervention and ICU care. At any time, however, practitioners are often left making care decisions without knowledge of the current patient physiologic state and which parameters are important at that moment. Left with this uncertainty and awash in constantly evolving multivariate data, practitioners make decisions based on clinical gestalt, a few favorite variables, and rules of thumb developed from clinical experience. To aid in prediction, the medical literature is filled with scoring systems and published associations between these variables (physiology, biomarker, demographic, etc.) and outcomes of interest (Krumrei et al., 2012; Lesko et al., 2012; MacFadden et al., 2012; Nuñez et al., 2009; Schöchl et al., 2011). While numerous, these published statistical associations, given the reported methodology, often report misspecified and overfit models. In addition most of these statistical predictive models do not account for the rapidly changing dynamics of a severely injured patient, and fail to take into account the statistical issues discussed in the previous paragraphs. An ideal system would mimic the clinical decision making of an experienced practitioner by providing dynamic prediction (changing prediction at iterative time points) while evaluating the dynamic importance of each variable over time (Buchman, 2010). This then would mimic the implicit understanding a clinician brings to a patient where it is clear that the necessary focus of care must change over time.

The chapter is organized as follows. In section 5.1 we describe the structure of the data and introduce the statistical problem using causal inference tools to define statistical parameters that measure the importance of a variable with respect to an outcome of interest. In section 5.2 we present various estimators for the variable importance parameters previously defined, and briefly describe the super learner (van der Laan, Polley, and Hubbard, 2007), an ensemble learner whose asymptotic performace is optimal for prediction. In section 5.3 we describe the problem of prognosis for trauma patients and the dynamic importance of clinical factors, demonstrate the use of the methods previously presented, and compare the results with an approach that utilizes stepwise regression to estimate VIM measures and provides a comparison with common statistical practice. Finally, in section 5.4 we provide some concluding remarks.

5.1 Data, problem formulation, and parameters of interest

In order to estimate the effect of a variable A on an outcome Y controlling for a set of variables W, it is common practice among data analysts to estimate the parameter β in a parametric regression model $E(Y|A,W) = m(A,W|\beta)$ for a known function m, for example,

$$E(Y|A,W) = \beta_0 + \beta_1 A + \beta_2 W. \tag{5.1}$$

It is also common to assume more complex models for the relation between (A,W) and Y (e.g., by varying the amount of interaction terms, functional form of m, or by using smoothing techniques), but the linear regression example suffices to introduce the problem. Under model (5.1), the estimate of β_1 is interpreted as the expected change in Y given a change of one unit in A:

$$\beta_1 = E\{E(Y|A+1,W) - E(Y|A,W)\}, \tag{5.2}$$

where we note that the interest of the researcher is to estimate the right hand side of this equation, since under small violations to the assumptions of model (5.1) the estimate of β_1 cannot be interpreted as in (5.2) anymore. Consider for example the following models:

$$E(Y|A,W) = \beta_0^{(1)} + \beta_1^{(1)}A + \beta_2^{(1)}W + \beta_3^{(1)}AW \tag{5.3}$$

$$E(Y|A,W) = \beta_0^{(2)} + \beta_1^{(2)}\log(A) + \beta_2^{(2)}W. \tag{5.4}$$

If the true conditional expectation is given by model (5.3), but (5.1) is estimated instead, neither the estimate of β_1 in model (5.1) nor $\beta_1^{(1)}$ in (5.3) represent the quantity in the right hand side of (5.2), which is now given by $\beta_1^{(1)} + \beta_3^{(1)}E(W)$. On the other hand, if the true model is (5.4), the parameter of interest is now given by $\beta_1^{(2)}\{E(\log(A+1)) - E(\log A)\}$.

In order to avoid these flaws, in this paper we will define parameters in terms of characteristics of the probability distribution of the data under a non-parametric model, as in equation (5.2). This practice allows the definition of the parameter of interest independently of (possibly) misspecified parametric models, and avoids dealing with different interpretations of regression parameters under incorrect model specifications.

The causal interpretation of statistical parameters (e.g., 5.2) requires additional untestable assumptions about the distribution of counterfactual outcomes under a hypothetical interventions that are often encoded in a structural equation model (Pearl, 2000).

In the remaining of the section we will describe the observed data, and use a nonparametric structral equation model (NPSEM) in order to define the VIM measures in terms of modifications to the assumed causal model. We will now introduce the example that motivated the development of these tools, and that will be analyzed in section 5.3.

Example The data analyzed in this example were collected as part of the Activation of Coagulation and Inflammation in Trauma (ACIT) study, which is a prospective cohort study of severe trauma patients admitted to a single level 1 trauma center. Several physiological and clinical measurements were recorded at several time points for each patient after arrival to the emergency room. These variables include demographic variables (e.g., age, gender, etc.), baseline risk factors (e.g., asthma, chronic lung disease, Glasgow coma scale, diabetes, injury mechanism, injury severity score, etc.), longitudinally measured variables that account for the patient's treatment and health status history (e.g., respiratory and heart rate, platelets, coagulation measures like prothrombin time and INR, activated protein C, etc.), and an indicator of the occurrence of death at each time interval. Because these data are often collected in a high-stress environment, it is common that some variables are missing for some patients at a given time point. The list of variables we analyzed presented in table 5.1.

Assume that observations on each patient are recorded at times $t_0, t_1, \ldots, t_J$, where $t_0 = 0$, and let T denote the time of death of a patient. The observed data for each patient is given by the random variable

$$O = (L_0, C_1, L_1, Y_1, \ldots, C_J, L_J, Y_J),$$

Variable	Type	Description
Age	Baseline	Age in years
GCS	Baseline/Treatment	Arrival Glascow Comma Score
ISS	Baseline/Treatment	Injury Severity Score
Asthma	Baseline	Indicator of previous Asthma
COPD	Baseline	Indicator of previous Chronic Obstructive Pulmonary Disease
OCLG	Baseline	Indicator of Other Chronic Lung Disease
CAD	Baseline	Coronary Artery Diseae
CHF	Baseline	Congestive Heart Failure
ESRD	Baseline	End Stage Renal Disease
CIRR	Baseline	Cirrhosis
DIAB	Baseline	Diabeted
HPAN	Baseline	Hypoalbuminemia
Gender	Baseline	Gender
MECH	Baseline	Injury mechanism: blunt or penetrating
HR	Treatment	Heart Rate
RR	Treatment	Respiratory Rate
SBP	Treatment	Spontaneous Bacterial Peritonitis
BDE	Treatment	Base Deficit/Excess
BUN	Treatment	Blood Urea Nitrogen
CREA	Treatment	Creatinine
HGB	Treatment	Hemoglobin
HCT	Treatment	Hematocrit
PLTS	Treatment	Platelets
PT	Treatment	Prothrombin Time
PTT	Treatment	Partial Prothrombin Time
INR	Treatment	International Normalized Ratio
FV	Treatment	Factor III
FVIII	Treatment	factor VIII
ATIII	Treatment	Antithrombin III
PC	Treatment	Protein C
DDIM	Treatment	D-Dimer
TPA	Treatment	Tissue Plasminogen Activator
PAI	Treatment	Plasminogen Activator Inhibitor
SEPCR	Treatment	Soluble Endothelial Protein C Receptor
STM	Treatment	Soluble Thrombomodulin
APC	Treatment	Activated Protein C

Table 5.1: Variables in the ACIT data set

where L_0 denotes a set of baseline variables recorded at admision to the hospital, $L_j = (L_{j1}, \ldots, L_{jK})$ denotes a set of variables measured at time t_j, $C_j = (C_{j1}, \ldots, C_{jK})$ where C_{jk} denotes an indicator of missingness of L_{jk}, and $Y_j = I(t_j < T \leq t_{j+1})$ denotes an indicator of death occuring in the interval $(t_j, t_{j+1}]$, for $j = 0, \ldots, J-1$. Once death occurs the random variables in the remaining time points of the vector O become degenerate so that this structure is well defined.

In order to introduce the VIM measures we will temporarily assume that the observed data were generated by the following non-parametric structural equation model (NPSEM Pearl, 2000).

If the assumptions of NPSEM are correct, the VIM parameter can be causally interpreted as the expected change in the outcome caused by a care intervention on the variable of interest. If the assumptions of the model are false, the VIM measure provides a measure of the importance of each variable for predicting the outcome if no other variable was collected at the time point of interest, as we will see below. The NPSEM is given by

$$\begin{aligned}
L_0 =& f_{L_0}(U_{L_0}) & \\
C_{jk} =& f_{C_{jk}}(C_{j-1}, L_{j-1}, L_0, U_{C_j}) & j = 1,\dots,J;\ k = 1,\dots,K \\
L_{jk} =& C_{jk} f_{L_{jk}}(C_{j-1}, L_{j-1}, L_0, U_{L_j}) & j = 1,\dots,J;\ k = 1,\dots,K \\
Y_j =& f_{Y_j}(\bar{C}_j, \bar{L}_j, L_0, U_{Y_j}) & j = 1,\dots,J,
\end{aligned} \tag{5.5}$$

where, for a random variable X, f_X denotes an unknown but fixed function, U_X denotes all the unmeasured factors that are causally related to X, and $\bar{X}_j = (X_1,\dots,X_j)$ denotes the history of X up until time t_j. As pointed out by Pearl (2000), this model assumes that the data O for each patient are generated by the mechanistic process implied by the functions f_{X_j} with a temporal order dictated by the ordering of the time points t_j. In addition, this NPSEM encodes two important conditional independence assumptions:

$$L_{jk} \perp\!\!\!\perp L_{jk^*} | (L_0, L_{j-1}) \quad \forall j,\ k^* \neq k, \tag{5.6}$$

$$L_{jk} \perp\!\!\!\perp \bar{L}_{j-2} | (L_0, L_{j-1}) \quad \forall j,\ k. \tag{5.7}$$

Assumption (5.6) means that the variables L_{jk} at time t_j are drawn simultaneously as a function of the past only, and the contemporary variables do not interact with each other. Assumption (5.7) means that the value of a variable L_{jk} is only affected by the immeadiate past, and is not directly affected by any of the variables measured before or on time t_{j-2}.

We will define VIM measures in terms of the causal effect of L_{jk} on $Y_{j'}$, for all $j' \geq j$ and for all k. That is, we are interested in estimating the effect of a variable recorded at time point t_j on the hazard of death in each of the subsequent time intervals $(t_j, t_{j+1}], \dots, (t_{J-1}, t_J]$. This approach has the advantage that VIM can be seen as a dynamic process in which the factors that are decisive for developing a clinical outcome change as a function of time.

As a consequence of assumption (5.6), the problem of estimating the causal effect of each L_{jk} on each $Y_{j'}$ for $j' \geq j$ can be seen as a series of cross-sectional problems as follows. Note that L_{jk^*} for $k^* \neq k$ are not confounders of the causal relation between L_{jk} and $Y_{j'}$. To illustrate this, consider the NPSEM encoded in the directed aclyclic graph of Figure 5.1, in which for simplicity we assume that all covariates are observed (i.e., C variables are not present) and that $J = K = 2$. It stems from the graph that the variable L_{22} plays no role as a confounder of the causal effect of L_{21} on Y_2. Thus, for fixed j, $j' \geq j$, and k, and for each patient still at risk at $t_{j'}$, denoting $A \equiv L_{jk}$, $C \equiv C_{jk}$, $W \equiv (L_0, C_{j-1}, L_{j-1})$, and $Y_{j'} \equiv Y$, it suffices to consider the following simplified NPSEM:

$$W = f_W(U_W),\ C = f_C(W, U_C),\ A = C f_A(W, U_A),\ Y = f_Y(A, C, W, U_Y), \tag{5.8}$$

where the U variables denote all the exogenous, unobserved factors associated to each of the observed variables, and the functions f are deterministic but unknown and completely unspecified. Thus, from now on we will focus on the study of this data structure with observed data

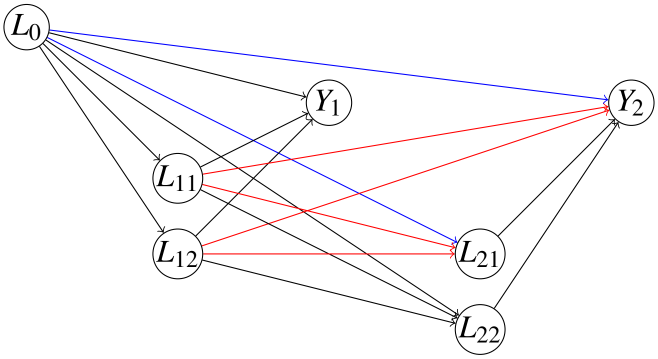


Figure 5.1: Directed acyclic graph, the arrows in blue and red denote the relations that confound the causal effect of L_{21} on Y_2

$O = (W,C,A,Y)$; the analysis will be done repatedly for each combination (j, j', k) of time points and variable of interest.

Some additional consequences of NPSEM (5.8) are:

(1) The missingness indicator C is allowed to depend on the covariates measured in the previous time point. In this way we take into account that a variable can be missing as a result of the previous health status of the patient, and also that it can be strongly correlated with previous missingness indicators.

(2) Missingness is informative. A patient's missingness indicator C is allowed to affect the way Y is generated, therefore acknowledging that missingness can contain information about the health outcome (e.g., sicker patients who will die earlier might be more likely to have missing values because information stops being recorded during life-threatening situations).

Without loss of generality we will assume that the variable A is either binary or continuous in the interval $(0,1)$. The true distribution of O will be denoted P_0, with $\bar{Q}_0(A,C,W) \equiv E_0(Y|A,C,W)$, $g_0(A|C,W) \equiv P_0(A|C,W)$, $\phi_0(C|W) \equiv P_0(C|W)$ and $Q_{W,0}(W) \equiv P_0(W)$.

