The International Journal of Biostatistics

Volume 2, *Issue* 1 2006 *Article* 11

Targeted Maximum Likelihood Learning

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

Recommended Citation:
van der Laan, Mark J. and Rubin, Daniel (2006) "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 11.
DOI: 10.2202/1557-4679.1043

Targeted Maximum Likelihood Learning

Mark J. van der Laan and Daniel Rubin

Abstract

Suppose one observes a sample of independent and identically distributed observations from a particular data generating distribution. Suppose that one is concerned with estimation of a particular pathwise differentiable Euclidean parameter. A substitution estimator evaluating the parameter of a given likelihood based density estimator is typically too biased and might not even converge at the parametric rate: that is, the density estimator was targeted to be a good estimator of the density and might therefore result in a poor estimator of a particular smooth functional of the density. In this article we propose a one step (and, by iteration, k-th step) targeted maximum likelihood density estimator which involves 1) creating a hardest parametric submodel with parameter epsilon through the given density estimator with score equal to the efficient influence curve of the pathwise differentiable parameter at the density estimator, 2) estimating epsilon with the maximum likelihood estimator, and 3) defining a new density estimator as the corresponding update of the original density estimator. We show that iteration of this algorithm results in a targeted maximum likelihood density estimator which solves the efficient influence curve estimating equation and thereby yields a locally efficient estimator of the parameter of interest, under regularity conditions. In particular, we show that, if the parameter is linear and the model is convex, then the targeted maximum likelihood estimator is often achieved in the first step, and it results in a locally efficient estimator at an arbitrary (e.g., heavily misspecified) starting density.

We also show that the targeted maximum likelihood estimators are now in full agreement with the locally efficient estimating function methodology as presented in Robins and Rotnitzky (1992) and van der Laan and Robins (2003), creating, in particular, algebraic equivalence between the double robust locally efficient estimators using the targeted maximum likelihood estimators as an estimate of its nuisance parameters, and targeted maximum likelihood estimators. In addition, it is argued that the targeted MLE has various advantages relative to the current estimating function based approach. We proceed by providing data driven methodologies to select the initial density estimator for the targeted MLE, thereby providing data adaptive targeted maximum likelihood estimation methodology. We illustrate the method with various worked out examples.

KEYWORDS: causal effect, cross-validation, efficient influence curve, estimating function, locally efficient estimation, loss function, maximum likelihood estimation, sieve, targeted maximum likelihood estimation, variable importance

Author Notes: This research was supported by NIH grant R01 GM071397.

1 Introduction

Let $O_1,\ldots,O_n$ be n independent and identically distributed (i.i.d.) observations of an experimental unit O with probability distribution $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the statistical model. For the sake of presentation, we will assume that $\mathcal{M}$ is dominated by a common measure μ so that we can identify each possible probability measure $P \in \mathcal{M}$ by its density $p = dP/d\mu$. In the discussion we point out that our methods are not restricted to models dominated by a single measure. Let P_n be the empirical probability distribution of $O_1,\ldots,O_n$ which puts mass $1/n$ on each of the n observations. Let $p_0 = \frac{dP_0}{d\mu}$ be the density of p_0 with respect to a dominating measure μ, and let p_n be a density estimator of p_0. For example, $p_n \equiv \Phi(P_n)$ could be the maximum likelihood estimator defined by the following mapping Φ

$$p_n = \Phi(P_n) \equiv \arg\max_{P\in\mathcal{M}} \sum_{i=1}^n \log \frac{dP}{d\mu}(O_i).$$

Alternatively, if the model $\mathcal{M}$ is too large in the sense that the maximum likelihood estimator is too variable or even inconsistent, then one typically proposes a sieve $\mathcal{M}_s \subset \mathcal{M}$, indexed by indices s, approximating $\mathcal{M}$, and computes candidate maximum likelihood estimators

$$p_{ns} = \Phi_s(P_n) \equiv \arg\max_{P\in\mathcal{M}_s} \sum_{i=1}^n \log \frac{dP}{d\mu}(O_i).$$

