University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 159

Data Adaptive Estimation of the Treatment Specific Mean

Yue Wang* Oliver Bembom†

Mark J. van der Laan‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, ywang@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, bembom@gmail.com

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

Data Adaptive Estimation of the Treatment Specific Mean

Yue Wang, Oliver Bembom, and Mark J. van der Laan

Abstract

An important problem in epidemiology and medical research is the estimation of the causal effect of a treatment action at a single point in time on the mean of an outcome, possibly within strata of the target population defined by a subset of the baseline covariates. Current approaches to this problem are based on marginal structural models, i.e., parametric models for the marginal distribution of counterfactural outcomes as a function of treatment and effect modifiers. The various estimators developed in this context furthermore each depend on a high-dimensional nuisance parameter whose estimation currently also relies on parametric models. Since misspecification of any of these models can lead to severely baised estimates of causal effects, the dependence of current methods on such parametric models represents a major limitation. In this article we introduce estimators that allow the marginal structural model as well as the parametric model for the relevant nuisance parameter to be selected data-adaptively. Our methodology is based on the unified loss-based estimation approach recently developed by van der Laan and Dudoit (2003) that in particular extends loss-based estimation to missing data problems. We study the practical performance of our proposed estimators in an extensive simulation study and also apply them to data derived from an epidemiologic study to assess the causal effect of forced expiratory volume on mortality in the elderly. All of the estimators presented in this article are made publicly available in the R package cvDSA.

1 Introduction

Epidemiologists and medical researchers frequently wish to estimate the causal effect of a treatment action A at a single time point on the mean of a subsequently measured outcome Y. Often it is desirable to estimate this causal effect within strata of the target population defined by a subset V of the baseline covariates W. For example, a researcher may be interested in estimating the causal effect of administering a certain vaccine on the subsequent risk of infection with the disease of interest within different age groups.

Such causal effects are defined through the notion of a counterfactual outcome $Y(a)$ that we would have measured had the subject, possibly contrary to the fact, followed treatment a (Rubin, 1978). Causal inference is now based on a hypothetical full data structure X, that contains for each subject the entire collection of counterfactual outcomes $Y(a)$ as a ranges over the set of possible treatment actions $\mathcal{A}$: $X = ((Y(a) : a \in \mathcal{A}), W)$. Given this full data structure, the average marginal causal effect of A on Y within strata defined by V can now be defined as the effect of a on $E[Y(a) \mid V]$. The observed data O, of course, only contain the single counterfactual outcome $Y(A)$ that corresponds to the treatment the subject actually followed: $O = (W, A, Y \equiv Y(A))$. Within the counterfactual framework, causal inference can thus be viewed as a missing data problem, with A playing the role of a missingness variable.

Current methods for estimation of the parameter $E[Y(a) \mid V]$ are based on parametric models, referred to in this context also as marginal structural models (Robins, 2000a,b; van der Laan and Robins, 2002). The resulting causal analyses are thus based on the assumption that $E[Y(a) \mid V] = m(a, V \mid \beta)$ for some value of β, where $m(a, V \mid \beta)$ is a particular functional form that is specified *a priori* by the researcher. A simple choice, for example, might be a linear model including a one-way interaction term: $E[Y(a) \mid V] = \beta_0 + \beta_1 a + \beta_2 V + \beta_3 aV$.

Unfortunately, it is generally unrealistic to arrive at an appropriate parametric model for $E[Y(a) \mid V]$ on the basis of subject-matter knowledge alone. Since model misspecification can lead to severely biased estimates of the causal effect of A on Y, the dependence of current methods on such parametric models represents a major limitation.

In other regression problems, a number of model selection techniques have been studied that are aimed at selecting the appropriate functional form data-adaptively. One approach to this problem relies on optimizing a penalized version of the likelihood function, such as Akaike's Information Criterion (Akaike, 1973) or the Bayesian Information Criterion (Schwarz, 1978). These model selection techniques fail if estimation is not based on the likelihood principle, as for example in the context of many semi-parametric or non-parametric models. An attractive alternative in these situations is based on minimizing the expectation of an appropriately specified loss function that measures the performance of a candidate model at a particular observation. For the purpose of estimating $E[Y(a) \mid V]$, for example, one might use a squared-error loss function for repeated measures, $L(X, m) = \sum_{a \in \mathcal{A}} (Y(a) - m(a, V \mid \beta))^2 h(a, V)$, where $h(a, V)$ denotes a particular choice of weight function.

This approach appears to be infeasible if the observed data represent only a coarsened

version of some hypothetical full data structure that is required to define the parameter of interest. In such instances, loss functions appropriate for estimating the parameter of interest are functions of that full data structure, but not functions of the observed data. In our example, we do not observe $Y(a)$ for all possible treatments a so that the loss function $L(X, m)$ cannot be evaluated using only the observed data.

van der Laan and Dudoit (2003), however, recently demonstrated how full data loss functions can be mapped into corresponding observed data loss functions with the same expectation, making it possible to apply the general loss-based estimation methodology in the context of missing data problems. We here apply this approach to the data-adaptive estimation of $E[Y(a) \mid V]$ in the setting of a point treatment study, a problem that has previously not been addressed in the causal inference literature.

Current methods in this area in fact rely not only on a correctly specified marginal structural model for $E[Y(a) \mid V]$, but also on a correctly specified parametric model for a high-dimensional nuisance parameter that is involved in the estimation process. For Inverse-Probability-of-Treatment-Weighted (IPTW) estimators, this nuisance parameter consists of the probability of following a given treatment a given the collection of baseline covariates W; G-computation estimators rely on a regression of the outcome Y on baseline covariates W and treatment A; double robust estimators, finally, involve both of these nuisance parameters (Robins, 2000a,b; van der Laan and Robins, 2002). In contrast, the methodology we present here does not rely on correctly specified parametric models for these nuisance parameters, but rather applies the general loss-based estimation approach described above to select an appropriate model data-adaptively.

Another limitation of the methods currently in use is the lack of a publicly available implementation. While IPTW estimators and G-computation estimators are fairly straightforward to implement from scratch in a point treatment setting, the double robust estimator is involved enough to deter a number of potential users. Along with this article, we introduce the publicly available `R` package `cvDSA` that implements both the standard parametric approach as well as our novel data-adaptive methodology. We hope that this `R` package makes these general causal inference tools available to a broader class of epidemiologists and medical researchers.

The remainder of this article is organized as follows: We first formally define our data structure, model assumptions, and parameter of interest. The following section then describes in detail how we apply the general loss-based estimation road map laid out in van der Laan and Dudoit (2003) to this particular problem. In section 4 we present a simulation study aimed at comparing the behavior of the resulting three estimators. Our methodology is then applied to a data set derived from an epidemiologic study in an attempt to study the causal effect of forced expiratory volume on mortality in the elderly. We conclude with a brief discussion of the benefits of as well as potential extensions to the data-adaptive methods we develop in this article.

2 A nonparametric marginal structural model

In this section, we formally define our data structure, model assumptions, and parameter of interest. The hypothetical full data structure $X = ((Y(a) : a \in \mathcal{A}), W) \sim F_{X_0}$ contains for each subject baseline covariates W as well as the entire collection of counterfactual outcomes $Y(a)$ as a ranges over the set of possible treatment actions $\mathcal{A}$. We denote the conditional probability distribution of the treatment variable A by $g_0(a \mid X) \equiv Pr(A = a \mid X), a \in \mathcal{A}$. The observed data now consist of n i.i.d. observations of $O = (W, A, Y \equiv Y(A))$, containing only the single counterfactual outcome $Y(A)$ that corresponds to the treatment the subject actually followed. Hence O represents a coarsened version of the full data structure X, with A playing the role of a missingness variable. Consequently, we can parameterize the distribution of the observed data by the full data distribution F_{X_0} and the treatment mechanism g_0: $P_0 = P_{F_{X_0},g_0}$.

We make no assumptions about the full data distribution F_{X_0} and denote the corresponding non-parametric full data model by $\mathcal{M}^F$. The treatment mechanism is left unspecified as well except for the following two standard assumptions which are necessary to make the parameter of interest identifiable. First, the randomization assumption (RA) requires that treatment is randomized within strata of W:

$$g_0(a \mid X) = g_0(a \mid W) \text{ for all } a \in \mathcal{A}.$$

This requirement is equivalent with assuming that the missingness mechanism satisfies coarsening at random (van der Laan and Robins, 2002). Second, the experimental treatment assignment assumption (ETA) requires that within each stratum of W each possible treatment a is assigned with positive probability:

$$\inf_{a \in \mathcal{A}} g_0(a \mid W) > 0, \ F_{0W}\text{-a.e.}$$

In our case, it will be sufficient to rely on a weaker V-specific version of this assumption according to which there exists a conditional density $g(\cdot \mid V)$ such that

$$\sup_{a \in \mathcal{A}} \frac{g(a \mid V)}{g(a \mid W)} < \infty, \ F_{0W}\text{-a.e.}$$

Under these two assumptions on the treatment mechanism, the density of O factorizes as

$$p(O) = p(W)p(Y \mid A, W)g(A \mid W) = p(W)\, p(Y(a) \mid W)|_{a=A}\, g(A \mid X)$$

where the first two terms represent the F_X-part of the density of O and the last term corresponds to the treatment mechanism.