For the analysis of the ACIT data we have classified the variables L_{jk} in two non-mutually exclusive categories: baseline and treatment variables. Baseline variables (L_0) are causally related to the outcome but can seldom be manipulated by the physician and are rarely of interest as possible care targets. Although baseline variables are not of interest in themselves, controlling for them is crucial when estimating the effect of treatment variables, which are often longitudinal variables that represent possible targets for clinical care. The label of each variable according to this classification is shown in table 5.1.

We will now define the VIM parameter for continuous and binary outcomes.

Continuous Variables. Consider an intervened system in which the variables are generated by the following system of equations

$$\begin{aligned} W &= f_W(U_W) \\ C^I &= 1 \\ A^I &= f_A(W, U_A) + \delta \\ Y^I &= f_Y(A^I, C^I, W, U_Y), \end{aligned} \tag{5.9}$$

which, for a small positive δ, can be interpreted as a model in which there is no missingness, and the distribution of the exposure variable A^I is shifted to the right by δ units. This type of intervention has been previously discussed in the literature (Díaz and van der Laan, 2011a), and belongs to a wider class of interventions known as stochastic interventions (Korb et al., 2004; Didelez, Dawid, and Geneletti, 2006; Dawid and Didelez, 2010). The parameter

$$E(Y^I) - E(Y)$$

can be causally interpreted as the expected reduction in mortality rate gained by an increase of δ units in the variable A for each patient. Since the counterfactual data $O^I = (W, C^I, A^I, Y^I)$ are not observed, $E(Y^I)$ is not estimable without further untestable assumptions. Under the randomization assumption (see, e.g., Rubin, 1978; Pearl, 2000) that

$$(C, A) \perp Y^I | W, \tag{5.10}$$

and the positivity assumption

$$g_0(A|W) > 0, \text{ and } \phi_0(1|W) > 0 \text{ for all } A \text{ and } W, \tag{5.11}$$

the expectation $E(Y^I)$ is identified as $E(Y^I) = E_W E\{\bar{Q}(A + \delta, C, W)|C = 1, W\}$, and the parameter of interest is defined as

$$\Psi_c(\bar{Q}, Q_W, g) \equiv E_W E\{\bar{Q}(A + \delta, 1, W)|C = 1, W\} - E(Y), \tag{5.12}$$

The true value of this parameter will be denoted $\psi_{c,0}$. A proof of this result under the randomization assumption is presented by Díaz and van der Laan (2011a). That proof follows the arguments for identification of general causal parameters given by Pearl (2000), who provides a unified framework for identification of counterfactual parameters as function of the observed data generating mechanism. Equation (5.12) defines the VIM measure for continuous exposures.

Binary Variables For binary variables, following the structural causal model described in (5.8), the VIM parameter is defined according to the following intervened system:

$$\begin{aligned} W &= f_W(U_W) \\ C^I &= 1 \\ A^I &= \begin{cases} 1 & \text{with probability } g(1|1, W) + \delta \\ 0 & \text{with probability } g(0|1, W) - \delta \end{cases} \\ Y^I &= f_Y(A^I, C^I, W, U_Y), \end{aligned}$$

where $0 < \delta < \sup_w g(0|1,w)$ is a user-given value. Under randomization assumption (5.10), and the positivity assumption

$$0 < g_0(1|W) < 1, \text{ and } \phi_0(1|W) > 0 \text{ for all } W, \tag{5.13}$$

the expectation of Y^I is identified as a function of the observed data generating mechanism as $E(Y^I) = E_W E\{\bar{Q}(A,C,W)|C=1,W\} + \delta\{E[\bar{Q}(1,1,W) - \bar{Q}(1,0,W)]\}$, and the parameter of interest is defined as

$$\Psi_b(\bar{Q},Q_W,g) \equiv E_W E\{\bar{Q}(A,C,W)|C=1,W\} + \delta\{E[\bar{Q}(1,1,W) - \bar{Q}(1,0,W)]\} - E(Y), \tag{5.14}$$

The true value of this parameter will be denoted $\psi_{b,0}$. Equation (5.14) defines the VIM measure for binary exposures that we will use in this paper.

Comparability We argue that the previous variable importance measures for continuous and binary VIM are comparable up to first order. First of all, note that, under the appropriate differentiability assumptions, for continuous A we have

$$\Psi_c(\bar{Q},Q_W,g) \approx E_W\{\bar{Q}(A,1,W)|C=1,W\} + \delta\frac{d}{d\delta}E_W E\{\bar{Q}(A+\delta,1,W)|C=1,W\}\bigg|_{\delta=0}.$$

This expression and (5.14) both have the form $a + \delta \times b$, where b can be seen as the appropriate slope of $E\{\bar{Q}(A,C,W)\}$ as a function of its first argument, providing an argument that, at least in first order, these two VIM measures are comparable.

Causal interpretation If model (5.5) does not hold, $\psi_{c,0}$ does not have a causal interpretation and must not be used to make treatment decisions. In that case, there are two main uses of this parameter. First, it can be interpreted as the importance of variable L_{jk} for predicting death in the interval (t_j,t_{j+1}) when only the patient's history $W = (L_0,C_{j-1},L_{j-1})$ and $A = L_{jk}$ have been measured. This prediction VIM measure may be used as a tool for determining the best set of prediction variables by ruling out those whose change from A to $A+\delta$ would not induce a considerable change in the expected prediction of a patient's outcome. Second, this VIM parameter may be used as a tool for formulating causal hypothesis that may be tested in a subsequent randomized study or in an observational study in which the necessary causal assumptions are met. An analogous argument is valid for the VIM parameter $\psi_{b,0}$ for binary variables.

In the following sections we will discuss double robust estimation methods for these parameters.

5.2 Estimation and prediction methods

We will first discuss the consistent and efficient estimation of the VIM parameters defined in the previous section, and then we will procede to discuss prediction methods for $\bar{Q}_0$, g_0 and ϕ_0.

VIM estimation

In order to define semi-parametric VIM estimates that have optimal asymptotic properties we first need to talk about the efficient influence function. The efficient influence function is a known function D of the data O and P_0, and is a key element in semi-parametric efficient estimation, since it defines the linear approximation of all efficient regular asymptotically linear estimators (Bickel et al., 1997). This means that the variance of the efficient influence function provides a lower bound for the variance of all regular asymptotically linear estimators, analogously to the Cramer-Rao lower bound in parametric models.

The efficient influence function of parameters (5.12) and (5.14) are given by

$$D_c(\bar{Q},Q_W,g,\phi)(O) = D_{c1}(\bar{Q},g,\phi)(O) + D_{c2}(\bar{Q},g,\phi)(O) + D_{c3}(\bar{Q},Q_W,g)(O) \tag{5.15}$$

$$D_b(\bar{Q},Q_W,g,\phi)(O) = D_{b1}(\bar{Q},g,\phi)(O) + D_{b2}(\bar{Q},g,\phi)(O) + D_{b3}(\bar{Q},Q_W,g)(O), \tag{5.16}$$

respectively, where

$$\begin{aligned} D_{c1}(\bar{Q},g,\phi)(O) &= \frac{C}{\phi(1|W)} \frac{g(A-\delta|1,W)}{g(A|1,W)} \{Y - \bar{Q}(A,1,W)\} \\ D_{c2}(\bar{Q},g,\phi)(O) &= \frac{C}{\phi(1|W)} \left[\bar{Q}(1,A+\delta,W) - E_g\{\bar{Q}(1,A+\delta,W)|C=1,W\}\right] \\ D_{c3}(\bar{Q},Q_W,g)(O) &= E_g\{\bar{Q}(1,A+\delta,W)|C=1,W\} - Y - \Psi_c(\bar{Q},Q_W,g), \end{aligned} \tag{5.17}$$

and

$$\begin{aligned} D_{b1}(\bar{Q},g,\phi)(O) &= \frac{C}{\phi(1|W)} \left(\delta \frac{2A-1}{g(A|1,W)} + 1\right) \{Y - \bar{Q}(A,1,W)\} \\ D_{b2}(\bar{Q},g,\phi)(O) &= \frac{C}{\phi(1|W)} [\bar{Q}(A,1,W) - E_g\{\bar{Q}(A,1,W)|C=1,W\}] \\ D_{b3}(\bar{Q},Q_W,g)(O) &= \delta\{\bar{Q}(1,1,W) - \bar{Q}(0,1,W)\} + E_g\{\bar{Q}(A,1,W)|C=1,W\} - Y - \Psi_b(\bar{Q},Q_W,g). \end{aligned} \tag{5.18}$$

Result 5 provides the conditions under which these estimating equations have expectation zero, therefore leading to consistent, triply robust estimators.

Result 5. *Let D be either D_c or D_b presented in equations (5.15) and (5.16). We have that*

$$E_{P_0}\{D(O|\phi,g,\bar{Q},\psi_0)\} = 0$$

if either ($\bar{Q} = \bar{Q}_0$ and $\phi = \phi_0$) or ($\bar{Q} = \bar{Q}_0$ and $g = g_0$) or ($g = g_0$ and $\phi = \phi_0$).

Recall that an estimator that solves an estimating equation will be consistent if the expectation of the estimating equation equals zero. As a consequence of this result, and under the conditions on $\bar{Q}$, g and ϕ stated in Theorem 5.11 and 6.18 of van der Vaart (2002), an estimator that solves the efficient influence function D will be consistent if either two of the three initial estimators are consistent, and it will be efficient if all of them are consistently estimated. Mathematical proofs of the efficiency of these estimators are out of the scope of this paper, but the general theory underlying their asymptotic properties can be found in van der Laan and Robins (2003), among others.

We will use targeted minimum loss based estimators (TMLE, van der Laan and Rubin, 2006; Rose and van der Laan, 2011) of the parameters Ψ_c and Ψ_b. TMLE is a substitution/plug-in estimation method that, given initial estimators $(\bar{Q}_n, Q_{W,n}, g_n)$ of $(\bar{Q}, Q_W, g)$, finds updated estimators $(\bar{Q}_n^*, Q_{W,n}^*, g_n^*)$ and defines the estimator of Ψ as

$$\psi_n = \Psi(\bar{Q}_n^*, Q_{W,n}^*, g_n^*).$$

TMLE is an estimation method that enjoys the best properties of both G-computation estimators (Robins, 1986) and the estimating equation methodology (see e.g., van de Geer, 2000; van der Laan and Robins, 2003). On one hand, TMLE is similar to G-computation estimators (e.g., $\Psi(\bar{Q}_n, Q_{W,n}, g_n)$) in that it is a plug-in estimator, and therefore produces estimates that are always within the range of the parameter of interest (e.g., it is always in the interval $[0,1]$ if the estimand is a probability). On the other hand, under regularity conditions and consistency of $(\bar{Q}_n, g_n, \phi_n)$, it is assymptotically linear with influence function equal to the efficient influence function:

$$\psi_n - \psi_0 = \sum_{i=1}^{n} D(P_0)(O_i) + o_P(1/\sqrt{n}).$$

As a consequence, TMLE has the following properties:

- It is a substitution/plug-in estimator.
- It is efficient if $\bar{Q}_n, g_n$, and ϕ_n are consistent for $\bar{Q}_0, g_0$, and ϕ_0, respectively.
- It is consistent if either $\bar{Q}_n$ or both g_n and ϕ_n are consistent. This property is refered to as double robustness.
- It is more robust to empirical violations of the positivity assumptions (5.11) and (5.13).

In the next subsection we will describe an iterative procedure that transforms the initial estimates $\bar{Q}_n$ and g_n into targeted estimates $\bar{Q}_n^*$ and g_n^* such that $\Psi(\bar{Q}_n^*, g_n^*, Q_{W,n}^*)$ is a TMLE of $\Psi(\bar{Q}_0, g_0, Q_{W,0})$.

TMLE algorithm

In order to define a targeted maximum likelihood estimator for ψ_0, we need to define three elements: (1) A loss function $L(Q)$ for the relevant part of the likelihood required to evaluate $\Psi(P)$, which in this case is $Q = (\bar{Q}, g, Q_W)$. This function must satisfy $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$, where Q_0 denotes the true value of Q; (2) An initial estimator Q_n^0 of Q_0; (3) A parametric fluctuation $Q(\varepsilon)$ through Q_n^0 such that the linear span of $\frac{d}{d\varepsilon} L\{Q(\varepsilon)\}|_{\varepsilon=0}$ contains the efficient influence curve $D(P)$ defined by either (5.15) or (5.16), depening on wheter A is continuous or binary. These elements are defined below:

Loss Function

As loss function for Q, we will consider $L(Q) = L_Y(\bar{Q}) + L_A(g) + L_W(Q_W)$, where $L_Y(\bar{Q}) = Y\log\{\bar{Q}(A,W)\} + (1-Y)\log\{1-\bar{Q}(A,W)\}$, $L_A(g) = -\log g(A|W)$, and $L_W(Q_W) = -\log Q_W(W)$. It can be easily verified that this function satisfies $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$.

Parametric Fluctuation

Given an initial estimator Q_n^k of Q_0, with components $(\bar{Q}_n^k, g_n^k, Q_{W,n}^k)$, we define the $(k+1)$th fluctuation of Q_n^k as follows:

$$\begin{aligned} \operatorname{logit}\bar{Q}_n^{k+1}(\varepsilon_1)(A,W) &= \operatorname{logit}\bar{Q}_n^k(A,W) + \varepsilon_1 H_1^k(C,A,W) \\ g_n^{k+1}(\varepsilon_1)(A|W) &\propto \exp\{\varepsilon_1 H_2^k(A,W)\} g_n^k(A|W) \\ Q_{W,n}^{k+1}(\varepsilon_2)(W) &\propto \exp\{\varepsilon_2 H_3^k(W)\} Q_{W,n}^k(W), \end{aligned}$$

where the proportionality constants are so that the left hand side terms integrate to one, for continuous A

$$H_1^k(A,C,W) = \frac{C}{\phi_n(1|W)} \frac{g_n^k(A-\delta|,1W)}{g_n^k(A|,1,W)},$$

for binary A

$$H_1^k(A,C,W) = \frac{C}{\phi_n(1|W)} \left(\delta \frac{2A-1}{g_n^k(A|1,W)} + 1\right),$$

$H_2^k(A,W) = D_2(P^k)(O)$, and $H_3(W) = D_3(P^k)(O)$, with D_2 and D_3 defined as in (5.17) and (5.18). We define these fluctuations using a two-dimensional ε with two different parameters ε_1 and ε_2, though it is theoretically correct to define these fluctuations using any dimension for ε, as far as the condition $D(P) \in < \frac{d}{d\varepsilon} L\{Q(\varepsilon)\}|_{\varepsilon=0} >$ is satisfied, where $< \cdot >$ denotes linear span. The convenience of the particular choice made here will be clear once the targeted maximum likelihood estimator (TMLE) is defined.