In such a setting it remains to data adaptively select s. For example, one could use likelihood based cross-validation to select s:

$$s_n = \arg\max_s E_{B_n} \sum_{i:B_n(i)=1} \log \Phi_s(P^0_{n,B_n})(O_i),$$

where $B_n \in \{0,1\}^n$ is a random vector of binary variables defining a random split in a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, and P^0_{n,B_n}, P^1_{n,B_n} denote the empirical probability distributions of the training and validation sample, respectively. Now, one would define the estimator of p_0 as the cross-validated maximum likelihood estimator given by

$$p_n = \Phi(P_n) \equiv p_{ns_n} = \Phi_{s_n}(P_n).$$

It is common practice to evaluate one or many Euclidean valued smooth functionals $\Psi(p_n)$ of the density estimator p_n and view them as estimators of the parameter $\Psi(p_0)$ for given parameter mappings $\Psi : \mathcal{M} \to \mathbb{R}^d$. Although this method is known to result in efficient estimators of $\Psi(p_0)$ in parametric models (i.e., $\mathcal{M}$ in the above definition of p_n is a parametric model), in general, such substitution estimators are not correctly trading off bias and variance with respect to the parameter of interest $\psi_0 = \Psi(p_0)$. For example, a univariate (standard) kernel density estimator optimizing the mean squared error with respect to p_0, assuming a continuous second derivative, can have bias of the order $n^{-2/5}$ based on an optimal bandwidth of the order $n^{-1/5}$. The corresponding substitution estimator of the cumulative distribution function at a point can have bias which converges to zero at the same rate $n^{-2/5}$, but a variance of $O(1/n)$, so that the substitution estimator has a variance $(1/n)$ which is smaller than the square bias $(n^{-4/5})$ by an order of magnitude. In particular, the smoothed empirical cumulative distribution functions would not even converge at root-n rate due to the fact that $\sqrt{n}$ times the bias $n^{-2/5}$ does not converge to zero: that is, in this kernel density estimator example $\sqrt{n}n^{-2.5} \to \infty$, so that the relative efficiency of the empirical cumulative distribution function and this smooth cumulative distribution function converges to zero. This shows that substitution estimators based on optimal (*for the purpose of the density itself*) density estimators of the cumulative distribution function are typically theoretically inferior to other more targeted estimators of the parameter of interest. In general, substitution estimators based on density estimators might simply not be very good estimators, and, in particular, likelihood based substitution estimators will often fail to be asymptotically efficient due to the bias caused by the curse of dimensionality: the kernel density example already shows the failure of likelihood based learning of smooth parameters of a density of a univariate random variable, and it gets much worse for densities of multivariate random variables. This issue has been stressed repeatly by Robins and co-authors (see e.g., Robins and Rotnitzky (1992) and van der Laan and Robins (2003)). This article proposes a method which, given a particular pathwise differentiable parameter of interest, allows one to map a density estimator (such as p_n or p_{ns} for each s) into a targeted maximum likelihood density estimator so that the corresponding substitution estimator of ψ_0 is locally efficient, under reasonable conditions: that is, if the starting density estimator is consistent, it will typically be efficient, and otherwise in certain classes of problems it might still be consistent and asymptotically linear.

Specifically, in this article we propose a one step maximum likelihood density estimator which involves 1) creating a parametric model with Euclidean parameter ϵ (e.g., the same dimension d as the parameter ψ_0) through a given density estimator p_n^0 (e.g., s-specific MLE p_{ns}) at $\epsilon = 0$ whose scores include the components of the efficient influence curve of the pathwise differentiable parameter at the density estimator p_n^0, 2) estimating ϵ with the maximum likelihood estimator of this parametric model, and 3) defining a new density estimator p_n^1 as the corresponding fluctuation of the original density estimator p_n^0. In addition, iterating this process results in a sequence of p_n^k with increasing log-likelihood converging to a solution of the efficient influence curve estimating equation, and thereby typically results in a locally efficient substitution estimator of ψ_0. We refer to this solution as the targeted maximum likelihood estimator based on the initial p_n^0. We provide various examples in which this targeted maximum likelihood estimator is achieved at the first step of the algorithm.