Let $\mathcal{G}(RA)$ be the model for the treatment mechanism $g(A \mid X)$ implied by RA and ETA. Our model for the observed data structure O is then given by the model implied by the nonparametric full data model for F_X and $\mathcal{G}(RA)$ for the treatment mechanism $g(A \mid X)$: $\mathcal{M} = \{P_{F_X,g} : F_X \in \mathcal{M}^F, g \in \mathcal{G}(RA)\}$. We will refer to this model as the nonparametric marginal structural model.

The parameter of interest is given by the treatment-specific mean of the counterfactual outcomes $Y(a)$, possibly within strata defined by a subset V of the baseline covariates W. Note that this is a parameter of the full data distribution F_{X_0}. To emphasize that it is a mapping from the full data model $\mathcal{M}^F$ to the class $D(\mathcal{S})$ of real-valued functions defined on a Euclidean set $\mathcal{S}$ containing all possible outcomes of (A, V), we denote this parameter by $\Psi^F : \mathcal{M}^F \to D(\mathcal{S})$, $\Psi^F(F_X)(a, V) = E_{F_X}(Y(a) \mid V)$.

Under the two assumptions we make about the treatment mechanism, it follows that $E(Y(a) \mid V) = E(E(Y \mid A = a, W) \mid V)$, implying that the full data parameter Ψ^F is identified by the observed data. In fact, we have that $\Psi^F(F_X) = \Psi(P_{F_X,g})$ where the observed data model parameter Ψ is defined by $\Psi : \mathcal{M} \to D(\mathcal{S})$, $\Psi(P_{F_X,G})(a, V) = E_{P_{F_X,G}}(E_{P_{F_X,G}}(Y \mid A = a, W) \mid V)$. We denote the true value of this parameter by $\psi_0 = \Psi(P_0)$ and the corresponding parameter space by $\mathbf{\Psi} \equiv \{\Psi(P) : P \in \mathcal{M}\} = \{\Psi^F(F_X) : F_X \in \mathcal{M}^F\}$. We note that Ψ only depends on the F_X-part of the data-generating distribution since the conditional expectations used to define it only involve the a-specific conditional distributions of $Y(a)$ given W, $a \in \mathcal{A}$, and the marginal distribution of W.

Our goal is to use a sample of n i.i.d. observations $O_1, \ldots, O_n$ to estimate the parameter ψ_0. The information contained in the sample can be summarized through its empirical distribution P_n, i.e. the distribution that places probability $1/n$ on each realization O_i. In the following, it will be useful to distinguish between an estimator $\hat{\Psi}$ and its realization $\hat{\psi}$: An estimator $\hat{\Psi}$ is a mapping from empirical distributions to the parameter space $\mathbf{\Psi}$ while $\hat{\psi} = \hat{\Psi}(P_n)$ is the value of this mapping applied to the actual empirical distribution of our sample.

3 Data-adaptive loss-based estimation

van der Laan and Dudoit (2003) present a general road map for unified loss-based estimation that in particular allows this approach to be applied to missing data problems. The general cross-validation selector proposed by these authors is shown to be asymptotically optimal in the following sense: Given a collection of $K(n)$ candidate estimators for a given sample size n, the cross-validation selector performs asymptotically as well as an oracle selector that is allowed to choose among these estimators based on knowledge about the true data-generating distribution. The number of candidate estimators $K(n)$ is allowed to grow polynomially in n, enabling the cross-validation selector to search very aggressively through a large space of candidate estimators as sample size increases. This capability makes unified loss-based estimation a very appealing approach for the problem at hand.

In this section, we develop estimators for our parameter of interest by applying the steps of the road map laid out by van der Laan and Dudoit (2003) to our particular problem. The first step consists of specifying a function whose expectation is minimized at the true parameter value. In our discussion, we will focus on three possible classes of such loss functions that each depend on a nuisance parameter, namely Inverse-Probability-of-Treatment-Weighted (IPTW) loss functions, G-computation based loss functions, and double robust (DR) loss functions. The second step requires us to define a sequence of subspaces of the parameter

space that can approximate the parameter space arbitrarily well. Here we construct such a sieve based on various restrictions on the tensor products of polynomial basis functions we use to parameterize our parameter space. We now define for each subspace a corresponding candidate estimator as the minimizer of the empirical mean of a given loss function over that subspace. To find these candidate estimators, we first obtain an estimate of the nuisance parameter indexing our loss function and then carry out the minimization problem by applying the Deletion/Substitution/Addition algorithm developed by Sinisi and van der Laan (2004). Finally, we use cross-validation to select among these candidate estimators.

3.1 Defining the parameter of interest in terms of a loss function

We first need to specify a function of the observed data whose expectation is minimized at the true parameter value. If we observed the full-data structure X, a possible choice for this loss function would be given by

$$L_h(X, \psi) = \sum_{a \in \mathcal{A}} (Y(a) - \psi(a, V))^2 h(a, V),$$

where $h(a, V)$ denotes a particular weight function. Since we only observe one of the counterfactual outcomes $Y(a)$ for each subject, however, $L_h(X, \psi)$ is not a function of the observed data structure O.

A fundamental objective of the estimating function theory for censored data structures presented in van der Laan and Robins (2002) consists of mapping an estimating function of a full data structure into an observed data function with the same expectation. To this end, van der Laan and Robins (2002) provide an Inverse-Probability-of-Treatment-Weighted (IPTW) mapping as well as a double robust (DR) mapping that is optimal in the sense that it yields observed data estimating functions with minimal variance. van der Laan and Dudoit (2003) show that such mappings form the basis for extending loss-based estimation to missing data problems since they allow us to map infeasible full data loss functions into functions of the observed data whose expectation is likewise minimized at the true parameter value.

In contrast to loss functions used previously, however, loss functions obtained through such mappings are unknown in the sense that they depend on a nuisance parameter. Therefore, it will be necessary to first state precisely how the definitions currently used in loss-based estimation are extended to this more general situation.

For this purpose, let $\Upsilon : \mathcal{M} \rightarrow \mathcal{D}_{nuis}(\mathcal{S}_{nuis})$ be a nuisance parameter that maps our model $\mathcal{M}$ into a space $\mathcal{D}_{nuis}(S_{nuis})$ of real-valued functions on a Euclidean set S_{nuis}. We denote its true value by $\upsilon_0 = \Upsilon(P_0)$. A loss function $(O, \psi, \upsilon) \rightarrow L(O, \psi \mid \upsilon) \in \mathbb{R}$ is now a function that maps an observation O, candidate parameter value $\psi \in \mathbf{\Psi}$, and nuisance parameter value $\upsilon \in \mathbf{\Upsilon}$ into a real number whose expectation at the true nuisance parameter value υ_0 is minimized at ψ_0:

$$\begin{aligned}\psi_0 &= \text{argmin}_{\psi\in\mathbf{\Psi}} \int L(o,\psi \mid \upsilon_0) dP_0(o) \\ &= \text{argmin}_{\psi\in\mathbf{\Psi}} E_0 L(O,\psi \mid \upsilon_0)\end{aligned}$$

In accordance with terminology used in the prediction literature, we define the risk of a candidate parameter value $\psi \in \mathbf{\Psi}$ as $E_{P_0} L(O,\psi \mid \upsilon_0)$. Furthermore, we will refer to the difference between the risk at ψ and the minimal risk at ψ_0 as the risk difference at ψ:

$$d(\psi,\psi_0) \equiv E_0\{L(O,\psi \mid \upsilon_0) - L(O,\psi_0 \mid \upsilon_0)\}$$

We note that $d(\psi,\psi_0) \geq 0$ for all $\psi \in \mathbf{\Psi}$, and if the minimum ψ_0 is unique, then $d(\psi,\psi_0) = 0$ if and only if $\psi = \psi_0$.

In the following, we describe how the IPTW and DR mappings are applied to $L_h(X,\psi)$ to obtain two classes of observed data loss functions. We furthermore provide a mapping that is based on the G-computation formula to obtain a third class of observed data loss functions. As in van der Laan and Robins (2002), we denote the mappings from full data functions to observed data functions with $L \rightarrow IC(O \mid Q, g, L)$, where Q denotes a nuisance parameter based on the F_X-part of the likelihood and g denotes the treatment mechanism.