Targeted Maximum Likelihood Estimator

The TMLE is defined by the following iterative process:

1. Initialize $k = 0$.
2. Estimate ε as $\varepsilon_n^k = \arg\min_\varepsilon P_n L\{Q_n^k(\varepsilon)\}$.
3. Compute $Q_n^{k+1} = Q_n^k(\varepsilon_n^k)$.
4. Update $k = k+1$ and iterate steps 2 through 4 until convergence (i.e., until $\varepsilon_n^k = 0$)

First of all, note that the value of ε_2 that minimizes the part of the loss function corresponding to the marginal distribution of W in the first step (i.e., $-P_n \log Q_{W,n}^1(\varepsilon_2)$) is $\varepsilon_2^1 = 0$. Therefore, the iterative estimation of ε only involves the estimation of ε_1. The kth step estimation of ε_1 is obtained by minimizing $P_n(L_Y(\bar{Q}_n^k(\varepsilon_1)) + L_A(g_n^k(\varepsilon_1)))$, which implies solving the estimating equation

$$S^k(\varepsilon_1) = \sum_{i=1}^{n} \Bigg\{ \Big[Y_i - \text{expit}\{\text{logit}\,\bar{Q}_n^k(A_i,W_i) + \varepsilon_1 H_1^k(O_i)\}\Big] H_1^k(O_i) + D_2(P_n^k)(O_i) - \frac{\int_{\mathscr{A}} D_2(P_n^k)(Y_i,a,W_i)\ \exp\{\varepsilon_1 D_2(P_n^k)(Y_i,a,C_i,W_i)\}\ g_n^k(a|1,W_i)\,d\mu(a)}{\int_{\mathscr{A}} \exp\{\varepsilon_1 D_2(P_n^k)(Y_i,a,C_i,W_i)\}\ g_n^k(a|1,W_i)\,d\mu(a)} \Bigg\} \tag{5.19}$$

where

$$D_2(P_n^k)(O) = \bar{Q}_n^k(A+\delta,1,W) - \int_{\mathscr{A}} \bar{Q}_n^k(a+\delta,1,W) g_n^k(a|1,W)\,d\mu(a).$$

The TMLE of ψ_0 is defined as $\psi_n \equiv \lim_{k\to\infty} \Psi(P_n^k)$, assuming this limit exists. In practice, the iteration process is carried out until convergence in the values of ε_k is achieved, and an estimator Q_n^* is obtained. Under the conditions of Theorem 2.3 of van der Laan and Robins (2003), a conservative estimator of the variance of ψ_n is given by

$$\frac{1}{n}\sum_{i=1}^{n} D^2(\bar{Q}_n^*, Q_{W,n}, g_n^*, \phi_n)(O_i).$$

An R function that computes the TMLE of ψ_0 can be found in appendix A.2.

Augmented IPTW, stepwise, and undadjusted estimators. In addition to the TMLE we will compute three additional estimates of the VIM, for comparison with other estimation methods. The first estimator, the augmented IPTW (AIPTW), is an estimator that uses the efficient influence function of the parameter in order to define the estimator as the solution of the corresponding estimating equation. Because the AIPTW is also asymptotically linear with influence function equal to the efficient influence function, it is consistent and assymptotically efficient. However, the estimating equation that defines the AIPTW may not have a solution in the parameter space, in which case the AIPTW does not exist. Additionally, the AIPTW is more sensitive to practical violations to the positivity assumption conmpared to the TMLE. The second estimator, a g-computation formula based on stepwise regression (SW) represents common practice in statistics. The SW estimator requires initial estimators ϕ_n and $\bar{Q}_n$ of ϕ_0 and $\bar{Q}_0$, and is defined as

$$\begin{aligned}
\psi_{c,n,SW} =& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{C_i}{\phi_n(W_i)}\bar{Q}_n(A_i+\delta,1,W_i) - Y_i\right\} \\
\psi_{b,n,SW} =& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{C_i}{\phi_n(W_i)}\bar{Q}_n(A_i,1,W_i) + \delta[\bar{Q}_n(1,1,W_i) - \bar{Q}_n(0,1,W_i)] - Y_i\right\},
\end{aligned}$$

for Ψ_c and Ψ_b, respectively. The unadjusted estimator is identical to the SW estimator but including only the intercept term in the vector W.

Since the consistency of the initial estimators of $\bar{Q}_0$, g_0 and ϕ_0 is key to attain estimators with optimal statistical properties (i.e., consistency and efficiency), we will carefully discuss the construction of such estimators in the next subsection. In particular, the next subsection deals with the construction of an estimator for $\bar{Q}_0$, the predictor of death in our working example.

Prediction

As explained in the previous section, the consistency of the initial estimators $\bar{Q}_n$, g_n and ψ_n determine the statistical properties of the estimators of $\psi_{c,0}$ and $\psi_{b,0}$. Common practice in statistics involves the estimation of models like

$$\operatorname{logit} \bar{Q}(A,W) = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 AW. \tag{5.20}$$

This approach that has gained popularity among researchers in epidemiology and biostatistics, partly because of the analysis of its statistical properties requires simple mathematical methods, and partly because it is readily available in every statistical software. Nevertheless, as it is also well known among their users, parametric models of the type described by (5.20) are rarely correct, and their choice is merely based on their computational advantages and other subjective criteria. This practice leads to regression estimator whose usefulness is highly questionable given that the assumptions it entails (linearity, normality, link function, etc.) do not originate in legitimate knowledge about the phenomena under study, but rather come from analytical tractability and computational convenience.

In this paper we will use the super learner (van der Laan, Polley, and Hubbard, 2007) for estimation of $\bar{Q}_0$, g_0, and ψ_0. Super learner is a methodology that uses cross-validated risks to find an optimal combination of a list of user-supplied estimation algorithms. One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution), thus providing the closest approximation to the real data generating mechanism. Proofs and simulations regarding these and other asymptotic properties of the super learner can be found in van der Laan, Dudoit, and Keles (2004) and van der Laan and Dudoit (2003).

To implement the super learner predictor it is necessary to specify a library of candidate predictors algorithm. In the case of the conditional expectations $\bar{Q}_0$, ϕ_0, and g_0 for binary A, the candidates can be any regression or classification algorithm. Examples include random forests, logistic regression, k nearest neighbors, Bayesian models, etc. For estimation of the conditional densities g_0 we will also use the super learner, with candidates given by several histogram density estimators, which yields a piece-wise constant estimator of the conditional density. The choice of the number of bins and their location is indexed by two tuning parameters. The implementation of this density estimator is discussed in detail by Díaz and van der Laan (2011b), and will be ommited in this paper.

5.3 Data analysis

In this section we analyze the data described in the example of section 5.1. The sample size was $n = 918$ patients, and measurements of the variables described in table 5.1 were taken at 6, 12, 24, 48, and 72 hours after admision to the emergency room.

The main objective of the study was the construction of prediction models for the risk of death of a patient in a certain time interval given the variables measured up to the start of the interval, as well as the definition and estimation of VIM measures that provide an account of the longitudinal evolution of the relation between these physiological and clinical measurements and the risk of death at a certain time point.

The data set was partitioned in 6 different data sets according to the time intervals defined by the time points in which measurements were taken, each of these 6 datasets contained only the patients that were at risk of death (alive) at the start of the time interval. Each of the continuous covariates was rescaled by subtracting the minimum and dividing by the range so that all of the covariates range between zero and one. The methods described in the previous sections were applied to each variable in each of these datasets.

The candidate algorithms for prediction of death used in the super learner predictor are as follows:

- Logistic regression with main terms (GLM)
- Stepwise logistic regression (SW)
- Bayesian logistic regression (BLR)
- Generalized additive models (GAM)
- Earth (Earth)
- Sample mean (MEAN),

from which the first three represent common practice in epidemiology and statistics, GAM and Earth are algorithms that intend to capture non-parametric structures of the data, and the sample mean is included for contrast.

Table 5.2 shows the coefficients of each candidate algorithm in the super learner predictor of $E(Y_j|\bar{L}_j,\bar{C}_j,L_0)$. The variability in these coefficients shows that no single algorithm is optimal for prediction at each time point, and that each algorithm describes certain features of the data generating mechanism that the others are not capable of unveiling.

	0-6hr	6-12hr	12-24hr	24-48hr	48-72hr	72+hr
GLM	0.0000	0.0000	0.0000	0.0318	0.0259	0.0000
SW	0.0000	0.1889	0.0000	0.0000	0.2073	0.1787
BGLM	0.3318	0.0586	0.1049	0.1329	0.0313	0.2750
GAM	0.5118	0.7525	0.8951	0.8353	0.7201	0.2487
Earth	0.1563	0.0000	0.0000	0.0000	0.0154	0.1298
MEAN	0.0000	0.0000	0.0000	0.0000	0.0000	0.1678

Table 5.2: Coefficients in the Super Learner

Figure 5.2 presents the ROC curves for the cross-validated super learning predictions of death, as well as the cross-validated predictions based on a logistic model with AIC-based stepwise selection of variables, for comparison with common practice. The super learner prediction methods outperforms the stepwise prediction in all cases, with AUC ROC (area under the ROC curve) differences ranging from 0.02 to 0.07. Though this differences might be small, an interpretation of their meaning reveals the clinical relevance of a slight improvement in prediction. The AUC ROC can be interpreted as the proportion of times that a patient who will die obtains a higher prediction score than a patient who will survive. In practice, an AUC ROC difference of 0.02 means that in 100 pairs of patients (pairs formed by one patient who will die and one who will not) the super learner classifier will correctly classify two pairs more than the step-wise classifier, which could potentially lead to live-saving treatments for these two patients.

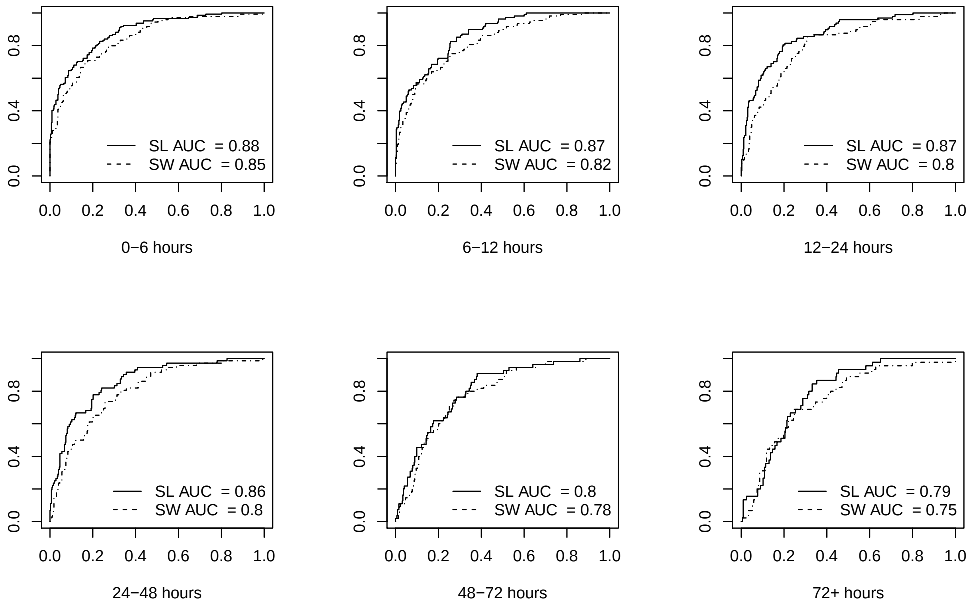


Figure 5.2: ROC curves of cross-validated prediction for the super learner (SL) and the logistic step-wise regression (SW), for different time intervals.

The VIM measures computed for each variable at each time point were ranked according to their p-value. Table 5.3 presents the first five most important variables for prediction of death at each time interval, according to the TML estimator presented in section 5.2. Recall that all the continuous variables were re-scaled between zero and one; the value $\delta = 0.01$ was used for all the estimates. The interpretation of the values in the first row of table 5.3, for example, is that if PT were to increase by 1% for every patient, the mortality rate in the first time interval would be augmented by 2.13%, according to the TMLE. The TMLE and the AIPTW produced generally similar results, whereas the SW estimator produced results that are not significant. This is because

TMLE and AIPTW are efficient estimators that make the best use of the information available in order to find precise and reliable estimates of the VIM measure.

In addition to the previous tables, Figure 5.3 shows heatmaps of the effect that a modification on each variable at a given time point would have in the mortality rates of the subsequent time intervals (i.e., the VIM measure). For example, Figure 5.3a shows the effect that an intervention on each of the variables measured at baseline will have on the outcome between 0-6 hours, 6-12 hours, 12-24 hours, 24-48 hours, 48-72 hours, and 72+ hours. Additionally, the dendrogram plotted in the left margin of Figure 5.3a shows a hierarchical clustering of the variables acording to the profile of their effect on the longitudinal outcome. For example, an increase in variables DDIM through CREA at baseline would cause an increase (significative most of the times) in mortality rate, particularly at earlier time points. Variables RR through SBP seem to have a small protective effect that is significative in few cases, and variables PF12MN through TPA have estimated protective effect whose statistical significance could not be established with this sample.

At each time point, variables that less than 15% of observed values were not included in the analysis. For this reason, and because missingness was more common in later measurement times, the number of variables included in Figure 5.3 decreases as the time of measurement increases. Additionally, the output for variables measured at 48 and 72 hours is not shown because none of the results were significant.