In particular, one can map each model based MLE p_{ns} into a targeted MLE p_{ns}^* (targeted towards ψ_0). We suggest that it is appropriate to select among this collection of targeted MLEs p_{ns}^* with likelihood based cross-validation, as explained heuristically in our accompanying technical report: targeted MLE's are comparable w.r.t. to being fully trained w.r.t. estimation of the parameter of interest, which makes the log-likelihood an appropriate criteria to select among them. That is, let $p_{ns}^* = \hat{\Phi}_s^*(P_n)$ be the s-specific targeted MLE applied to the initial density estimator p_{ns}. Let

$$s_n = \arg\max_s E_{B_n} \sum_{i:B_n(i)=1} \log \hat{\Phi}_s^*(P_{n,B_n}^0)(O_i),$$

where $B_n \in \{0,1\}^n$ is a random vector of binary variables defining a random split in a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, and P_{n,B_n}^0, P_{n,B_n}^1 denote the empirical probability distributions of the training and validation sample, respectively, as above. Now, likelihood cross-validated targeted MLE is defined as:

$$p_n^* = \hat{\Phi}(P_n) \equiv p_{ns_n}^* = \hat{\Phi}_{s_n}^*(P_n).$$

We also note that the candidate models indexed by s can be chosen to represent a sieve in a possibly misspecified (big) model $\mathcal{M}$, as long as this model $\mathcal{M}$ is still such that the Kullback-Leibler projection of the true density p_0 on this model identifies the parameter of interest $\Psi(p_0)$ correctly: for example, if the parameter of interest is a parameter of a regression of an outcome Y on covariates W, then one might select as big model the normal densities with unspecified conditional mean, given W, and certain possibly misspecified conditional variance, even though the true density p_0 is not a member of this model.

1.1 Organization of article.

In Section 2, given an initial density estimator p_n^0 (e.g., p_{ns}) of p_0, we formally define the k-th order targeted maximum likelihood density estimator p_n^k, and corresponding targeted maximum likelihood estimator $\Psi(p_n^k)$ of ψ_0. We illustrate the targeted MLE of the cumulative distribution function at a point in a nonparametric model. In this case, it appears that the first step targeted MLE of ψ_0 algebraically equals the empirical cumulative distribution function, for any given initial density estimator p_n^0. Thus, while the original substitution estimator of the cumulative distribution function would not converge at the parametric rate $1/\sqrt{n}$ due to it being too biased, the first order targeted bias corrected density estimator estimates the cumulative distribution function efficiently. In Section 3 we establish that the targeted MLE solves the efficient influence curve estimating equation, which provides the basis of its asymptotic efficiency for ψ_0. In Section 4 we present general templates for establishing consistency, asymptotic linearity and efficiency of the targeted MLE of ψ_0, which provides a particular powerful theorem for convex models and linear pathwise differentiable parameters stating that the targeted MLE will be consistent and asymptotically linear for an arbitrary starting density, and it will be efficient if the starting (or its targeted MLE version) density consistently estimates the efficient influence curve. We illustrate the latter result with two examples. In Section 5 we discuss the relation, and in particular, the algebraic equivalence, between targeted maximum likelihood estimation and estimating function based estimation if one estimates the nuisance parameters in the estimating functions with the targeted MLE. We point out that targeted MLE is more widely applicable by not relying on being able to map the efficient influence curve in a corresponding estimating function, and it deals naturally with the issue of multiple solutions of estimating equations. In Subsection 5.1 we focus on censored data models to make the comparison with the estimating function methodology in van der Laan and Robins (2003). In particular, we present the targeted MLE approach which results in algebraic equivalence between the Inverse Probability of Censoring Weighted estimator, the double robust IPCW estimator, and the targeted MLE of a parameter of the full data distribution based on observing n i.i.d. observations of a censored data structure under coarsening at random (CAR). These results show that the targeted MLE does not only provide a boost for likelihood based estimation, but it also provides an improvement relative to the current implementation of locally efficient estimation based on estimating function methodology. In Section 6 we present important examples illustrating the power and computational simplicity of this new targeted maximum likelihood estimator: estimation of a marginal causal effect, and the parametric component in a semiparametric regression model, and we present a simulation to illustrate the targeted MLE. In Section 7 we present a loss based approach of targeted MLE learning based on the unified loss function based approach in van der Laan and Dudoit

(2003). We end this article with a discussion in Section 8. In our accompanying technical report we show generalizations of the targeted MLE of pathwise differentiable parameters to targeted MLE of general parameters.