3.1.1 G-computation loss function

The G-computation loss function is defined as

$$\begin{aligned}L_{h,Gcomp}(O,\psi \mid \upsilon_0) &= IC_{Gcomp}(O \mid Q_0, L_h(\cdot,\psi)) \\ &\equiv \sum_{a\in\mathcal{A}} E_0((Y-\psi(A,V))^2 h(A,V) \mid A=a,W) \\ &= \sum_{a\in\mathcal{A}} \{Q_{02}(a,W) - 2Q_{01}(a,W)\psi(a,V) + \psi(a,V)^2\} h(a,V),\end{aligned}$$

where $Q_{01}(A,W) = E_0(Y \mid A,W)$, $Q_{02}(A,W) = E_0(Y^2 \mid A,W)$ and the nuisance parameter $\upsilon_0 = Q_0 = (Q_{01}, Q_{02})$.

We can verify explicitly that the G-computation mapping $L_h(\cdot,\psi) \rightarrow IC_{Gcomp}(O \mid Q_0, L_h(\cdot,\psi))$ maps the full data loss function $L_h(\cdot,\psi)$ into an observed data function with the same expectation:

$$\begin{aligned}E(L_{h,Gcomp}(O,\psi \mid \upsilon_0)) &= E_0\left(\sum_{a\in\mathcal{A}} E_0((Y-\psi(A,V))^2 h(A,V) \mid A=a,W)\right) \\ &= E_0 \sum_{a\in\mathcal{A}} \{Y(a)^2 - 2Y(a)\psi(a,V) + \psi(a,V)^2\} h(a,V) \\ &= E_0\left(\sum_{a\in\mathcal{A}} (Y(a)-\psi(a,V))^2 h(a,V)\right) \\ &= E_0 L_h(X,\psi)\end{aligned}$$

This implies in particular that the expectation of $L_{h,Gcomp}(O,\psi \mid \upsilon_0)$ is minimized at the true parameter value ψ_0.

3.1.2 IPTW loss function

Choosing $h(a,V) = g_0(a \mid V)$, the IPTW-loss function is defined as

$$\begin{aligned} L_{h,IPTW}(O,\psi \mid \upsilon_0) &= IC_{IPTW}(O \mid g_0, L_h(\cdot,\psi)) \\ &\equiv \frac{(Y-\psi(A,V))^2}{g_0(A \mid W)} g_0(A \mid V), \end{aligned}$$

where the nuisance parameter $\upsilon_0 = g_0 = (g_0(A \mid W), g_0(A \mid V))$.

To see that the expectation of $L_{h,IPTW}(O,\psi \mid \upsilon_0)$ is in fact minimized by ψ_0, we can again verify explicitly that the IPTW-mapping $L_h(\cdot,\psi) \rightarrow IC_{IPTW}(O \mid Q_0, L_h(\cdot,\psi))$ maps the full data loss function $L_h(\cdot,\psi)$ into an observed data function with the same expectation:

$$\begin{aligned} E(L_{h,IPTW}(O,\psi \mid \upsilon_0)) &= E_0\left(\frac{(Y-\psi(A,V))^2}{g_0(A \mid W)} g_0(A \mid V)\right) \\ &= E_0\left(E_0 \frac{(Y-\psi(A,V))^2}{g_0(A \mid W)} g_0(A \mid V) \mid X\right) \\ &\overset{ETA}{=} E_0\left(\sum_{a \in \mathcal{A}} \left\{\frac{(Y(a)-\psi(a,V))^2}{g_0(A=a \mid W)} g_0(A=a \mid V) \mid g_0(A=a \mid W)\right\}\right) \\ &= E_0 L_h(X,\psi) \end{aligned}$$

3.1.3 Double robust loss function

Choosing $h(a,V) = g_0(a \mid V)$, the double robust (DR) loss function is given by

$$\begin{aligned} L_{h,DR}(O,\psi \mid \upsilon_0) &= IC_{DR}(O \mid Q_0, g_0, L_h(\cdot,\psi)) \\ &\equiv \frac{(Y-\psi(A,V))^2}{g_0(A \mid W)} g_0(A \mid V) - \frac{g_0(A \mid V)}{g_0(A \mid W)} E_0\left[(Y-\psi(A,V))^2 \mid A, W\right] \\ &\quad + \sum_{a \in \mathcal{A}} E_0\left[(Y(a)-\psi(a,V))^2 \mid A=a, W\right] g_0(a \mid V), \end{aligned}$$

where

$$\begin{aligned} E_0[(Y-\psi(A,V))^2 \mid A,W] &= Q_{02}(A,W) - 2Q_{01}(A,W)\psi(A,V) + \psi(A,V)^2, \\ E[(Y(a)-\psi(a,V))^2 \mid A=a,W] &= Q_{02}(a,W) - 2Q_{01}(a,W)\psi(a,V) + \psi(a,V)^2, \end{aligned}$$

and the nuisance parameter υ_0 includes both g_0 and $Q_0 = (Q_{01}, Q_{02})$. Note that the conditional expectations in this observed data loss function are indeed identified by these first two conditional moments of the conditional distribution of Y given W.

It can be verified (van der Laan and Robins (2002), section 6.3) that for any treatment mechanism g_1 satisfying the experimental treatment assignment assumption, that is, $\inf_{a \in \mathcal{A}} g_1(a \mid W) > 0$ P_0-a.e., we have

$$E_{P_0} IC(O \mid Q_1, g_1, L_h(\cdot,\psi)) = E_{F_{X0}} L_h(X,\psi) \text{ if either } g_1 = g_0 \text{ or } Q_1 = Q_0$$

In van der Laan and Robins (2002) this identity is referred to as double robustness of the estimating function $IC(O \mid Q_0, g_0, L_h(\cdot \mid \psi))$ for $E_0 L_h(X, \psi)$ w.r.t. misspecification of Q_0, g_0. Thus, if υ is an element of

$$\Upsilon(P_0) \equiv \{(Q, g) : Q = Q_0 \text{ or } g = g_0\},$$

where g ranges over conditional distributions satisfying the ETA assumption, then

$$E_0 L_{h,DR}(O, \psi \mid \upsilon) = E_0 L_{h,DR}(O, \psi \mid \upsilon_0) = E_0 L_h(X, \psi),$$

implying in particular that

$$\psi_0 = \text{argmin}_\psi E_{P_0} L_{h,DR}(O, \psi \mid \upsilon)$$

3.2 Constructing a sieve

Let $P_n \to \hat{\Upsilon}(P_n)$ be a given estimator of the nuisance parameter υ_0 of the loss function. Let $\upsilon_n = \hat{\Upsilon}(P_n)$. We will present a data-adaptive estimator of υ_0 in section 3.5. The minimum empirical risk estimator

$$P_n \to \text{argmin}_{\psi \in \mathbf{\Psi}} \int L(o, \psi \mid \upsilon_n) dP_n(o)$$

suffers from the curse of dimensionality due to the size of the parameter space $\mathbf{\Psi}$ which is given by the space $D(\mathcal{S})$ of all real-valued functions on a Euclidean set $\mathcal{S} \subset \mathbb{R}^d$ containing all possible outcomes of $(A, V) = (A, V_1, \ldots, V_{d-1})$.

One general approach advocated in the theoretical literature for dealing with this problem is based on the construction of a sieve, i.e. a sequence of subspaces that can approximate the whole parameter space $\mathbf{\Psi}$ arbitrarily well (LeCam, 1986; Barron, 1989; LeCam and Yang, 1990; Barron, 1991; Shen, 1997; Barron et al., 1999). For each subspace a corresponding candidate estimator is defined as the minimizer of the empirical risk over that subspace. The final estimator is then selected from among these candidate estimators as the minimizer of an appropriately penalized empirical risk or a cross-validated empirical risk. We here follow this approach by first parameterizing the parameter space in terms of linear combinations of a particular choice of basis functions and then defining subspaces based on various restrictions on these linear combinations.

A general class of basis functions for the space $D(\mathbb{R}^d)$ of real-valued functions on $\mathbb{R}^d$ can be obtained from tensor products of univariate basis functions. We here choose the polynomial powers $e_1 = 1, e_2 = x, e_3 = x^2, \ldots$ as univariate basis functions, but we note that many other choices are possible. One may for example consider spline basis functions of fixed degree with corresponding fixed set of knot points or wavelet basis functions. Given a vector $\vec{p} = (p_1, \ldots, p_d) \in \mathbb{N}^d$, we let $\phi_{\vec{p}}(A, V) = e_{p_1}(A) \times e_{p_2}(V_1) \times \ldots \times e_{p_d}(V_{d-1})$ denote the tensor product of univariate basis functions identified by $\vec{p}$. Using polynomial powers as univariate basis functions, we have $\phi_{\vec{p}}(A, V) = A^{p_1} V_1^{p_2} \ldots V_{d-1}^{p_d}$. The collection $\{\phi_{\vec{p}} : \vec{p}\}$ provides now a basis for the space $D(\mathbb{R}^d)$ of real-valued functions on $\mathbb{R}^d$.