5.4 Discussion

The techniques presented in this paper provide a methodology for computing prediction algorithms and VIM measures that overcome the difficulties presented in the introduction, namely the lack of interpretability and optimality of the estimates. The superiority of the super learner as a prediction method has been proven analytically by several authors, and is corroborated in our example by comparison with current practice in epidemiology and biostatistics. The VIM measures that were defined provide parameters with a valid causal and statistical interpretation, independent of the statistical model or prediction method selected. The use of consistent and efficient estimates of the VIM parameters was also demonstrated, and the importance of using these optimal estimators was exemplified in the application section.

VarName	VarTime	TMLE	IPTW	StepwiseGLM	Unadjusted
PT	00	0.0213(<0.001)	0.0215(<0.001)	-0.0011(0.698)	0.0400(<0.001)
ISS	00	-0.0320(<0.001)	-0.0322(<0.001)	-0.0242(<0.001)	-0.0244(0.002)
FVIII	00	0.0105(0.027)	0.0108(0.023)	0.0085(0.839)	0.0827(<0.001)
APC	00	0.0205(0.028)	0.0202(0.031)	0.0235(0.386)	0.1063(<0.001)
INR	00	0.0109(0.113)	0.0065(0.343)	-0.0011(0.722)	0.0345(<0.001)

(a) Death between 0 and 6 hours

VarName	VarTime	TMLE	IPTW	StepwiseGLM	Unadjusted
CREA	00	0.0289(<0.001)	0.0289(<0.001)	0.0013(0.020)	0.0338(<0.001)
HCT	00	0.0383(<0.001)	0.0383(<0.001)	-0.0005(0.384)	0.0281(<0.001)
HGB	00	0.0396(<0.001)	0.0396(<0.001)	-0.0004(0.485)	0.0281(<0.001)
BUN	00	0.0260(<0.001)	0.0258(<0.001)	0.0009(0.441)	0.0347(<0.001)
PT	00	0.0088(<0.001)	0.0086(<0.001)	0.0019(0.426)	0.0376(<0.001)

(b) Death between 6 and 12 hours

VarName	VarTime	TMLE	IPTW	StepwiseGLM	Unadjusted
CREA	00	0.0281(<0.001)	0.0282(<0.001)	0.0012(0.033)	0.0314(<0.001)
HCT	00	0.0559(<0.001)	0.0558(<0.001)	-0.0006(0.278)	0.0262(<0.001)
HGB	00	0.0578(<0.001)	0.0577(<0.001)	-0.0004(0.467)	0.0264(<0.001)
BUN	00	0.0240(<0.001)	0.0238(<0.001)	0.0003(0.799)	0.0319(<0.001)
PT	00	0.0077(<0.001)	0.0074(0.001)	0.0019(0.377)	0.0357(<0.001)

(c) Death between 12 and 24 hours

VarName	VarTime	TMLE	IPTW	StepwiseGLM	Unadjusted
CREA	00	0.0565(<0.001)	0.0565(<0.001)	0.0013(0.051)	0.0266(<0.001)
HCT	00	0.0566(<0.001)	0.0564(<0.001)	-0.0002(0.732)	0.0224(<0.001)
HGB	00	0.0389(<0.001)	0.0390(<0.001)	-0.0001(0.862)	0.0224(<0.001)
APC	00	0.0532(<0.001)	0.0539(<0.001)	0.0226(0.310)	0.0761(<0.001)
ISS	00	-0.0233(<0.001)	-0.0236(<0.001)	-0.0211(<0.001)	-0.0155(0.020)

(d) Death between 24 and 48 hours

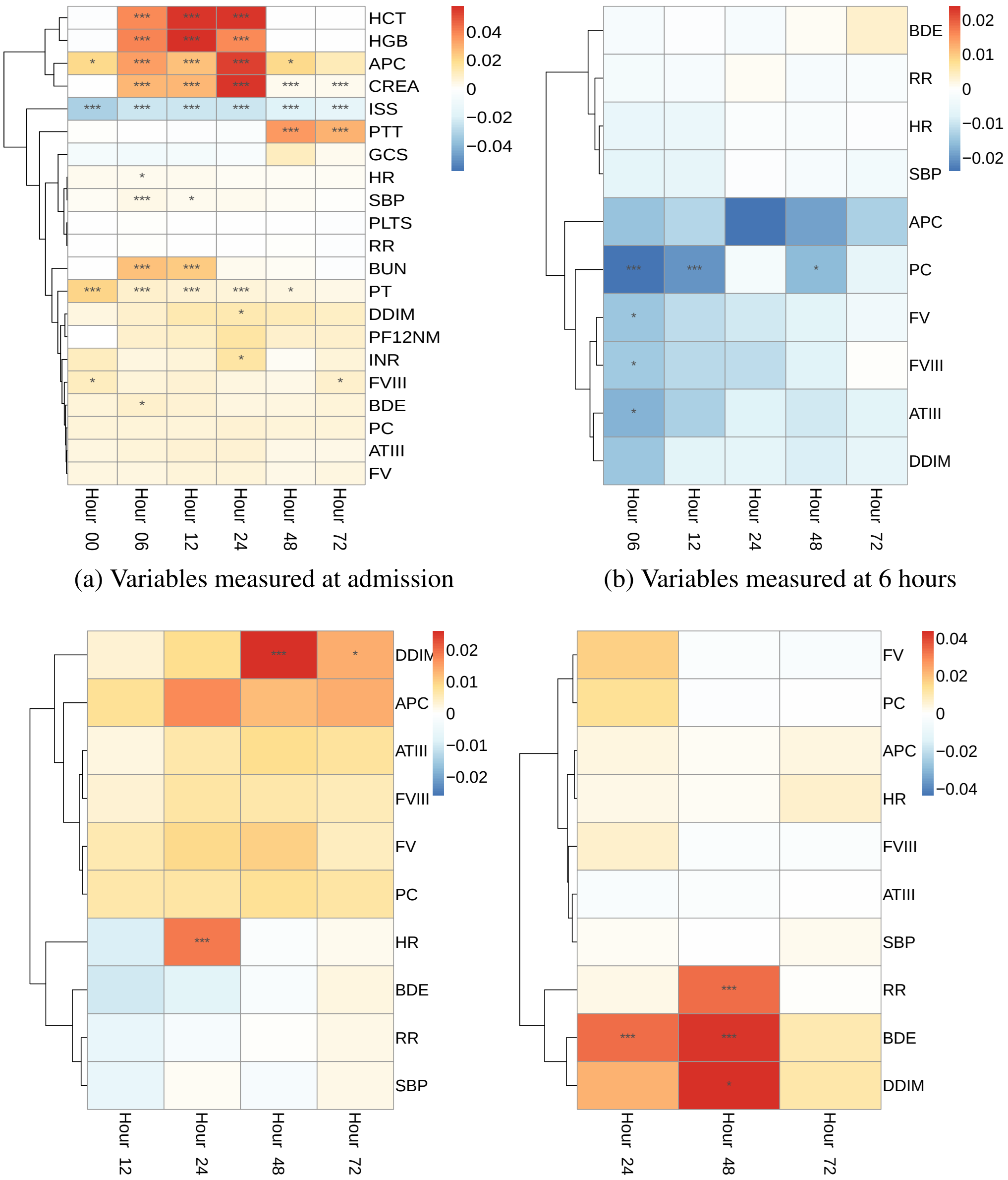
VarName	VarTime	TMLE	IPTW	StepwiseGLM	Unadjusted
PTT	00	0.0351(<0.001)	0.0350(<0.001)	0.0018(0.016)	0.0261(<0.001)
RR	24	0.0339(<0.001)	0.0304(<0.001)	0.0057(0.749)	0.0644(<0.001)
BDE	24	0.0426(<0.001)	0.0501(<0.001)	0.0140(0.870)	0.0828(<0.001)
DDIM	12	0.0260(<0.001)	0.0263(<0.001)	0.0481(0.885)	0.0605(<0.001)
ISS	00	-0.0195(<0.001)	-0.0202(<0.001)	-0.0180(<0.001)	-0.0116(0.049)

(e) Death between 48 and 72 hours

VarName	VarTime	TMLE	IPTW	StepwiseGLM	Unadjusted
PTT	00	0.0300(<0.001)	0.0298(<0.001)	0.0017(0.012)	0.0220(<0.001)
CREA	00	0.0026(0.004)	0.0026(0.004)	0.0007(0.291)	0.0168(<0.001)
ISS	00	-0.0140(0.008)	-0.0148(0.004)	-0.0145(0.008)	-0.0085(0.124)
DDIM	12	0.0138(0.017)	0.0116(0.049)	0.0669(0.693)	0.0604(<0.001)
FVIII	00	0.0087(0.034)	0.0097(0.015)	0.0046(0.525)	0.0500(<0.001)

(f) Death after 72 hours

Table 5.3: VIM estimates for the five most important variables for prediction of death at each time interval according to TML estimate (p-values in parentheses and truncated at 0.001).



(a) Variables measured at admission

(b) Variables measured at 6 hours

(c) Variables measured at 12 hours

(d) Variables measured at 24 hours

Figure 5.3: VIM estimates of measured variables according to TMLE. ‘***’ indicates p-value $\leq$ 0.001, ‘**’ indicates 0.001 $<$ p-value $\leq$ 0.01, and ‘*’ indicates 0.01 $<$ p-value $\leq$ 0.05.

Appendix A

R functions

A.1 R function `tmle.shift()` for chapter 2

Arguments

Argument	Description
Y	Outcome vector.
A	Treatment vector.
W	Covariates matrix.
Qn	An initial estimator of $\bar{Q}_0$ in the form of a function that takes a vector **A** and a matrix **W** and returns the vector of conditional expectations of Y given **A** and **W**.
gn	An initial estimator g_0 that takes as input a vector **A** and a matrix **W** and returns the density of A conditional on W at points **A**.
delta	A function of W defining the parameter of interest.
tol	Tolerance value for the convergence of ε.
max.iter	Maximum of iterations allowed.
Aval	A vector with equally spaced values indicating a partition of the support of A over which to compute Riemann sums to approximate the integrals involved in the estimation process.

Table A.1: Arguments of the R function tmle.shift

Code

```
tmle.shift <- function(Y, A, W, Qn, gn, delta, tol = 1e-5, iter.max = 5, Aval){
  # interval partition length
  h.int <- Aval[3]-Aval[2]
  # this function takes as input initial estimator of Q and g and returns
  # their updated value
  f.iter <- function(Qn, gn, gn0d = NULL, prev.sum = 0, first = FALSE){
      # numerical integrals and equation (7)
      Qnd <- t(sapply(1:nrow(W), function(i)Qn(Aval + delta, W[i,])))
      gnd <- t(sapply(1:nrow(W), function(i)gn(Aval, W[i,])))
      gnd <- gnd/rowSums(gnd)
      if(first) gn0d <- gnd
      EQnd <- rowSums(Qnd*gnd)*h.int
      D2   <- Qnd - EQnd
      QnAW <- Qn(A, W)
      H1   <- gn(A - delta, W)/gn(A, W)
      # equation (8)
      est.equation <- function(eps){
        sum((Y - (QnAW + eps*H1)) * H1 + (Qn(A + delta, W) - EQnd) -
        rowSums(D2*exp(eps*D2 + prev.sum)*gn0d)/rowSums(exp(eps*D2 + prev.sum)*gn0d))
      }
      eps  <- uniroot(est.equation, c(-1, 1))$root
      # updated values
      gn.new   <- function(a, w)exp(eps*Qn(a + delta, w)) * gn(a, w)
      Qn.new   <- function(a, w)Qn(a, w) + eps * gn(a - delta, w)/gn(a, w)
      prev.sum <- prev.sum + eps*D2
      return(list(Qn = Qn.new, gn = gn.new, prev.sum =
                       prev.sum, eps = eps, gn0d = gn0d))
  }
  ini.out <- f.iter(Qn, gn, first = TRUE)
  gn0d    <- ini.out$gn0d
  iter = 0
  # iterative procedure
  while(abs(ini.out$eps) > tol & iter <= iter.max){
    iter = iter + 1
    new.out <- f.iter(ini.out$Qn, ini.out$gn, gn0d, ini.out$prev.sum)
    ini.out <- new.out
  }
Qnd <- t(sapply(1:nrow(W), function(i)ini.out$Qn(Aval + delta, W[i,])))
gnd <- t(sapply(1:nrow(W), function(i)ini.out$gn(Aval, W[i,])))
gnd <- gnd/rowSums(gnd)
# plug in tmle
psi.hat <- mean(rowSums(Qnd*gnd)*h.int)
# influence curve of tmle
IC      <- (Y - ini.out$Qn(A, W))*ini.out$gn(A - delta, W)/ini.out$gn(A, W) +
           ini.out$Qn(A + delta, W) - psi.hat
var.hat <- var(IC)/length(Y)
return(c(psi.hat = psi.hat, var.hat = var.hat, IC = IC))
}
```

Example

Here is an example of how to use the previous function based on the data generating mechanism presented in the simulation

```
n  <- 100
W <- data.frame(W1 = runif(n), W2 = rbinom(n, 1, 0.7))
A  <- rpois(n, lambda = exp(3 + .3*log(W$W1) - .2*exp(W$W1)*W$W2))
Y  <- rbinom(n, 1, plogis(-1 + .05*A - .02*A*W$W2 + .2*A*tan(W$W1^2) -
      .02*W$W1*W$W2 + 0.1*A*W$W1*W$W2))
fitA.0 <- glm(A ~ I(log(W1)) + I(exp(W1)):W2, family = poisson, data =
              data.frame(A, W))
fitY.0 <- glm(Y ~ A + A:W2 + A:I(tan(W1^2)) + W1:W2 + A:W1:W2, family =
          binomial, data = data.frame(A, W))
gn.0  <- function(A = A, W = W)dpois(A, lambda = predict(fitA.0,
                  newdata = W, type = "response"))
Qn.0 <- function(A = A, W = W)predict(fitY.0, newdata = data.frame(A, W,
        row.names = NULL), type = "response")
tmle00 <- tmle.shift(Y, A, W, Qn.0, gn.0, delta=2, tol = 1e-4, iter.max = 5,
          Aval = seq(1, 60, 1))
```