If the conditional expectation of $Y(a)$ given V can only take on values in some proper subset $\mathcal{Y}$ of $\mathbb{R}$, the parameter space is a proper subset of the space of real-valued functions on $\mathcal{S}$. If Y is binary, for example, the parameter space consists of all functions $\mathcal{S} \to \mathcal{Y} = [0, 1]$. To respect such constraints, we parameterize $\mathbf{\Psi}$ in terms of a known transformation $t : \mathbb{R} \to \mathcal{Y}$ of linear combinations of basis functions for $D(\mathbb{R}^d)$. These transformations represent the familiar link functions used in generalized linear models, as for example the logit link function for binary outcomes. This allows us to write

$$\mathbf{\Psi} = \left\{ t\left(\sum_{\vec{p} \in I} \beta_{\vec{p}} \phi_{\vec{p}} \right) : I \subset \mathcal{I}, \beta \right\},$$

where I denotes an index set containing a collection of vectors $\vec{p}$, $\mathcal{I}$ denotes the collection of indices of the allowed subsets of basis functions, and β ranges over Euclidean sets guaranteeing that the linear combinations are contained in the parameter space. Thus, for each subset of basis functions $I \subset \mathcal{I}$, we have that $\beta \in B_I \equiv \{\beta : \sum_{\vec{p} \in I} \beta_{\vec{p}} \phi_{\vec{p}} \in \mathbf{\Psi}\}$. In the following, we will use

$$\psi_{I,\beta} \equiv t\left(\sum_{\vec{p} \in I} \beta_{\vec{p}} \phi_{\vec{p}} \right)$$

to denote the element of the parameter space $\mathbf{\Psi}$ corresponding to a given index set I and given $\beta \in B_I$.

Given this parameterization, we can now define subspaces $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ by requiring the subsets I of basis functions to be contained in $\mathcal{I}_s \subset \mathcal{I}$ or requiring the values for the corresponding coefficients $(\beta_{\vec{p}} : \vec{p} \in I)$ to be contained in $B_{I,s} \subset B_I$:

$$\mathbf{\Psi}_s = \{\psi_{I,\beta} : \ I \in \mathcal{I}_s \subset \mathcal{I}, \ \beta \in B_{I,s} \subset B_I\}$$

In our `R` package `cvDSA`, a subspace $\mathbf{\Psi}_s$ is currently indexed by $s = (k_1, k_2)$, with k_1 giving the maximum number of basis functions that each linear combination may consist of and k_2 giving the maximum number $m(\vec{p})$ of non-zero components in $\vec{p}$:

$$\mathbf{\Psi}_{k_1,k_2} = \left\{ t(\sum_{\vec{p} \in I} \beta_{\vec{p}} \phi_{\vec{p}}) : \mid I \mid \leq k_1, \ \max_{\vec{p} \in I} m(\vec{p}) \leq k_2 \right\}$$

One might also consider other functions $m(\vec{p})$ to measure the 'complexity' of a given basis function. In addition, one could introduce another parameter k_3 to restrict the parameter vector β, letting for example $B_{I,s=(k_1,k_2,k_3)} = \{\beta \in B_I : \sum_{\vec{p} \in I} \mid \beta_{\vec{p}} \mid \leq k_3\}$. Finally, we note that the set A_n containing all possible values of s may be allowed to depend on the sample size n.

3.3 Generating candidate estimators

For each subspace $\mathbf{\Psi}_s$, we now define a corresponding estimator $\hat{\Psi}_s$ as the minimizer of the empirical risk over that subspace:

$$\hat{\Psi}_s(P_n) \equiv \text{argmin}_{\psi \in \Psi_s} \int L(o, \psi \mid \upsilon_n) dP_n(o)$$

The task of computing $\hat{\Psi}_s(P_n)$ is naturally split into two sequential steps. Given a particular subset $I \in \mathcal{I}_s$ of basis functions, it is straightforward to obtain an estimate of β by minimizing the empirical mean of the loss function of interest:

$$\beta(P_n \mid I, s) \equiv \text{argmin}_{\beta \in B_{I,s}} \int L\left(o, \psi_{I,\beta} \mid \upsilon_n\right) dP_n(o)$$

For the IPTW and DR loss functions, it is easily seen that this estimator is equivalent with solving the IPTW and DR estimating equations for the parametric marginal structural model $E(Y(a) \mid V) = \psi_{I,\beta}(a, V)$, treating I as given (van der Laan and Robins, 2002). The estimate $\beta_{IPTW}(P_n \mid I, s)$ based on the IPTW loss function, for example, is thus obtained through a weighted least-squares regression of Y on A and V according to the model $E(Y \mid A, V) = \psi_{I,\beta}(A, V)$, with weights given by $w_i = \frac{g_n(A_i|V_i)}{g_n(A_i|W_i)}$, $i = 1, \ldots, n$.

For each I, this first step results in an estimator $\psi_{I,s,n} = \hat{\Psi}_{I,s}(P_n) \equiv \psi_{I,\beta(P_n|I,s)}$. Now, it remains to minimize the empirical risk over all allowed subsets $I \in \mathcal{I}_s$ of basis functions. Specifically, one needs to minimize the function $f_E : \mathcal{I}_s \to \mathbb{R}$ defined by

$$f_E(I) \equiv \int L\left(o, \widehat{\Psi}_{I,s}(P_n)\right) dP_n(o)$$

This step corresponds to searching through candidate functional forms for the marginal structural model and is carried out here by using the Deletion/Substitution/Addition (D/S/A) algorithm developed by Sinisi and van der Laan (2004). In contrast to previously proposed forward/backward selection approaches, this algorithm performs an extensive and aggressive search over the set of candidate functional forms that is truly aimed at identifying the corresponding minimum empirical risk. This provides us now with the empirical risk minimizer $\psi_{s,n} = \hat{\Psi}_s(P_n)$ for each choice of subspace $\mathbf{\Psi}_s$, where

$$\hat{\Psi}_s(P_n) = \hat{\Psi}_{I_s(P_n),s}(P_n) = \psi_{I_s(P_n),\beta(P_n|I_s(P_n),s)}$$

These estimators $\hat{\Psi}_s(P_n)$, $s \in \mathcal{A}_n$, represent the set of candidate estimators of ψ_0.

3.4 Selecting among candidate estimators by cross-validation

We select among these candidate estimators by cross-validation. The general idea of this approach is to divide the original dataset into a training set that is used to arrive at a candidate estimator $\hat{\Psi}_s$ and a validation set that is then used to obtain an independent estimate of the risk of that estimator.

For this purpose, let $B_n \in \{0,1\}^n$ be a random vector whose observed value defines a split of the observed data $O_1, \ldots, O_n$ into a validation sample and a training sample. If $B_n(i) = 0$, then observation i is placed in the training sample, and if $B_n(i) = 1$, it is placed in the validation sample. We will denote the empirical distribution of the data in the training sample and validation sample with P^0_{n,B_n} and P^1_{n,B_n}, respectively. The proportion of observations in the validation sample is denoted by $p = \sum_i B_n(i)/n$.

The choice of distribution for B_n corresponds now with different possible cross-validation schemes presented in the literature such as V-fold cross-validation or Monte-Carlo cross-validation. In our application, we focus on V-fold cross-validation by dividing the data set into V separate groups and letting the split vector B_n have a uniform distribution on V different realizations such that each of the V groups is used exactly once as the validation sample.

After applying the estimator $\hat{\Psi}_s$ to a given training sample, its performance is measured through the empirical mean of the loss function over the corresponding validation sample. We note that the nuisance parameter υ_0 indexing the loss function is estimated by applying the corresponding estimator $\hat{\Upsilon}$ only to the training sample as well rather than to the entire data set. The cross-validated empirical risk is then given by the expectation of this estimated empirical risk over different realizations of the split vector B_n; in the case of V-fold cross-validation, this expectation corresponds to the simple mean of the estimated empirical risk over the V different splits of the observed data. We now choose the index s that minimizes the cross-validated empirical risk:

$$\hat{S}(P_n) \equiv \text{argmin}_{s \in \mathcal{A}_n} E_{B_n} \int L(o, \hat{\Psi}_s(P^0_{n,B_n}) \mid \hat{\Upsilon}(P^0_{n,B_n})) dP^1_{n,B_n}(o)$$

Our final estimator of ψ_0 is obtained by applying the corresponding estimator $\hat{\Psi}_{\hat{S}(P_n)}$ to the entire sample:

$$\psi_n = \hat{\Psi}(P_n) \equiv \hat{\Psi}_{\hat{S}(P_n)}(P_n)$$

3.5 Estimating the nuisance parameter υ_0 data-adaptively

It remains to present an estimator for the nuisance parameter υ_0 indexing the loss function. For the IPTW loss functions, this nuisance parameter consists of the treatment mechanism $g_0 = (g_0(A \mid V), g_0(A \mid W))$; for the G-computation loss functions, it is given by the regressions $Q_0 = (E_0(Y \mid A, W), E_0(Y^2 \mid A, W))$; the DR loss functions, finally, rely on both of these nuisance parameters.

We estimate these nuisance parameters data-adaptively by applying the same loss-based estimation approach used to estimate the parameter of interest $E[Y(a) \mid V]$. If the treatment variable A is binary, $g_0(A \mid V)$ and $g_0(A \mid W)$) are densities and we accordingly rely on the $-\log(\cdot)$ loss function; if the treatment variable A is continuous, these nuisance parameters are regressions and we rely on the squared-error loss function, which is also used for estimating $E_0(Y \mid A, W)$ and $E_0(Y^2 \mid A, W)$. In contrast to the loss functions used to estimate our parameter of interest, these loss functions do not depend on a nuisance parameter, thus simplifying the estimation problem somewhat. As above, we use a polynomial sieve and

search aggressively over candidate subspaces by applying the D/S/A algorithm. We note that since the resulting estimator $\hat{\Upsilon}$ of the nuisance parameter υ_0 is itself based on cross-validation, the selection of the index s described in section 3.4 involves in fact a nested cross-validation, consisting of an inner one for the estimation of υ_0 and an outer one for the selection of s.

By the G-computation formula for point-treatment studies

$$\psi_0 = E_0[Y(a) \mid V] = E_0[E_0[Y \mid A, W] \mid V] = E_0[Q_{01}(A, W) \mid V]$$

so that ψ_0 is a function of the nuisance parameter Q_{01}. Thus it may happen that the model selected for Q_{01} is in fact not compatible with the model selected for ψ_0. Furthermore, the model selected for $Q_{02} = E[Y^2 \mid A, W]$ may not be compatible with the model selected for $Q_{01} = E[Y \mid A, W]$. Such incompatibilities do not represent a major problem, however, since Q_{01} and Q_{02} are only involved in estimating the loss function and are thus not used to make any claims about the functional form of ψ_0. Forcing the models for Q_{01} and Q_{02} to be compatible with each other and the model for ψ_0 would complicate the the model selection process considerably and might in fact result in worse estimation of the loss function.

3.6 Assessing the performance of an estimator

The performance of machine-learning algorithms is commonly assessed on the basis of an estimated risk. Our `R` package reports the cross-validated empirical risk for a given estimator, obtained through V-fold cross-validation, We thus apply the entire estimation procedure described above to each of V learning samples separately and evaluate the resulting estimators on the corresponding validation samples, forcing us to add another layer of cross-validation to those already inherent in the estimation procedure itself. We note that the loss function used to obtain this empirical risk estimate need not be the same as the loss function indexing the estimator itself.

A general problem with data-adaptive model selection techniques lies in the difficulty to obtain honest measures of statistical variability and significance. Our `R` package currently does not report standard errors or confidence intervals for parameter estimates, but we note that, where desired, these may be obtained by the user through re-sampling based methods such as the bootstrap. Optimistic inference can furthermore be obtained by treating all selected models as fixed and given.

4 Simulation study

In the previous section we developed three data-adaptive estimators of the treatment-specific mean that differ in the type of loss function they employ: A G-computation loss function, an IPTW loss function, or a DR loss function. Each of these loss functions depends on a different nuisance parameter υ_0, consisting either of the regressions $Q_0 = (E_0[Y \mid A, W], E_0[Y^2 \mid A, W])$, the treatment mechanism $g_0 = (g_0(A \mid V), g_0(A \mid W))$, or both. The performance of the corresponding estimators can thus be expected to vary depending on how well these

nuisance parameters can be estimated using the polynomial sieve proposed here. The IPTW estimator can furthermore be expected to be sensitive to violations of the experimental treatment assignment (ETA) assumptions. In this section, we present a simulation study that is aimed at examining the impact of these different dependencies in practice.

In each simulation, the data set consists of continuous baseline covariates $W = (V, W_1, W_2)$, a binary treatment variable A, and a continuous outcome Y. The baseline covariates are independent of each other and distributed as

$$\begin{aligned} V &\sim U(0,1) \\ W_1 &\sim U(-10,10) \\ W_2 &\sim U(0,1) \end{aligned}$$

The counterfactual outcomes $Y(a)$ are distributed as $Y(a) = Q_0(a,W) + \varepsilon,\ \varepsilon \sim N(0,1)$, where we consider two choices for $Q_0(A,W) \equiv E_0[Y \mid A, W]$:

$$\begin{aligned} Q_{0,1}(A,W) &= -2 + 0.7A + V + 0.5V^2 + 0.2W_1 - W_2 \\ Q_{0,2}(A,W) &= -2 + 0.7A + \sin 10V + 0.5V^2 + \sin W_1 - W_2 \end{aligned}$$

Note that $Q_{0,1}$ can be much more easily approximated by a polynomial sieve than $Q_{0,2}$, suggesting that the G-computation loss function and possibly also the DR loss function will be more reliably estimated under $Q_{0,1}$ than under $Q_{0,2}$. The corresponding true treatment-specific mean $\psi_0(A,V)$ is given by

$$\begin{aligned} \psi_{0,1}(A,V) &= -2.5 + 0.7A + V + 0.5V^2 \\ \psi_{0,2}(A,V) &= -2.5 + 0.7A + \sin 10V + 0.5V^2 \end{aligned}$$

Since $\psi_{0,1}$ can again be much more easily captured by a polynomial sieve than $\psi_{0,2}$, we expect in fact that all three estimators will behave more poorly in selecting an appropriate marginal structural model under $Q_{0,2}$ than under $Q_{0,1}$. We consider four choices for the treatment mechanism $g_0(W) \equiv P(A = 1 \mid W)$ that differ both in how well they can be captured by a polynomial sieve and in the extent to which they violate the ETA assumption:

$$\begin{aligned} g_{0,1}(W) &= \text{logit}^{-1}(-1 + 2V + 0.0004W_1 - 0.5W_2) \\ g_{0,2}(W) &= \text{logit}^{-1}(-1 - 1.5V - 0.02W_1 + 3W_2) \\ g_{0,3}(W) &= \text{logit}^{-1}(0.2 + 0.5 \times \sin 10V + 0.5 \times \sin W_1 - 0.7 \times \sin 10W_2) \\ g_{0,4}(W) &= \text{logit}^{-1}(1 + \sin 10V + 1.5 \times \sin W_1 - 0.5 \times \sin 10W_2) \end{aligned}$$

The first two treatment mechanisms are much more easily captured by a polynomial sieve than the remaining two. The first and the third treatment mechanism generate no observations with $g_0(A \mid W) < 0.1$ or $g_0(A \mid W) > 0.9$), representing practical violations of the ETA assumption, while $g_{0,2}$ and $g_{0,4}$ generate approximately 35% and 22% of practical ETA violations, respectively. The IPTW loss function can thus be expected to be reliably estimated only under $g_{0,1}$, with the remaining three choices representing more difficult scenarios.

We assess the performance of a given estimator using the risk dissimilarity

$$\begin{aligned} d(\hat{\Psi}(P_n), \psi_0) &= E_0 L(X, \hat{\Psi}(P_n)) - E_0 L(X, \psi_0) \\ &= \sum_{a \in \{0,1\}} \int (\hat{\Psi}(P_n)(a, V) - \psi_0(a, V))^2 dF_0(V), \end{aligned}$$

where we approximate the minimum of the full data risk function

$$\begin{aligned} E_0 L(X, \psi_0) &= E[\sum_{a=\{0,1\}} (Y(a) - \psi_0(a, V))^2] \\ &= \sum_{a=\{0,1\}} E(Q_0(a, W) + \varepsilon - \psi_0(a, V))^2 \end{aligned}$$

numerically as the mean of $L(X, \psi_0)$ over 20,000 realizations of X.