A.2 R functions for chapter 5

```
## Qn, phin, and gn are the initial estimators of Q, phi, and g.
## Qn and gn should be given in form of a list with n elements, each
## containing a function of A. For example Qn[[1]]
## should be a function of a containing E(Y|A=a, C=1,W_1).
## phin is a vector of size n containing P(C=1|W).
## Aval is a range of values of A used to integrate over, only useful
## for continuous A, but included
## in the function for binary A for compatibility.
## tol is the tolerance on epsilon
## iter.max is the maximum of iterations of the TMLE algorithm

## Function for binary exposures
tmle.bin <- function(Y, A, C, Qn, phin, gn, delta, tol = 1e-5,
                iter.max = 5, Aval=c(0,1), lengths=diff(Aval)){

    n  <- length(Y)
    wC <- C/phin
    wC <- as.vector(wC[,1])

    f.iter <- function(Qn, gn, gn0d = NULL, prev.sum = 0, first = FALSE, iter){

        ## computation of different parts of the efficient influence function
        Qnd <- t(sapply(1:n,function(i)sapply(Aval, function(a)Qn[[i]](a))))
        gnd <- t(sapply(1:n,function(i)sapply(Aval, function(a)gn[[i]](a))))
        gnd <- gnd/rowSums(t(t(gnd) * lengths))

        EQnd <- rowSums(t(t(Qnd * gnd) * lengths))
        Qn1W <- sapply(1:n, function(i){Qn[[i]](1)})
        Qn0W <- sapply(1:n, function(i){Qn[[i]](0)})
        QnAW <- sapply(1:n, function(i){A[i]*Qn1W[i] + (1-A[i])*Qn0W[i]})
        D2AW <- wC * sapply(1:n, function(i){Qn[[i]](A[i]) - EQnd[i]})
        H1   <- wC * sapply(1:n, function(i){delta * (2*A[i] - 1) /
                                             gn[[i]](A[i]) + 1})

        IPTW <- varIPTW <- NULL

        ## if it is the first iteration, compute the augmented IPTW
        if(first){
            gn0d    <- gnd
            IPTW    <- mean(H1 * (Y - QnAW) + D2AW + EQnd +
                                 delta * (Qn1W - Qn0W) - Y)
            varIPTW <- var( H1 * (Y - QnAW) + D2AW + EQnd +
                                 delta * (Qn1W - Qn0W) - Y)/n
        }

        D2   <- wC * (Qnd - EQnd)

        ## estimating equation
        est.equation <- function(eps){
            sum((Y - expit(logit(QnAW) + eps * H1)) * H1 + D2AW -
                rowSums(D2 * exp(eps * D2 + prev.sum) * gn0d) /
                        rowSums(exp(eps * D2 + prev.sum) * gn0d))
        }

        ## solve the estimating equation
        eps <- try(uniroot(est.equation, c(-1, 1))$root, silent = TRUE)

        gn.new   <- Qn.new <- list()
        file <- paste('Qngn', iter, sample(1e5, 1), '.r', sep='')
        if(file.exists(file))file.remove(file)

        ## update g and Q
        for(i in 1:n){
            cat("gn.new[[",i,"]] <- function(A)exp(eps * wC[", i,"] *
                                      Qn[[", i, "]](A + delta)) *
                                      gn[[", i, "]](A)\n",
                sep = '', file = file, append = TRUE)
            cat("Qn.new[[",i,"]] <- function(A)expit(logit(Qn[[", i, "]](A))
                                       + eps * wC[", i,"] * gn[[", i, "]](A - delta) /
                                       gn[[", i, "]](A))\n",
                sep = '', file = file, append = TRUE)
        }

        source(file, local = T)
        file.remove(file)

        gn.new <- lapply(gn.new, Vectorize)
        Qn.new <- lapply(Qn.new, Vectorize)

        ## this avoids recursive computation of numerical integrals
        prev.sum <- prev.sum + eps * D2

        return(list(Qn = Qn.new, gn = gn.new, prev.sum = prev.sum,
                    eps = eps, gn0d = gn0d, IPTW = IPTW, varIPTW = varIPTW))

    }

    ## initiate
    iter = 0
    ini.out <- f.iter(Qn = Qn, gn = gn, gn0d = NULL, prev.sum = 0,
                      first = TRUE, iter = 0)

    ## extract A-IPTW
    gn0d    <- ini.out$gn0d
    IPTW    <- ini.out$IPTW

    ## iterate
    while(abs(ini.out$eps) > tol & iter <= iter.max){

        cat("iter ", iter, " started at ", date(), file = "iters.txt", append = TRUE)
        iter = iter + 1
        new.out <- f.iter(ini.out$Qn, ini.out$gn, gn0d, ini.out$prev.sum, FALSE, iter)
        ini.out <- new.out

    }

    ## compute TMLE and influence function
    Qn <- ini.out$Qn
```

```
    gn <- ini.out$gn

    Qnd <- t(sapply(1:n,function(i)sapply(Aval, function(a)Qn[[i]](a))))
    gnd <- t(sapply(1:n,function(i)sapply(Aval, function(a)gn[[i]](a))))
    gnd <- gnd/rowSums(t(t(gnd) * lengths))

    EQnd <- rowSums(t(t(Qnd * gnd) * lengths))
    Qn1W <- sapply(1:n, function(i){Qn[[i]](1)})
    Qn0W <- sapply(1:n, function(i){Qn[[i]](0)})
    QnAW <- sappli(1:n, function(i){A[i]*Qn1W[i] + (1-A[i])*Qn0W[i]})
    D2AW <- wC * sapply(1:n, function(i){Qn[[i]](A[i]) - EQnd[i]})
    H1   <- wC * sapply(1:n, function(i){delta * (2*A[i] - 1) / gn[[i]](A[i]) + 1})

    psi.hat <- mean(EQnd + delta * (Qn1W - Qn0W) - Y)
    IC      <- H1 * (Y - QnAW) + D2AW + EQnd + delta * (Qn1W - Qn0W) - Y

    var.hat <- var(IC)/n
    meanIC  <- mean(IC)

    return(c(psi.hat = psi.hat, var.hat = var.hat, meanIC = meanIC, iter = iter,
           IPTW = IPTW, varIPTW = varIPTW))

}

tmle.shift <- function(Y, A, C, Qn, phin, gn, delta, tol = 1e-5, iter.max = 5,
                Aval, lengths=diff(Aval)){

    n  <- length(Y)
    wC <- C/phin
    wC <- as.vector(wC[,1])

    wA <- function(gn, A, delta){
        ## ifelse(gn(A) == 0, 1, gn(A-delta)/gn(A))
        gn(A-delta)/gn(A)
    }

    f.iter <- function(Qn, gn, gn0d = NULL, prev.sum = 0, first = FALSE, iter){

        Qnd <- t(sapply(1:n,function(i)sapply(Aval, function(a)Qn[[i]](a))))
        gnd <- t(sapply(1:n,function(i)sapply(Aval, function(a)gn[[i]](a))))
        gnd <- gnd/rowSums(t(t(gnd) * lengths))

        EQnd <- rowSums(t(t(Qnd * gnd) * lengths))
        QnAW <- sapply(1:n, function(i){Qn[[i]](A[i])})
        D2AW <- wC * sapply(1:n, function(i){Qn[[i]](A[i] + delta) - EQnd[i]})
        H1   <- wC * sapply(1:n, function(i){wA(gn[[i]], A[i], delta)})

        IPTW <- varIPTW <- NULL

        if(first){
            gn0d    <- gnd
            IPTW    <- mean(H1 * (Y - QnAW) + D2AW + EQnd - Y)
            varIPTW <- var(H1 * (Y - QnAW) + D2AW + EQnd - Y)/n
        }

        D2   <- wC * (Qnd - EQnd)

        est.equation <- function(eps){
            sum((Y - expit(logit(QnAW) + eps * H1)) * H1 + D2AW -
                rowSums(D2 * exp(eps * D2 + prev.sum) * gn0d) /
```

```
                    rowSums(exp(eps * D2 + prev.sum) * gn0d))
    }

    eps <- try(uniroot(est.equation, c(-1, 1))$root, silent = TRUE)

    cat(’eps = ’, eps, ’\n’, sep = ’’)

    ##
    if(inherits(eps, ’try-error’)) return(list(est.eq = est.equation))

    gn.new   <- Qn.new <- list()
    file <- paste(’Qngn’, iter, sample(1e5, 1),’.r’, sep=’’)
    if(file.exists(file))file.remove(file)

    for(i in 1:n){
        cat("gn.new[[",i,"]] <- function(A)exp(eps * wC[", i,"] *
                          Qn[[", i, "]](A + delta)) * gn[[", i, "]](A)\n",
            sep=’’, file = file, append = TRUE)
        cat("Qn.new[[",i,"]] <- function(A)expit(logit(Qn[[", i, "]](A))
               + eps * wC[", i,"] * wA(gn[[", i, "]], A, delta))\n",
            sep=’’, file = file, append = TRUE)
    }

    cat(file)
    source(file, local = T)
    file.remove(file)

    prev.sum <- try(prev.sum + eps * D2)

    ##
    if(inherits(prev.sum, ’try-error’)) return(list(est.eq = est.equation))

    return(list(Qn = Qn.new, gn = gn.new, prev.sum = prev.sum, eps= eps,
          gn0d = gn0d, IPTW = IPTW, varIPTW = varIPTW, est.eq = est.equation))

}

iter = 0
ini.out <- f.iter(Qn = Qn, gn = gn, gn0d = NULL, prev.sum = 0, first = TRUE, iter = 0)

gn0d    <- ini.out$gn0d
IPTW    <- ini.out$IPTW
varIPTW <- ini.out$varIPTW

##
save(list = ls(), file = ’iter0.RData’)
if(inherits(try(abs(ini.out$eps)), ’try-error’)) return(ini.out)

while(abs(ini.out$eps) > tol & iter < iter.max){

    cat("iter ", iter, " started at ", date(), file = "iters.txt", append = TRUE)
    new.out <- f.iter(ini.out$Qn, ini.out$gn, gn0d, ini.out$prev.sum, FALSE, iter)
    ini.out <- new.out
    iter = iter + 1

    ##
    save(list = ls(), file = paste(’iter’, iter, ’.RData’, sep = ’’))
    if(inherits(try(abs(ini.out$eps)), ’try-error’)) return(ini.out)

}
```

```
    Qn <- ini.out$Qn
    gn <- ini.out$gn

    Qnd <- t(sapply(1:n, function(i)sapply(Aval, function(a)Qn[[i]](a))))
    gnd <- t(sapply(1:n, function(i)sapply(Aval, function(a)gn[[i]](a))))
    gnd <- gnd/rowSums(t(t(gnd) * lengths))

    EQnd <- rowSums(t(t(Qnd * gnd) * lengths))
    QnAW <- sapply(1:n, function(i){Qn[[i]](A[i])})
    D2AW <- wC * sapply(1:n, function(i){Qn[[i]](A[i] + delta) - EQnd[i]})
    H1   <- wC * sapply(1:n, function(i){gn[[i]](A[i] - delta)/gn[[i]](A[i])})

    psi.hat <- mean(EQnd - Y)
    IC      <- H1 * (Y - QnAW) + D2AW + EQnd - psi.hat - Y

    var.hat <- var(IC)/n
    meanIC  <- mean(IC)

    return(c(psi.hat = psi.hat, var.hat = var.hat, meanIC = meanIC,
           iter = iter, IPTW = IPTW, varIPTW = varIPTW))

}
```

Appendix B

Proofs

B.1 Theorem 1

This proof follows closely the proofs presented in Zheng and van der Laan, 2010 for a general CV-TMLE. Those proofs rely heavily on empirical process theory, of which van der Vaart and Wellner, 1996 provide a complete study.