For each of the eight possible choices for (Q_0, g_0), we generated 100 data sets each consisting of 1000 i.i.d. observations. We note that the number of replications we were able to carry out for this simulation study was limited by the considerable time requirements of the proposed data-adaptive methodology. Table 1 summarizes the mean risk dissimilarities for each of the three estimators.

Table 1: Mean risk dissimilarities $\int(\hat{\Psi}(P_n) - \psi_0)^2$ based on 100 samples of size $n = 1000$

	Q_0	g_0	ETA	Estimator		
(Q_0, g_0)	polynomial	polynomial	holds	G-comp	IPTW	DR
$(Q_{0,1}, g_{0,1})$	✓	✓	✓	0.0215	0.0325	0.0236
$(Q_{0,1}, g_{0,2})$	✓	✓		0.0486	0.1532	0.0641
$(Q_{0,1}, g_{0,3})$	✓		✓	0.0179	0.0305	0.0275
$(Q_{0,1}, g_{0,4})$	✓			0.0216	0.0416	0.0351
$(Q_{0,2}, g_{0,1})$		✓	✓	0.7915	0.5151	0.4440
$(Q_{0,2}, g_{0,2})$		✓		0.8124	0.7487	0.6537
$(Q_{0,2}, g_{0,3})$			✓	0.9148	0.8813	0.8839
$(Q_{0,2}, g_{0,4})$				1.2604	1.2021	1.1946

As is to be expected, all three estimators perform considerably better if ψ_0 is easily approximated by sums of polynomial powers than if this is not the case. Thus the mean risk dissimilarities under $Q_{0,1}$ are about 30 to 50 times smaller than under $Q_{0,2}$.

The G-computation estimator depends only on a good estimate of Q_0 and thus performs very well if Q_0 is easily captured by a polynomial sieve, but rather poorly if this is not the case. If $Q_0 = Q_{0,1}$, this estimator is in fact more efficient than either of the two remaining estimators, regardless of the choice for the treatment mechanism. If $Q_0 = Q_{0,2}$, however, it is the least efficient of the three estimators.

The IPTW estimator, depending only on g_0, is expected to work well if the treatment mechanism satisfies the ETA assumption and is easily approximated by sums of polynomial powers. Hence it performs competitively under $g_0 = g_{0,1}$, but considerably worse under the remaining treatment mechanisms. It appears particularly sensitive to the large proportion of practical ETA violations caused by $g_0 = g_{0,2}$.

The DR estimator relies on estimating either Q_0 or g_0 sufficiently well, giving it a broader range of data-generating distributions under which it performs favorably. Under $Q_0 = Q_{0,1}$, it is slightly less efficient than the G-computation estimator, but more efficient than the IPTW estimator. If the treatment mechanism violates the ETA assumption or is hard to approximate by sums of polynomial powers, it performs considerably better than the IPTW estimator by being able to rely on a polynomial Q_0. Under $Q_0 = Q_{0,2}$ and $g_0 = g_{0,1}$, it gains considerably in efficiency relative to the G-computation estimator by being able to rely on a polynomial treatment mechanism satisfying the ETA assumption. Under these circumstances, it is also slightly more efficient than the IPTW estimator. If both Q_0 and g_0 are difficult to capture by a polynomial sieve, the DR estimator is still somewhat more efficient than the G-computation estimator, although the gains are far less pronounced.

These results confirm our theoretical understanding and are also in agreement with simulation studies presented by Neugebauer and van der Laan (2004) and Yu and van der Laan (2003) that compare the performance of these three estimators assuming pre-specified parametric models for Q_0, g_0, and ψ_0.

5 Application to SPPARCS data

In this section we apply our methodology to a data set derived from the "Study of Physical Performance and Age-Related Changes in Sonomans" (SPPARCS) (Tager et al., 1998). The SPPARCS project is a population-based study of people aged 55 years and older living in and around Sonoma, CA, that is aimed at investigating the effect of physical activity and cardiopulmonary function on morbidity and mortality in the elderly.

We here wish to assess the causal effect of differences in lung function, as measured by forced expiratory volume (FEV), on the survival experience of this population. While it is hard to imagine an intervention that would allow one to set FEV to any desired level, estimation of this causal effect is still interesting from a mechanistic point of view in that it might shed light on the extent to which changes in FEV lie on the causal pathway to mortality. We are furthermore interested in identifying factors that might modify the effect of FEV on survival.

5.1 Data structure

A population-based sample of 846 men and 1,246 women aged 55 years and older was recruited during the years of 1993 and 1994. Table 2 summarizes the available baseline covariates W that will be used in our analysis. The set V of potential stratification variables contains all variables in W except for *diabet*, *ldl*, *hdl*, and *bmi*. The treatment variable A is

given by the forced expiratory volume in one second (FEV1), measured in L/min.

Table 2: Definition of baseline covariates W. ETS stands for environmental tobacco smoke.

Variable	Definition
age	Age in years
$female$	Indicator for female sex
$cardio$	Indicator for diagnosis with cardiovascular condition
$diabet$	Indicator for diagnosis with diabetes
ldl	Indicator for high LDL cholesterol
hdl	Indicator for high HDL cholesterol
bmi	Body mass index
$currsmk$	Indicator for currently smoking
$pastsmk$	Indicator for previously smoking
$etshome$	Number of years of ETS exposure at home
$etsspouse$	Number of years of ETS exposure from spouse
$etswork$	Number of years of ETS exposure at work

The survival status of the entire cohort is available until the end of the study in August 2003. A total of 541 subjects died during the study period. Let S denote survival time measured in months since the beginning of the study. Let C denote follow-up time measured in months from recruitment until the end of the study. Due to censoring by C, the mean of S is not identifiable from the data that have been collected. A common solution to this problem is to model a comparatively low quantile of the survival time distribution, the first decile, for example, as a function of the explanatory variables of interest. Although the methodology presented here could easily be extended to such models, our current implementation, is restricted to modelling the treatment-specific *mean* of a particular outcome,

We here define a truncated survival time T as the minimum of S and C and let $Y = \log(T)$ be our outcome of interest. The parameter of interest is now given by the treatment-specific mean of this log-transformed truncated survival time T. In contrast to the treatment-specific mean of the original survival times S, this parameter is identifiable. While causal effects of FEV on T do not have as straightforward an interpretation as causal effects of FEV on S, we further note that this parameter is capable of capturing causal effects of FEV on S that manifest themselves within the nine or ten years comprising the study period. In future work we also plan to consider the outcome $Y = I(S > C)$, with the corresponding parameter of interest given by the treatment-specific probability of survival over the entire study period. In fact, this approach can be applied to estimate the treatment-specific probability of survival past an arbitrary time point during the study period, allowing us to obtain an estimate of the entire treatment-specific survival function.

5.2 Missing covariate and treatment measurements

A significant proportion of subjects have missing covariate or treatment measurements. As mentioned above, the outcome Y is observed for the entire cohort. We handle missing values differently for stratification variables V, baseline covariates W that are not included in V, and the treatment variable A.

Fortunately, the stratification variables V have only a very small proportion of missing values. We therefore simply remove the subjects with missing values for V from our analysis. For baseline covariates W that are not included in V, we re-define the variables to include an indicator Δ_j that equals 1 if the measurement for variable W_j is available: $W_j \equiv (\Delta_j, \Delta_j W_j)$. This is useful since the missingness of baseline covariates W may itself contain valuable information about a subject. We note that the randomization assumption now requires that treatment assignment is conditionally independent of the counterfactual outcomes given this re-defined collection of baseline covariates W.