*Proof o*f Theorem 1. First of all note that

$$R(\psi, Q) - R(\psi, Q_0) = -P_0 D\{\bar{Q}, Q_W, g_0, \psi\},$$

which implies that

$$\hat{R}(\hat{\Psi}_k) - \hat{R}_0(\hat{\Psi}_k) = -E_S P_0 D\{\hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), g_0, \hat{\Psi}_k(\mathbb{P}_T)\}.$$

Note that

$$E_S \mathbb{P}_V D\{\hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} = 0,$$

so that we can write

$$\begin{aligned}
\hat{R}(\hat{\Psi}_k) - \hat{R}_0(\hat{\Psi}_k) &= E_S(\mathbb{P}_V - P_0) D\{\hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} \\
&\quad + E_S P_0 \Big[D\{\hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} - D\{\hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), g_0, \hat{\Psi}_k(\mathbb{P}_T)\} \Big] \\
&= E_S(\mathbb{P}_V - P_0) D\{\hat{\bar{Q}}^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} \qquad \text{(B.1)} \\
&\quad + E_S P_0 \frac{g_0 - \hat{g}(\mathbb{P}_T)}{g_0 \hat{g}(\mathbb{P}_T)} h \Big[\{\bar{Q}_{2,0} - \hat{\bar{Q}}_2^*(\mathbb{P}_T)\} - 2\hat{\Psi}_k(\mathbb{P}_T)\{\bar{Q}_{1,0} - \hat{\bar{Q}}_1^*(\mathbb{P}_T)\} \Big]. \qquad \text{(B.2)}
\end{aligned}$$

We will first handle the term (B.2). By Cauchy-Schwartz, (B.2) can be bounded by

$$\begin{aligned}
&\sup_{(a,v)}|h(a,v)|\left|\left|\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)}\right|\right|_{0,S}||\hat{\bar{Q}}_2^*(P_0)-\hat{\bar{Q}}_2^*(\mathbb{P}_T)||_{0,S}\\
&\qquad\qquad+E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)}h\{\bar{Q}_{2,0}-\hat{\bar{Q}}_2^*(P_0)\}\\
&\quad-2\sup_{(s,a,v)}|h(a,v)\hat{\Psi}_k(\mathbb{P}_{T_S})(a,v)|\left|\left|\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)}\right|\right|_{0,S}||\hat{\bar{Q}}_1^*(P_0)-\hat{\bar{Q}}_1^*(\mathbb{P}_T)||_{0,S}\\
&\qquad\qquad\qquad-2E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)}h\hat{\Psi}_k(\mathbb{P}_T)\{\bar{Q}_{1,0}-\hat{\bar{Q}}_1^*(P_0)\}. \qquad \text{(B.3)}
\end{aligned}$$

Using a similar argument, the last term can be bounded (using assumption 1 and up to universal constants) by

$$\begin{aligned}
&||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}||\hat{\Psi}_k(\mathbb{P}_T)-\psi_0||_{0,S}+\\
&\qquad\qquad E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0^2}h\psi_0\{\bar{Q}_{1,0}-\hat{\bar{Q}}_1^*(P_0)\}+||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}^2,
\end{aligned}$$

whereas the second term in (B.3) is bounded by

$$E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0^2}h\{\bar{Q}_{2,0}-\hat{\bar{Q}}_2^*(P_0)\}+||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}^2.$$

Therefore

$$\begin{aligned}
\hat{R}(\hat{\Psi}_k)-\hat{R}_0(\hat{\Psi}_k)=&E_S(\mathbb{P}_V-P_0)D\{\hat{\bar{Q}}^*(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\hat{\Psi}_k(\mathbb{P}_T)\} \qquad \text{(B.4)}\\
&+E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0^2}h\left[\{\bar{Q}_{2,0}-\hat{\bar{Q}}_2^*(P_0)\}-2\psi_0\{\bar{Q}_{1,0}-\hat{\bar{Q}}_1^*(P_0)\}\right]\\
&+Rem_n,
\end{aligned}$$

where

$$\begin{aligned}
Rem_n\leq ||g_0-\hat{g}(\mathbb{P}_T)||_{0,S}\big\{&a||\bar{Q}_2^*(P_0)-\bar{Q}_2^*(\mathbb{P}_T)||_{0,S}-b||\bar{Q}_1^*(P_0)-\hat{\bar{Q}}_1^*(\mathbb{P}_T)||_{0,S}+\\
&c||\hat{\Psi}_k(\mathbb{P}_T)-\psi_0||_{0,S}+d||g_0-\hat{g}(\mathbb{P}_T)||_{0,S}\big\},
\end{aligned}$$

for constants a,b,c,d.

On the other hand, since $E_S\mathbb{P}_Vf=\mathbb{P}f$ when f does not depend on S, we may rewrite (B.4) as

$$\begin{aligned}
&E_S(\mathbb{P}_V-P_0)D\{\hat{\bar{Q}}^*(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\hat{\Psi}_k(\mathbb{P}_T)\}\\
&=E_S(\mathbb{P}_V-P_0)\left[D\{\hat{\bar{Q}}^*(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\hat{\Psi}_k(\mathbb{P}_T)\}-D\{\bar{Q}^*(P_0),Q_{W,0},g_0,\psi_0\}\right] \qquad \text{(B.5)}\\
&\quad+(\mathbb{P}-P_0)D\{\bar{Q}^*(P_0),Q_{W,0},g_0,\psi_0\}
\end{aligned}$$

Following similar arguments to those presented by Zheng and van der Laan, 2010, and using the assumptions of the theorem, it can be proven that (B.5) is $o_P(1/\sqrt{n})$, which implies

$$\begin{aligned} &E_S(\mathbb{P}_V - P_0)D\{\hat{\bar{Q}}(\mathbb{P}_T)(\hat{\varepsilon}), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} = \\ &\qquad\qquad (\mathbb{P} - P_0)D\{Q(P_0)(\varepsilon_0), \hat{g}(P_0), \psi_0\} + o_P(1/\sqrt{n}). \end{aligned}$$

This result, together with (B.4) and assumptions 2, 3 and 4, yields

$$\hat{R}(\hat{\Psi}_k) - \hat{R}_0(\hat{\Psi}_k) = (\mathbb{P} - P_0)\,[D\{Q(P_0), Q_{W,0}, g_0, \psi_0 + IC_g(P_0)] + o_P(1/\sqrt{n}).$$

□

B.2 Theorem 2

Before proceeding to prove Theorem 2, we will first present and prove the following useful theorem.

Theorem 3. *Define* $\hat{R}(\hat{\Psi}_k, \psi_0) \equiv \hat{R}(\hat{\Psi}_k) - \hat{R}(\psi_0)$ *and* $\hat{R}_0(\hat{\Psi}_k, \psi_0) \equiv \hat{R}_0(\hat{\Psi}_k) - R_0(\psi_0)$*, where*

$$\hat{R}(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), \hat{\bar{Q}}^*_k(\mathbb{P}_T), \mathbb{P}_V\}$$

is the TMLE of the true conditional risk

$$\hat{R}_0(\hat{\Psi}_k) = E_S R\{\hat{\Psi}_k(\mathbb{P}_T), Q_0\}.$$

Assume

$$(\hat{R} - \hat{R}_0)(\hat{\Psi}_k, \psi_0) = E_S(\mathbb{P}_V - P_0)D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k) + Rem_k$$

for some function D_k *that depends on* $(\mathbb{P}_T, P_0, \hat{\varepsilon}_k)$ *such that*

$$E_S P_0 D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k) = \hat{R}_0(\hat{\Psi}_k, \psi_0),$$

$||D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k)||_\infty < M_1 < \infty$*, and* $P_0\{D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k)\}^2 \leq M_2 P_0 D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k)$*. Assume also that* $\hat{\varepsilon}_k$ *falls in a finite set with maximally* n^c *points for some finite c, and denote* $M_n = n^c K_n$*. Then,*

$$\begin{aligned} &E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) \lesssim (1 + 2\delta)E\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0) + \\ &\qquad\qquad c(M_1, M_2, \delta)\frac{1 + \log M_n}{n} + (1 + \delta)\{ERem_{\tilde{k}} - ERem_{\hat{k}}\}, \end{aligned}$$

where $c(M_1, M_2, \delta) = (1 + \delta)^2(M_1/3 + M_2/\delta)$

*Proof o*f Theorem 3.

$$\begin{aligned}
0 \leq & \hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) \\
= & \hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) - (1+\delta)\hat{R}(\hat{\Psi}_{\hat{k}}, \psi_0) + (1+\delta)\hat{R}(\hat{\Psi}_{\hat{k}}, \psi_0) \\
\leq & \hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) - (1+\delta)\hat{R}(\hat{\Psi}_{\hat{k}}, \psi_0) + (1+\delta)\hat{R}(\hat{\Psi}_{\tilde{k}}, \psi_0) \\
= & \hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) \\
& + (1+\delta)\{\hat{R}(\Psi_{\tilde{k}}, \psi_0) - \hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0)\} \\
& - (1+\delta)\{\hat{R}(\Psi_{\hat{k}}, \psi_0) - \hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0)\} \\
& + (1+\delta)\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0) \\
& - (1+\delta)\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) \\
= & (1+2\delta)\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0) + H_{\hat{k}} + T_{\tilde{k}},
\end{aligned}$$

where

$$\begin{aligned}
H_k &= -(1+\delta)(\hat{R}-\hat{R}_0)(\hat{\Psi}_k, \psi_0) - \delta\hat{R}_0(\hat{\Psi}_k, \psi_0) \\
T_k &= \ (1+\delta)(\hat{R}-\hat{R}_0)(\hat{\Psi}_k, \psi_0) - \delta\hat{R}_0(\hat{\Psi}_k, \psi_0).
\end{aligned}$$

By using the assumptions of the theorem we get

$$\begin{aligned}
H_k = & -(1+\delta)E_S(P_V - P_0)D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k) - \delta E_S P_0 D_k(\mathbb{P}_T, P_0, \hat{\varepsilon}_k) - (1+\delta)Rem_k \\
\equiv & H_k^* - (1+\delta)Rem_k.
\end{aligned}$$

Following arguments similar to those presented by Dudoit and van der Laan, 2005 and van der Vaart, Dudoit, and van der Laan, 2006 we have that

$$EH_{\hat{k}}^* \lesssim c(M_1, M_2, \delta)\frac{1+\log M_n}{n}.$$

The same bound applies to $ET_{\tilde{k}}^*$. As a consequence we obtain the desired result. □

*Proof o*f Theorem 2. Recall from Theorem 1 that

$$\begin{aligned}
\hat{R}(\hat{\Psi}_k) - \hat{R}_0(\hat{\Psi}_k) = E_S(\mathbb{P}_V - P_0)D\{\hat{\bar{Q}}_k^*(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} + \\
2E_S P_0 \frac{g_0 - \hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)} h\hat{\Psi}_k(\mathbb{P}_T)\{\hat{\bar{Q}}_k^*(\mathbb{P}_T) - \bar{Q}_0\}, \quad \text{(B.6)}
\end{aligned}$$

where D is the efficient influence function

$$\begin{aligned}
D(Q, Q_W, g, \psi)(O) = -2\frac{h(A,Z)\psi(A,Z)}{g(A,W)}\{Y - \bar{Q}(A,W)\} + \\
\int_{\mathscr{A}} \psi(a,Z)\{\psi(a,Z) - 2\bar{Q}(a,W)\}h(a,Z)d\mu(a) - R(\psi)(\bar{Q}, Q_W).
\end{aligned}$$

Applying this same equality to the constant algorithm ψ_0 and subtracting it from (B.6) yields

$$(\hat{R}-\hat{R}_0)(\hat{\Psi}_k,\psi_0) = \tag{B.7}$$

$$E_S(\mathbb{P}_V-P_0)\left[D\{\hat{\bar{Q}}^*_k(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\hat{\Psi}_k(\mathbb{P}_T)\}-D\{\hat{\bar{Q}}^*_0(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\psi_0\}\right]$$

$$+2E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)}h\{\hat{\Psi}_k(\mathbb{P}_T)-\psi_0\}\{\hat{\bar{Q}}^*_0(\mathbb{P}_T)-\bar{Q}_0\} \tag{B.8}$$

$$+2E_SP_0\frac{g_0-\hat{g}(\mathbb{P}_T)}{g_0\hat{g}(\mathbb{P}_T)}h\hat{\Psi}_k(\mathbb{P}_T)\{\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\hat{\bar{Q}}^*_0(\mathbb{P}_T)\} \tag{B.9}$$

$$=E_S(\mathbb{P}_V-P_0)\left[D\{\hat{\bar{Q}}^*_k(\mathbb{P}_T),Q_W(\mathbb{P}_V),g_0,\hat{\Psi}_k(\mathbb{P}_T)\}-D\{\hat{\bar{Q}}^*_0(\mathbb{P}_T),Q_W(\mathbb{P}_V),g_0,\psi_0\}\right] \tag{B.10}$$

$$\begin{aligned}&+E_S(\mathbb{P}_V-P_0)\Bigg(\left[D\{\hat{\bar{Q}}^*_k(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\hat{\Psi}_k(\mathbb{P}_T)\}-D\{\hat{\bar{Q}}^*_k(\mathbb{P}_T),Q_W(\mathbb{P}_V),g_0,\hat{\Psi}_k(\mathbb{P}_T)\}\right]\\&\qquad-\left[D\{\hat{\bar{Q}}^*_0(\mathbb{P}_T),Q_W(\mathbb{P}_V),\hat{g}(\mathbb{P}_T),\psi_0\}-D\{\hat{\bar{Q}}^*_0(\mathbb{P}_T),Q_W(\mathbb{P}_V),g_0,\psi_0\}\right]\Bigg)\end{aligned} \tag{B.11}$$

$$+Rem_{k,1}+Rem_{k,2}$$

$$\equiv E_S(\mathbb{P}_V-P_0)D_k(\mathbb{P},P_0)+Rem_{k,1}+Rem_{k,2}+Rem_{k,3} \tag{B.12}$$

where $D_k(\mathbb{P},P_0)$ denotes the function inside square brackets in (B.10), and $Rem_{k,1}$, $Rem_{k,2}$ and $Rem_{k,3}$ denote (B.8), (B.9) and (B.11), respectively. From the definition of the efficient influence function D, note that $D\{\bar{Q},Q_W,g,\psi\}=L_{\bar{Q},g}(\psi)-R(\psi,\bar{Q},Q_W)$, which implies

$$\begin{aligned}E_S(\mathbb{P}_V-P_0)D_k(\mathbb{P},P_0)=E_S(\mathbb{P}_V-P_0)\left[L_{\hat{\bar{Q}}^*_k(\mathbb{P}_T),g_0}(\hat{\Psi}_k(\mathbb{P}_T))-L_{\hat{\bar{Q}}^*_0(\mathbb{P}_T),g_0}(\psi_0)\right]-&\\E_S(\mathbb{P}_V-P_0)\left[R\{\Psi(\mathbb{P}_T),\hat{\bar{Q}}^*_k(\mathbb{P}_T),Q_W(\mathbb{P}_V)\}-R\{\psi_0,\hat{\bar{Q}}^*_0(\mathbb{P}_T),Q_W(\mathbb{P}_V)\}\right]&,\end{aligned} \tag{B.13}$$

where the term inside square brackets in (B.13) is a constant, and (B.13) equals zero. Note that $\hat{\bar{Q}}^*_k(\mathbb{P}_T)\equiv\hat{\bar{Q}}(\mathbb{P}_T)(\hat{\varepsilon}_k)$ depends on $\mathbb{P}_V$ only through $\hat{\varepsilon}_k$, thus allowing us to rewrite

$$(\hat{R}-\hat{R}_0)(\hat{\Psi}_k,\psi_0)=E_S(\mathbb{P}_V-P_0)D_k(\mathbb{P}_T,P_0,\hat{\varepsilon}_k)+Rem_{k,1}+Rem_{k,2}+Rem_{k,3},$$

for

$$D_k(\mathbb{P}_T,P_0,\hat{\varepsilon}_k)\equiv L_{\hat{\bar{Q}}^*_k(\mathbb{P}_T),g_0}(\hat{\Psi}_k(\mathbb{P}_T))-L_{\hat{\bar{Q}}^*_0(\mathbb{P}_T),g_0}(\psi_0).$$