This approach for handling missing values does not apply to missing treatment measurements since we are not interested in estimating the causal effect of FEV on survival conditional on having a valid treatment measurement. Instead we address such missing values using an inverse-probability-of-missingness approach. Specifically, we define an indicator ξ that equals one if FEV is observed and zero otherwise. The loss function of interest is then weighted by $I(\xi = 1)/\hat{P}(\xi = 1 \mid W)$, where $\hat{P}(\xi = 1 \mid W)$ is an estimate of the missingness mechanism $P_0(\xi = 1 \mid W)$. Like v_0, this additional nuisance parameter is estimated data-adaptively by loss-based estimation using a polynomial sieve, the D/S/A algorithm, and V-fold cross-validation. Ideally, the cross-validation procedure would estimate the missingness mechanism separately for each split into training and validation sample, but the current version of our `R` package `cvDSA` does not yet have this capability. Instead, the missingness mechanism is estimated once and for all at the beginning of the analysis and then treated as fixed for all remaining step. The fit we obtained is given by

$$\hat{P}(\xi = 1 \mid W) = \text{logit}^{-1}(4 - 0.06\, age + 0.02\, age \times hdl)$$

5.3 Data-adaptive fits for the nuisance parameters Q_0 and g_0

We obtained the following data-adaptive fits for the nuisance parameters g_0 and Q_0:

$$\begin{aligned}
\hat{g}(A \mid V) &= 6.02 - 0.043\, age - 0.94\, female - 0.46\, currsmk - \\
&\quad 0.20\, cardio \times pastsmk - 0.0083\, pastsmk \times etswork \\
\hat{g}(A \mid W) &= 5.91 - 0.043\, age - 0.93\, female - 0.0093\, pastsmk \times etswork - \\
&\quad 0.45\, currsmk \times bmi + 0.12\, bmi \times ldl \\
\hat{E}(Y \mid A, W) &= 6.97 - 0.55\, A - 0.036\, age - 0.67\, female + 0.0087\, A \times age - \\
&\quad 0.12\, A \times cardio - 0.0050\, age \times cardio + 0.011\, age \times female - \\
&\quad 0.43\, cardio \times diabet - 0.0051\, currsmk \times etswork - \\
&\quad 0.058\, female \times pastsmk \\
\hat{E}(Y^2 \mid A, W) &= 24.84 - 0.00068\, age^2 - 0.00026\, age^2 \times cardio
\end{aligned}$$

Since the survival time S of about 75% of the subjects is truncated by the end of the study, we might expect that the nuisance parameters $E_0[Y \mid A, W]$ and $E_0[Y^2 \mid A, W]$ are not easily captured by a polynomial sieve. We therefore examined the fit of the models we obtained for these parameters by plotting observed against fitted values. Plots (c) and (d) of figure 1 do in fact show that the models we obtained for $E_0[Y \mid A, W]$ and $E_0[Y^2 \mid A, W]$ fit the data quite poorly. The models for $g_0(A \mid W)$ and $g_0(A \mid V)$ on the other hand appear to fit the data quite well. We also examined the validity of the ETA assumption $\sup_a g_0(a \mid V)/g_0(a \mid W) < \infty$ by inspecting the observed weights $\hat{g}(A_i \mid V_i)/\hat{g}(A_i \mid W_i)$. Since these observed weights were nicely bounded, we are comfortable that the ETA assumption is not practically violated.

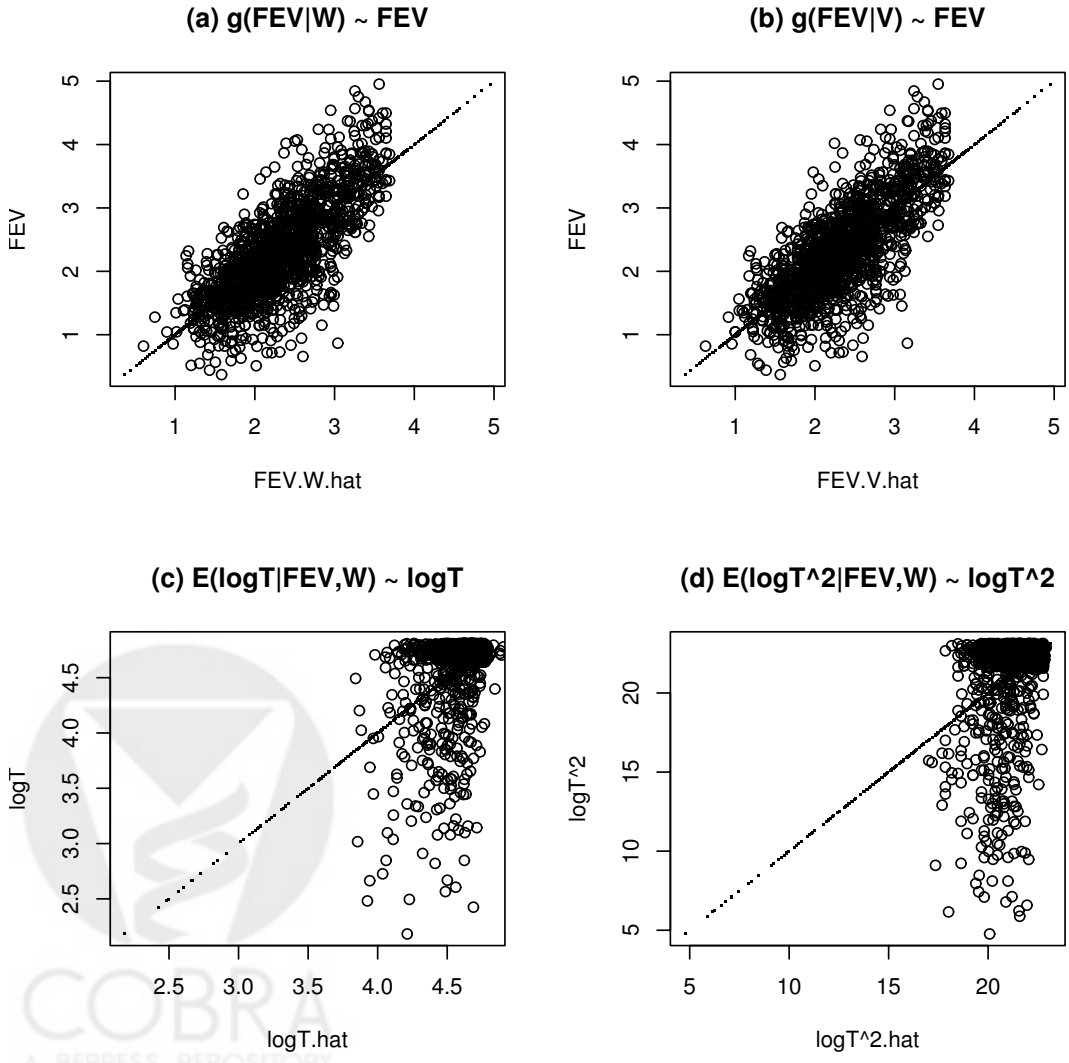

Figure 1: Plots of observed values against fitted values for the four nuisance parameters $g_0(A \mid W)$, $g_0(A \mid V)$, $E_0[Y \mid A, W]$, and $E_0[Y^2 \mid A, W]$

5.4 Data-adaptive fits for the marginal structural model

We obtained the following data-adaptive fits for the parameter of interest ψ_0. The performance of each estimator is measured through its cross-validated empirical risk based on the DR loss function, regardless of the loss function indexing the estimator.

- G-computation estimator:

  Cross-validated risk estimate = 1.3628

$$\begin{aligned} E(Y(a) \mid V) &= 4.69 - 2.52\, age \times cardio - 0.0049\, currsmk \times etswork + \\ &\quad 0.091\, cardio \times currsmk - 0.0000059\, age \times etshmtot - \\ &\quad 0.00097\, pastsmk \times etshmtot + 0.00085\, currsmk \times etshmtot - \\ &\quad 0.00030\, currsmk \times etsspouse \end{aligned}$$

- IPTW estimator:

  Cross-validated risk estimate = 0.7781

$$\begin{aligned} E(Y(a) \mid V) &= 5.14 - 0.011\, age - 0.0069\, age \times cardio + 0.0019\, age \times female + \\ &\quad 0.16\, a \times cardio + 0.0013\, a \times age \end{aligned}$$

- DR estimator:

  Cross-validated risk estimate = 1.3122

$$\begin{aligned} E(Y(a) \mid V) &= 7.24 - 0.62\, a - 0.040\, age - 0.77\, female - 0.48\, cardio + \\ &\quad 0.0099\, a \times age + 0.013\, age \times female - 0.062\, a \times currsmk + \\ &\quad 0.15\, a \times cardio - 0.056\, female \times pastsmk \end{aligned}$$

The relative performances of the three estimators are in line with our observations above regarding the fit of the data-adaptive models for Q_0 and g_0 as well as the validity of the ETA assumption. The good performance of the IPTW estimator is not surprising since the model for the treatment mechanism fits the data quite well and the ETA assumption appears to hold. The comparatively poor performance of the G-computation estimator, as evidenced in a nearly two-fold greater cross-validated risk, is likely due to the poor fit of the models we obtained for $E_0[Y \mid A, W]$ and $E_0[Y^2 \mid A, W]$. The DR estimator also suffers from the resulting poor estimate of Q_0, although it appears to benefit slightly from the good fit for the treatment mechanism.

Table 3 summarizes the coefficient estimates obtained by the IPTW estimator. Standard errors and p-values are based on the influence curve of this estimator, treating the models for ψ_0 and v_0 as given. Since they ignore the data-adaptive selection of these models, they can be expected to exaggerate the significance of the estimates obtained. The resulting p-values are small enough, however, to suggest that all parameter estimates are in fact significant at a conventional level of $\alpha = 0.05$.