From the identity $P_0L_{\bar{Q},g_0}(\psi)=P_0L_{\bar{Q}_0}(\psi)$ ($L_{\bar{Q}}$ denotes the g-comp loss function) we have that

$$E_SP_0D_k(\mathbb{P}_T,P_0,\hat{\varepsilon}_k)=\hat{R}_0(\hat{\Psi}_k,\psi_0).$$

This fact together with (B.12) prove that D_k satisfies the conditions of Theorem 3, with $Rem_k=Rem_{1,k}+Rem_{2,k}+Rem_{3,k}$. By application of Theorem 3 we obtain

$$\begin{aligned}E\hat{R}_0(\hat{\Psi}_{\hat{k}},\psi_0)\lesssim(1+2\delta)E\hat{R}_0(\hat{\Psi}_{\tilde{k}},\psi_0)+&\\c(M_1,M_2,\delta)\frac{1+\log M_n}{n}+(1+\delta)\{ERem_{\tilde{k}}-ERem_{\hat{k}}\}&.\end{aligned}$$

It remains to study $ERem_k$. Let us first consider $ERem_{1,k}$. Note that

$$||(\hat{\Psi}_k(\mathbb{P}_T)-\psi_0)/\sqrt{g_0}||^2_{0,S}=\hat{R}_0(\hat{\Psi}_k,\psi_0)$$

Since g_0 and $\hat{g}(\mathbb{P}_T)$ are assumed bounded away from zero (positivity assumption), we can apply the Cauchy-Schwartz inequality to obtain

$$ERem_{1,k}\lesssim E||(\hat{g}(\mathbb{P}_T)-g_0)(\hat{\bar{Q}}^*_0(\mathbb{P}_T)-\bar{Q}_0)||_{0,S}\sqrt{R_0(\hat{\Psi}_k,\psi_0)}.$$

We now consider $ERem_{2,k}$. By applying the Cauchy-Schwartz inequality we obtain

$$Rem_{2,k}\lesssim ||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}||\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\hat{\bar{Q}}^*_0(\mathbb{P}_T)||_{0,S}. \tag{B.14}$$

From the definition of $\hat{\varepsilon}_k$ in the CV-TMLE of $R(\hat{\Psi}_k)$, note that $\hat{\bar{Q}}^*_k$ satisfies the equation

$$E_S\mathbb{P}_V\frac{h}{\hat{g}(\mathbb{P}_T)}\hat{\Psi}_k(\mathbb{P}_T)\{Y-\hat{\bar{Q}}^*_k(\mathbb{P}_T)\}=0.$$

Applying the same equation for ψ_0 and $\hat{\bar{Q}}^*_0$, and subtracting it from the previous one yields

$$E_S\mathbb{P}_V\frac{h}{\hat{g}(\mathbb{P}_T)}\left(\hat{\Psi}_k(\mathbb{P}_T)\{Y-\hat{\bar{Q}}^*_k(\mathbb{P}_T)\}-\psi_0\{Y-\hat{\bar{Q}}^*_0(\mathbb{P}_T)\}\right)=0,$$

which can be written as

$$E_S\mathbb{P}_V\frac{h}{\hat{g}(\mathbb{P}_T)}\{\hat{\Psi}_k(\mathbb{P}_T)-\psi_0\}\{Y-\hat{\bar{Q}}^*_k(\mathbb{P}_T)\}=E_S\mathbb{P}_V\frac{h}{\hat{g}(\mathbb{P}_T)}\psi_0\{\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\hat{\bar{Q}}^*_0(\mathbb{P}_T)\},$$

which implies

$$\begin{aligned}E_SP_0\frac{h}{\hat{g}(\mathbb{P}_T)}\psi_0\{\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\hat{\bar{Q}}^*_0(\mathbb{P}_T)\}=&E_S(\mathbb{P}_V-P_0)\frac{h}{\hat{g}(\mathbb{P}_T)}\{\hat{\Psi}_k(\mathbb{P}_T)-\psi_0\}\{Y-\hat{\bar{Q}}^*_k(\mathbb{P}_T)\}\\&-E_S(\mathbb{P}_V-P_0)\frac{h}{\hat{g}(\mathbb{P}_T)}\psi_0\{\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\hat{\bar{Q}}^*_0(\mathbb{P}_T)\}\\&-E_SP_0\frac{h}{\hat{g}(\mathbb{P}_T)}\{\hat{\Psi}_k(\mathbb{P}_T)-\psi_0\}\{\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\bar{Q}_0\}.\end{aligned}$$

By empirical process theory (van der Vaart and Wellner, 1996, Theorem 2.14.1), noting that the first two terms are empirical processes applied to functions in a class of functions $\mathscr{F}=\{f(k,\varepsilon_k,\varepsilon_0):k,\varepsilon_k,\varepsilon_0\}$, the expectations of the first two terms are bounded by $(1+\log M_n)/\sqrt{n}$. The third term is bounded by $\sqrt{R_0(\hat{\Psi}_k,\psi_0)}||\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\bar{Q}_0||_{0,S}$. These facts together with (B.14) yield

$$\begin{aligned}ERem_{2,k}\lesssim E||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}\frac{1+\log M_n}{\sqrt{n}}+&\\&E||(\hat{g}(\mathbb{P}_T)-g_0)||_{0,S}||(\hat{\bar{Q}}^*_k(\mathbb{P}_T)-\bar{Q}_0)||_{0,S}\sqrt{R_0(\hat{\Psi}_k,\psi_0)},\end{aligned}$$

Finally, consider the term $Rem_{3,k}$. We can bound this term by $\max_{k,\varepsilon_k,\varepsilon_0} E_S(\mathbb{P}_V - P_0)D(\mathbb{P}_T, k, \varepsilon_k, \varepsilon_0)$, for

$$\begin{aligned} D(\mathbb{P}_T,k,\varepsilon_k,\varepsilon_0) =& D\{\hat{\bar{Q}}^*_k(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \hat{\Psi}_k(\mathbb{P}_T)\} - D\{\hat{\bar{Q}}^*_k(\mathbb{P}_T), Q_W(\mathbb{P}_V), g_0, \hat{\Psi}_k(\mathbb{P}_T)\} \\ & - D\{\hat{\bar{Q}}^*_0(\mathbb{P}_T), Q_W(\mathbb{P}_V), \hat{g}(\mathbb{P}_T), \psi_0\} + D\{\hat{\bar{Q}}^*_0(\mathbb{P}_T), Q_W(\mathbb{P}_V), g_0, \psi_0\} \\ =& 2h\frac{\hat{g}(\mathbb{P}_T) - g_0}{\hat{g}(\mathbb{P}_T)g_0}\{\hat{\Psi}_k(\mathbb{P}_T)(Y - \hat{\bar{Q}}^*_k(\mathbb{P}_T)) - \psi_0(Y - \hat{\bar{Q}}^*_0(\mathbb{P}_T))\} \end{aligned}$$

Let $F(\mathbb{P}_T)$ be the envelope of the class of functions $\mathscr{F}(\mathbb{P}_T) = \{D(\mathbb{P}_T, k, \varepsilon_k, \varepsilon_0) : k, \varepsilon_k, \varepsilon_0\}$, over which we take the maximum. We have $P_0F(\mathbb{P}_T)^2 \lesssim ||\hat{g}(\mathbb{P}_T) - g_0||^2_{0,S}$. We will apply the following inequality for empirical processes indexed by a finite class of functions $\mathscr{F}$:

$$E\max_{f\in\mathscr{F}}|(\mathbb{P} - P_0)f| \lesssim \frac{1}{\sqrt{n}}\sqrt{\log(\#\mathscr{F})}||F||_2,$$

where F is an envelope of $\mathscr{F}$. Thus, given $\mathbb{P}_T$, we can bound the conditional expectation of $Rem_{3,k}$ by $(1+\log M_n)||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}/\sqrt{n}$, which results in the following bound for the marginal expectation:

$$ERem_{3,k} \lesssim \frac{1+\log M_n}{\sqrt{n}}E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}.$$

Accumulation of these bounds for the different components of $Rem_{\hat{k}}$ and $\mathrm{Rem}_{\tilde{k}}$ yields the following inequality:

$$\begin{aligned} E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0) \lesssim & (1+2\delta)E\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0) \\ & + c(M,\delta)\frac{1+\log(K_n)}{n} \\ & + (1+\delta)E||(\hat{g}(\mathbb{P}_T) - g_0)(\hat{\bar{Q}}^*_0(\mathbb{P}_T) - \bar{Q}_0)||_{0,S}\sqrt{E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0)} \\ & + (1+\delta)E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}E||\hat{\bar{Q}}^*_{\hat{k}}(\mathbb{P}_T) - \bar{Q}_0||_{0,S}\sqrt{E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0)} \\ & + (1+\delta)E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}E||\hat{\bar{Q}}^*_{\tilde{k}}(\mathbb{P}_T) - \bar{Q}_0||_{0,S}\sqrt{E\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0)} \\ & + (1+\delta)\frac{1+\log M_n}{\sqrt{n}}E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S} \\ \lesssim & (1+2\delta)E\hat{R}_0(\hat{\Psi}_{\tilde{k}}, \psi_0) \\ & + c(M,\delta)\frac{1+\log(K_n)}{n} \\ & + (1+\delta)E||(\hat{g}(\mathbb{P}_T) - g_0)(\hat{\bar{Q}}^*_0(\mathbb{P}_T) - \bar{Q}_0)||_{0,S}\sqrt{E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0)} \\ & + (1+\delta)E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}E||\hat{\bar{Q}}^*_{\bar{k}}(\mathbb{P}_T) - \bar{Q}_0||_{0,S}\sqrt{E\hat{R}_0(\hat{\Psi}_{\hat{k}}, \psi_0)} \\ & + (1+\delta)E||\hat{g}(\mathbb{P}_T) - g_0||_{0,S}\frac{1+\log M(n)}{\sqrt{n}}, \end{aligned}$$

where $\bar{k}$ is either $\hat{k}$ or $\tilde{k}$, whichever gives the worst bound. This inequality can be written as $x^2 - bx \lesssim c$, for

$$\begin{aligned} b =&(1+\delta)E||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}||\hat{\bar{Q}}^*_{\bar{k}}(\mathbb{P}_T)-\bar{Q}_0||_{0,S}+ \\ &(1+\delta)E||(\hat{g}(\mathbb{P}_T)-g_0)(\hat{\bar{Q}}^*_0(\mathbb{P}_T)-\bar{Q}_0)||_{0,S} \\ c =&c(M,\delta)\frac{1+\log M_n}{\sqrt{n}}+(1+\delta)E||\hat{g}(\mathbb{P}_T)-g_0||_{0,S}\frac{1+\log M_n}{\sqrt{n}} \end{aligned}$$

and can be solved using the quadratic formula as $x \leq (b+\sqrt{b^2+4c})/2$, which in turn implies $x \leq b+\sqrt{c}$, proving Theorem 2. □

Bibliography

Bembom, O. and M. J. van der Laan (2007). “A practical illustration of the importance of realistic individualized treatment rules in causal inference”. In: *Electronic Journal of Statistics*. (Cit. on pp. 13, 17, 27, 28, 32).

Bickel, P. J. et al. (1997). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag. (Cit. on pp. 20, 35, 36, 38, 48, 51, 68, 76).

Breiman, L. (2001). “Random Forests”. In: *Machine Learning* 45 (1), pp. 5–32. ISSN: 0885-6125. (Cit. on p. 67).

Brotman, R. M. et al. (2008). “A Longitudinal Study of Vaginal Douching and Bacterial VaginosisA Marginal Structural Modeling Analysis”. In: *American Journal of Epidemiology* 168.2, pp. 188–196. DOI: 10.1093/aje/kwn103. eprint: http://aje.oxfordjournals.org/content/168/2/188.full.pdf+html. URL: http://aje.oxfordjournals.org/content/168/2/188.abstract. (Cit. on p. 39).

Bryan, J., Z. Yu, and M. J. van der Laan (2003). “Analysis of Longitudinal Marginal Structural Models”. In: *Biostatistics* 5.3, pp. 361–380. (Cit. on p. 6).

— (2004). “Analysis of longitudinal marginal structural models”. In: *Biostatistics* 5.3, pp. 361–380. DOI: 10.1093/biostatistics/kxg041. eprint: http://biostatistics.oxfordjournals.org/content/5/3/361.full.pdf+html. URL: http://biostatistics.oxfordjournals.org/content/5/3/361.abstract. (Cit. on p. 39).

Buchman, T. G. (2010). “Novel representation of physiologic states during critical illness and recovery”. In: *Crit. Care* 14, p. 127. (Cit. on p. 69).

Cain, L. E. et al. (2010). “When to Start Treatment? A Systematic Approach to the Comparison of Dynamic Regimes Using Observational Data”. In: *The International Journal of Biostatistics* 6. URL: http://www.bepress.com/ijb/vol6/iss2/18. (Cit. on pp. 18, 34).

Dawid, A. P. and V. Didelez (2010). “Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview”. In: *CoRR* abs/1010.3425. (Cit. on pp. 34, 37, 48, 74).

Denby, L. and C. Mallows (2009). “Variations on the Histogram”. In: *Journal of Computational and Graphical Statistics* Vol. 18, Iss. 1, pp. 21–31. (Cit. on p. 5).

Díaz, I and M. J. van der Laan (2011a). “Population Intervention Causal Effects Based on Stochastic Interventions”. In: *Biometrics*, In press. ISSN: 1541-0420. DOI: 10.1111/j.1541-0420.2011.01685.x. URL: http://dx.doi.org/10.1111/j.1541-0420.2011.01685.x. (Cit. on pp. 1, 13, 16, 74).