Table 3: Summary of MSM fit obtained by IPTW estimator

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	5.1397692	0.0979460	52.476	<2e-16
age	-0.0114990	0.0012458	-9.230	< 2e-16
$a \times cardio$	0.1585765	0.0305650	5.188	2.45e-07
$a \times age$	0.0012705	0.0002729	4.656	3.55e-06
$age \times cardio$	-0.0069347	0.0009418	-7.363	3.10e-13
$age \times female$	0.0018915	0.0003656	5.173	2.65e-07

The model suggests a positive effect of increased FEV on survival that is modified by age and whether or not a subject has been diagnosed with a cardiovascular condition. Table 4 summarizes the estimated proportional change in predicted truncated survival T caused by a 1L/min increase in FEV at different levels of these two effect modifiers.

Table 4: Estimated proportional change in predicted T caused by a 1L/min increase in FEV

Age in years	60	65	70	75	80
Cardio=0	1.079	1.086	1.093	1.100	1.107
Cardio=1	1.265	1.273	1.281	1.289	1.297

Our analysis suggests that the positive effect of greater FEV on survival is considerably greater among people who have previously been diagnosed with cardiovascular disease. Among this group of people, the predicted truncated survival T increases by 25% to 30% for a 1L/min increase in FEV whereas the corresponding increase among people without a previous diagnosis ranges only from 8% to 11%. Furthermore, the positive effect of FEV on survival appears to increase slightly with age. Finally, we note that none of the environmental tobacco smoke variables appear in the chosen marginal structural model, suggesting that past tobacco smoke exposure may exert most of its effect on survival through changes in FEV. Such a hypothesis may be more formally evaluated using statistical methods for the estimation of direct and indirect effects (Robins and Greenland, 1992; Pearl, 2000; van der Laan and Petersen, 2004).

6 Discussion

Current methodology for estimating the treatment-specific mean in a point-treatment study requires the researcher to specify *a priori* a marginal structural model for the dependence of counterfactual outcomes on treatment and effect modifiers as well as parametric models for various nuisance parameters. Since model misspecification can lead to severely biased causal

effect estimates, the dependence on such parametric models represents a major limitation. In this article, we introduce estimators that allow the marginal structural model as well as the parametric models for the relevant nuisance parameters to be selected data-adaptively.

The estimators we present are based on the unified loss-based estimation approach recently developed by van der Laan and Dudoit (2003) that in particular extends loss-based estimation to missing data problems such as causal inference. The key idea of this approach consists of mapping an infeasible full-data loss function into an observed-data loss function with the same expectation. Starting from a squared-error full-data loss function and applying three different mappings into an observed-data loss function, we thus obtain the G-computation, IPTW, and DR estimators. These estimators differ primarily in their dependence on two nuisance parameters Q_0 and g_0. Using both simulation studies and an actual data analysis, we demonstrate how their performances are differentially affected by how well these nuisance parameters can be estimated in a given situation.

While the data-adaptive methodology we develop here represents a major advance over current parametric methods, it still leaves room for a number of improvements and extensions. One of the greatest shortcomings lies certainly in the difficulty to arrive at honest measures of statistical significance. Currently, these can only be obtained through computationally intensive resampling-based approaches like the bootstrap. As we pointed out above, however, this represents a problem encountered with most machine-learning algorithms.

A more specific issue with the methodology presented here pertains to the experimental treatment assignment assumption. Estimators based on the IPTW mapping rely fundamentally on this assumption according to which, for any realization of baseline covariates W, all possible treatments must have positive probability of being observed. While IPTW estimators are becoming increasingly popular tools in the area of causal inference, this assumption is seldomly examined in a systematic manner, if at all. Most current approaches are based on examining fitted treatment probabilities over a range of plausible covariate values and ensuring that these are bounded away from zero and one. While such approaches may reveal whether or not the ETA assumption is practically violated, they do not give us a sense of how severely biased we may expect our parameter estimates to be as a consequence.

We are currently working on a simulation approach that aims to evaluate the validity of the ETA assumption exactly by estimating the size of this bias. For this purpose, we first carry out the entire data-adaptive estimation procedure described in this article to select a marginal structural model and to obtain fits for the nuisance parameters Q_0 and g_0. Estimating the distribution of baseline covariates W by the corresponding empirical distribution and assuming that continuous treatment or outcome variables are normally distributed, we then have an estimate of the entire data-generating distribution. We can now obtain a sampling distribution of IPTW estimates by applying the parametric IPTW estimator proposed by Robins (2000b) to data sets simulated according to this data-generating distribution. The true parameter values corresponding to this distribution can be approximated by using the parametric G-computation estimator (Robins, 2000b). Finally we can estimate the bias introduced by a potential violation of the ETA assumption by the difference between these G-computation estimates and the mean of the IPTW estimates. We are currently working on implementing this approach for a future version of the `R` package `cvDSA`.

Another issue regarding the estimators we present here pertains to the estimation of the nuisance parameters Q_0 and g_0. As described above, we currently employ the squared-error loss function or the $-\log(\cdot)$ loss function for this purpose. These loss functions are targeted at estimating Q_0 and g_0 efficiently. Our primary concern in estimating Q_0 and g_0, however, lies in estimating the risk of a given candidate estimator, which is a particular functional of Q_0 and g_0. We therefore believe that more efficient estimators of ψ_0 can be obtained by targeting the loss functions for estimating Q_0 and g_0 at this particular functional rather than at Q_0 and g_0 themselves. We are currently investigating this question by applying the estimating function based cross-validation methodology proposed by van der Laan and Rubin (2005).

Lastly, we would like to point out that some causal inference questions may be formulated more directly by considering parameters other than the treatment-specific mean. In particular, many causal analyses are concerned with estimating the causal effect of a treatment A on an outcome Y conditional on some baseline covariates V, but are not interested *per se* in understanding how Y is affected by changes in V alone. In interpreting the final model fit for the SPPARCS analysis, for example, we summarized the causal effect of A on Y conditional on V by collecting terms that contain A and evaluating them at different values of V. Terms containing only stratification variables V were of little interest. If the treatment variable A allows for the definition of an appropriate reference level $A = 0$, such causal analyses may be framed more directly by considering a causal effect parameter such as $\Psi(P)(a, V) = E_P[Y(a) - Y(0) \mid V]$.

Such parameters have two main advantages over the treatment-specific mean. First, since analyses based on the treatment-specific mean are not targeted directly at the question of interest, they run the risk of yielding results that in fact shed very little light on that question. This is for example the case if none of the terms in the data-adaptively chosen model contain the treatment variable. In such instances, one would be forced to conclude that the analysis suggests no causal effect of A on Y, but obtains no direct estimate of that causal effect. Second, the parameter space for $\Psi(P)(a, V) = E_P[Y(a) - Y(0) \mid V]$ does not include those functions of a and V that are constant in a and hence is considerably smaller than the parameter space for the treatment-specific mean. We therefore expect that causal effect parameters can be estimated more precisely than the treatment-specific mean.

We are currently working on implementing the variable importance methodology developed by van der Laan (2005) to estimate $\Psi(P)(a, V) = E_P[Y(a) - Y(0) \mid V]$ data-adaptively and hope to incorporate this approach in future versions of our `R` package `cvDSA`.

References

H. Akaike. *Information theory and an extension of the maximum likelihood principle*. Academiai Kiado, 1973.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:310–413, 1999.

A.R. Barron. Statistical properties of artificial neural networks. In *Proceedings of the 28th conference on decision theory and control, Tampa, Floria*, 1989.

A.R. Barron. *Nonparametric functional estimation and related topics*, chapter Complexity regularization with application to artificial neural networks, pages 561–576. Kluwer Academic Publishers, the Netherlands, 1991.

L.M. LeCam. *Asymptotic Methods in Statistical Decision Theory*. Springer Verlag, New York, 1986.

L.M. LeCam and G. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Verlag, New York, 1990.

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

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

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

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000b.

J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(0):143–155, 1992.

D.B. Rubin. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6:34–58, 1978.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

X. Shen. On methods of sieves and penalization. *Annals of Statistics*, 25:2555–2591, 1997.

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

I. Tager, M. Hollenberg, and W. Satariano. Self-reported leisure time physical activity and measures of cardiorespiratory fitness in an elderly population. *American Journal of Epidemiology*, 147:921–931, 1998.

M.J. van der Laan. Statistical inference for variable importance. Technical report 188, Division of Biostatistics, University of California, Berkeley, August 2005.

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

M.J. van der Laan and M.L. Petersen. Estimation of direct and indirect causal effects in longitudinal studies. Technical report, University of California, Berkeley, Division of Biostatistics, 2004.

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

M.J. van der Laan and D. Rubin. Estimating function based cross-validation and learning. Technical report 180, Division of Biostatistics, University of California, Berkeley, May 2005.

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