Díaz, I. and M. J. van der Laan (2011b). "Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models". In: *The International Journal of Biostatistics* 7.1, p. 38. (Cit. on p. 80).

Didelez, Vanessa, A. Philip Dawid, and Sara Geneletti (2006). "Direct and Indirect Effects of Sequential Treatments". In: *UAI*. (Cit. on pp. 34, 48, 74).

Dudoit, S. and M. J. van der Laan (2005). "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment". In: *Statistical Methodology* 2.2, pp. 131–154. (Cit. on pp. 49, 96).

Eberhardt, F. and R. Scheines (2006). "Interventions and Causal Inference". In: *Department of Philosophy. Paper 415*. URL: `http://repository.cmu.edu/philosophy/415`. (Cit. on p. 18).

Ghalanos, A. and S. Theussl (2010). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package*. (Cit. on p. 6).

Hafeman, D. M. and T. J. VanderWeele (2011). "Alternative Assumptions for the Identification of Direct and Indirect Effects". In: *Epidemiology* 22.6. (Cit. on p. 34).

Hess, J. R., J. B. Holcomb, and D. B. Hoyt (2006). "Damage control resuscitation: the need for specific blood products to treat the coagulopathy of trauma". In: *Transfusion* 46 (5), pp. 685–686. (Cit. on p. 68).

Holcomb, J. B. et al. (2007). "Causes of death in US Special Operations Forces in the global war on terrorism: 2001–2004". In: *Annals of surgery* 245.6, p. 986. (Cit. on p. 68).

Hubbard, A. E. and M. J. van der Laan (2005). *Population Intervention Models in Causal Inference*. Technical Report 191. Division of Biostatistics, University of California, Berkeley. (Cit. on p. 17).

Ishwaran, H. (2007). "Variable importance in binary regression trees and forests". In: *Electronic Journal of Statistics*, pp. 519–537. (Cit. on p. 67).

Joffe, M. M. et al. (2004). "Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications". English. In: *The American Statistician* 58.4, pp. 272–279. ISSN: 00031305. URL: `http://www.jstor.org/stable/27643582`. (Cit. on p. 39).

Korb, K. et al. (2004). "Varieties of Causal Intervention". In: *PRICAI 2004: Trends in Artificial Intelligence*. Ed. by Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap. Vol. 3157. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 322–331. (Cit. on pp. 18, 33, 48, 74).

Krumrei, N. J. et al. (2012). "Comparison of massive blood transfusion predictive models in the rural setting". In: *The Journal of Trauma and Acute Care Surgery* 72.1, p. 211. (Cit. on p. 69).

Lesko, M. M. et al. (2012). "Comparing Model Performance for Survival Prediction Using Total GCS and Its Components in Traumatic Brain Injury". In: *Journal of Neurotrauma* ja. (Cit. on p. 69).

MacFadden, L. N. et al. (2012). "A model for predicting primary blast lung injury". In: *The Journal of Trauma and Acute Care Surgery*. (Cit. on p. 69).

Mann, J. K. et al. (June 2010). "Short-Term Effects of Air Pollution on Wheeze in Asthmatic Children in Fresno, California". In: *Environ Health Perspect* 118.10. DOI: `10.1289/ehp.0901292`. URL: `http://dx.doi.org/10.1289%2Fehp.0901292`. (Cit. on p. 45).

McAlister, A. L. (1991). “Population Behavior Change: A Theory-Based Approach”. English. In: *Journal of Public Health Policy* 12.3, pp. 345–361. ISSN: 01975897. URL: `http://www.jstor.org/stable/3342846`. (Cit. on p. 33).

Neugebauer, R. and M. J. van der Laan (2007). “Nonparametric causal effects based on marginal structural models”. In: *Journal of Statistical Planning and Inference* 137.2, pp. 419 –434. ISSN: 0378-3758. DOI: `DOI:10.1016/j.jspi.2005.12.008`. URL: `http://www.sciencedirect.com/science/article/pii/S0378375806000334`. (Cit. on pp. 1, 2, 6, 22, 32, 34, 38, 43, 49, 63).

Nuñez, T. C. et al. (2009). “Early prediction of massive transfusion in trauma: simple as ABC (assessment of blood consumption)?” In: *The Journal of Trauma and Acute Care Surgery* 66.2, pp. 346–352. (Cit. on p. 69).

Olden, J. D. and D. A. Jackson (2002). “Illuminating “the black box”: a randomization approach for understanding variable contributions in artificial neural networks”. In: *Ecological Modelling* 154, pp. 135 –150. ISSN: 0304-3800. (Cit. on p. 67).

Olden, J. D., M. K. Joy, and R. G. Death (2004). “An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data”. In: *Ecological Modelling* 178, pp. 389 –397. ISSN: 0304-3800. (Cit. on p. 67).

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. (Cit. on pp. 6, 33, 36, 37, 48, 70–72, 74).

— (2001). *Cognitive systems Laboratory*. Tech. rep. University of California, Los Angeles, Department of Computer Science. (Cit. on p. 34).

— (2009). “Causal inference in statistics: An overview”. In: *Statistics Surveys*, p. 350. (Cit. on p. 33).

Petersen, M. L. et al. (2010). “Diagnosing and responding to violations in the positivity assumption.” In: *Stat Methods Med Res*. ISSN: 1477-0334. URL: `http://www.biomedsearch.com/nih/Diagnosing-responding-to-violations-in/21030422.html`. (Cit. on pp. 22, 39).

Porter, K. E. et al. (2011). “The Relative Performance of Targeted Maximum Likelihood Estimators”. In: *The International Journal of Biostatistics* 7.1, pp. 1–34. (Cit. on p. 40).

R. Mansson M.M. Joffe, W. Sun S. Hennessy (2007). “On the Estimation and Use of Propensity Scores in Case-Control and Case-Cohort Studies.” In: *Am J Epidemiol* 166.3, pp. 332–339. (Cit. on pp. 1, 48).

Robins, J. and T. Richardson (2010). *Alternative Graphical Causal Models and the Identification of Direct Effects*. Working Paper 100. Harvard School of Public Health. (Cit. on p. 34).

Robins, J. M. (1986). “A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect”. In: *Mathematical Modelling* 7, pp. 1393–1512. (Cit. on pp. 37, 48, 77).

Robins, J. M. and S. Greenland (1992). “Identifiability and exchangeability for direct and indirect effects”. In: *Epidemiology* 3.0, pp. 143–155. (Cit. on p. 34).

Robins, J. M., M. A. Hernan, and B. Brumback (2000). “Marginal structural models and causal inference in epidemiology”. In: *Epidemiology* 11.5, pp. 550–560. (Cit. on pp. 1, 6, 49, 63).

Rose, S. and M. J. van der Laan (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer. (Cit. on pp. 21, 35, 40, 42, 46, 48, 53–55, 68, 77).

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70, pp. 41–55. (Cit. on pp. 1, 48).

Rosenblatt, M. (1969). "Conditional probability density and regression estimates". In: *Multivariate Analysis II, Ed. P.R. Krishnaiah* 22, pp. 25–31. (Cit. on p. 2).

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology*. URL: `http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ118470`. (Cit. on pp. 33, 48).

— (1978). "Bayesian Inference for causal effects: the role of randomization". In: *Annals of Statistics* 6, pp. 34–58. (Cit. on pp. 33, 48, 74).

Rubin, D.B. (2006). *Matched Sampling for Causal Effects*. Cambridge, MA: Cambridge University Press. (Cit. on p. 48).

Schöchl, H. et al. (2011). "FIBTEM provides early prediction of massive transfusion in trauma". In: *Crit care* 15.6, R265. (Cit. on p. 69).

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley. ISBN: 9780471547709. (Cit. on p. 2).

Stitelman, O. M., A. E. Hubbard, and N. P. Jewell (2010a). "The Impact Of Coarsening The Explanatory Variable Of Interest In Making Causal Inferencegs: Implicit Assumptions Behind Dichotomizing Variables". In: *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 264. (Cit. on p. 67).

— (2010b). "The Impact Of Coarsening The Explanatory Variable Of Interest In Making Causal Inferences: Implicit Assumptions Behind Dichotomizing Variables". In: URL: `http://www.bepress.com/ucbbiostat/paper264`. (Cit. on pp. 18, 34, 39).

Strobl, C. et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: (cit. on p. 67).

Tager, I. B. et al. (2004). "Effects of Physical Activity and Body Composition on Functional Limitation in the Elderly: Application of the Marginal Structural Model". English. In: *Epidemiology* 15.4, pp. 479–493. ISSN: 10443983. URL: `http://www.jstor.org/stable/20485932`. (Cit. on p. 39).

Tager, I.B., M. Hollenberg, and W. Satariano (1998). "Self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population". In: (cit. on pp. 13, 17, 27).

Taubman, S. L. et al. (2009a). "Intervening on risk factors for coronary heart disease: an application of the parametric g-formula". In: *International Journal of Epidemiology* 38.6, pp. 1599–1611. DOI: `10.1093/ije/dyp192`. eprint: `http://ije.oxfordjournals.org/content/38/6/1599.full.pdf+html`. URL: `http://ije.oxfordjournals.org/content/38/6/1599.abstract`. (Cit. on p. 18).

Taubman, S. L et al. (2009b). "Intervening on risk factors for coronary heart disease: an application of the parametric g-formula". In: *International Journal of Epidemiology* 38.6, pp. 1599–1611. DOI: `10.1093/ije/dyp192`. eprint: `http://ije.oxfordjournals.org/content/38/6/`

1599.full.pdf+html. URL: http://ije.oxfordjournals.org/content/38/6/1599.abstract. (Cit. on p. 34).

Tian, J. (2008). "Identifying Dynamic Sequential Plans". In: *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*. Corvallis, Oregon: AUAI Press, pp. 554–561. (Cit. on p. 34).

Tsiatis, A.A. (2006). "Information based monitoring of clinical trials". In: *Statistics in Medicine*. (Cit. on p. 35).

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN: 9780521650021. (Cit. on p. 77).

van der Laan, M. J. and S. Dudoit (2003). *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. Division of Biostatistics, University of California, Berkeley. (Cit. on pp. 2, 23, 41, 49, 52, 55–57, 59, 68, 80).

van der Laan, M. J., S. Dudoit, and S. Keles (2004). "Asymptotic optimality of likelihood-based cross-validation". In: *Statistical Applications in Genetics and Molecular Biology* 3. (Cit. on pp. 2, 6, 23, 41, 49, 55, 59, 68, 80).

van der Laan, M. J., S. Dudoit, and A. W. van der Vaart (2006). "The Cross-Validated Adaptive Epsilon-Net Estimator". In: *Statistics and Decisions* 24.3, pp. 373–395. (Cit. on p. 49).

van der Laan, M. J. and M. L. Petersen (2012). "Targeted Learning". In: *Ensemble Machine Learning*. Ed. by Cha Zhang and Yunqian Ma. Springer US, pp. 117–156. ISBN: 978-1-4419-9326-7. (Cit. on p. 49).

van der Laan, M. J., E. Polley, and A. Hubbard (2007). "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.25. ISSN: 1. (Cit. on pp. 2, 3, 14, 23, 30, 41, 49, 59, 68, 69, 80).

van der Laan, M. J. and J.M. Robins (2003). *Unified methods for censored longitudinal data and causality*. Springer, New York. (Cit. on pp. 8, 9, 24, 35, 36, 41, 48, 53, 54, 68, 77, 79).

van der Laan, M. J. and D. Rubin (2006). "Targeted Maximum Likelihood Learning". In: *The International Journal of Biostatistics* 2.1. (Cit. on pp. 10, 22, 24, 42, 53, 77).

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, p. 62. (Cit. on pp. 1, 2).

— (2002). "Semiparameric Statistics". In: *Lectures on Probability Theory and Statistics*. Vol. 1781. Lecture Notes in Mathematics. Springer Berlin / Heidelberg, pp. 331–457. (Cit. on pp. 53, 54, 76).

— (2003). *Notes on Cross-validation*. Tech. rep. Department of Mathematics, Free University, Amsterdam. (Cit. on pp. 49, 55).

van der Vaart, A. W., S. Dudoit, and M. J. van der Laan (2006). "Oracle Inequalities for Multi-Fold Cross-Validation". In: *Statistics and Decisions* 24.3, pp. 351–371. (Cit. on pp. 2, 49, 96).

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Emprical Processes*. Springer-Verlag New York, p. 102. (Cit. on pp. 93, 98).

Wand, M.P. and M. C. Jones (1995). *Kernel smoothing*. Monographs on statistics and applied probability. Chapman & Hall. ISBN: 9780412552700. (Cit. on p. 2).

Wang, Y., O. Bembom, and M. J. van der Laan (2006). “Data adaptive estimation of the treatment specific mean”. In: *Journal of Statistical Planning and Inference*. (Cit. on pp. 51, 52, 56).

Yinyu, Y. (1987). “Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming”. PhD thesis. Department of EES, Stanford University. (Cit. on p. 6).

Young, J. et al. (2011). “Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric G-Formula”. In: *Statistics in Biosciences* 3 (1). 10.1007/s12561-011-9040-7, pp. 119–143. ISSN: 1867-1764. URL: `http://dx.doi.org/10.1007/s12561-011-9040-7`. (Cit. on p. 35).

Zheng, W. and M. J. van der Laan (2010). “Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation”. In: *Division of Biostatistics Working Paper Series*. (Cit. on pp. 56, 57, 93, 95).

— (2011a). “Cross-Validated Targeted Minimum-Loss-Based Estimation”. In: *Targeted Learning*. Ed. by P. J. Bickel et al. Springer Series in Statistics. Springer New York, pp. 459–474. ISBN: 978-1-4419-9782-1. (Cit. on p. 55).

— (2011b). *Targeted Maximum Likelihood Estimation of Natural Direct Effect*. Working Paper 288 http://www.bepress.com/ucbbiostat/paper288. U.C. Berkeley Division of Biostatistics Working Paper Series. (Cit. on p. 34).