1.2 Some relevant literature overview.

There exist various methods for construction of an efficient estimator of a parameter based on parametric models. In particular, Fisher's method of maximum likelihood estimation can be applied, or closely related M-estimate (i.e., estimators defined as solutions of estimating equations) methods which work under minimal conditions. Maximum likelihood estimation in semiparametric models has been an extensive research area of interest. Here we suffice with a referral to van der Vaart and Wellner (1996b) for a partial overview of the theory for the analysis of maximum likelihood. There are plenty of examples in which the straightforward semiparametric MLE even fails to be consistent, but often an appropriate regularization can be applied to repair the consistency of the semiparametric MLE: e.g., see van der Laan (1995) for such examples based on censored data. However, as argued above in the kernel density estimator example, maximum likelihood based smoothing/model selection will often provide the wrong trade-off of bias and variance for specific smooth parameters. The literature (notably Robins and co-authors) has recognized this problem with likelihood based estimation. For example, smoothing survival functions or smoothing the nonparametric components in a semiparametric regression model requires so called "under-smoothing" in order to obtain root-n consistency for the parameter of interest: see e.g., Cosslett (2004).

For an overview of the literature on efficient estimation of pathwise differentiable parameters in semiparametric models we refer to Bickel et al. (1993b). In particular, the latter presents the general one step estimator based on an estimate of the efficient influence curve: see e.g. Klaassen (1987). For an overview of the literature on locally efficient estimating function based estimation of pathwise differentiable parameters based on censored longitudinal data (starting with the ground breaking paper Robins and Rotnitzky (1992)), we refer to van der Laan and Robins (2003).

A unified loss function approach based methodology for estimation and estimator selection, and concrete illustration of this method in various examples is presented in van der Laan and Dudoit (2003). This methodology is general by allowing the loss function to be an unknown function of the experimental unit and the parameter values. van der Laan and Rubin (2005) and van der Laan and Rubin (2006) present an alternative unified estimating function methodology for both estimation and estimator selection. The latter two methodologies provide two general strategies for data adaptive estimation of any parameter in any model.

We note that these (unified) loss function and (unified) estimating function based approaches give up on using the log-likelihood as loss function for the purpose of estimator selection and estimation when the parameter of interest is not the actual density of the data, but a particular parameter of it: these methods replace the log-likelihood loss function by a loss function or an estimating function targeted at the parameter of interest. From that point of view, the current article shows that it is not necessary to replace the log-likelihood loss function by a targeted loss function, but that one can also target the directions in which one maximizes the log-likelihood.

2 Targeted maximum likelihood estimators.

Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ be a pathwise differentiable parameter at any density $p \in \mathcal{M}$, where $\mathcal{M}$ denotes the statistical model consisting of the possible densities $p = dP/d\mu$ of O with respect to some dominating measure μ. That is, given a sufficiently rich class of one-dimensional regular parametric submodels $\{p_\delta : \delta\}$ with parameter δ of $\mathcal{M}$ through the density p at $\delta = 0$, we have for each of these submodels p_δ with score s at $\delta = 0$ and $p_{\delta=0} = p$

$$\frac{d}{d\delta} \left. \Psi(p_\delta) \right|_{\delta=0} = E_p S(p)(O)s(O)$$

for some $S(p) \in (L_0^2(p))^d$, where $L_0^2(p)$ denotes the Hilbert space of functions of O with mean 0 and finite variance under P, endowed with inner product $\langle h_1, h_2 \rangle_P = E_p h_1(O) h_2(O)$. This random variable $S(p) \in (L_0^2(p))^d$ is called a gradient of the pathwise derivative at p. Let $T(p) \subset L_0^2(p)$ be the tangent space at p which is defined as the closure of the linear span of the scores s of this class of submodels through p. If the model is not locally saturated in the sense that $T(p) = L_0^2(p)$, then there can be many gradients. Let $T_{nuis}^{\perp}(p) \subset L_0^2(p)$ be the orthogonal complement of the so called nuisance tangent space, where the latter is defined as the closure of the linear span of all scores of p_δ for which the pathwise derivative equals 0 (see van der Laan and Robins (2003), Chapter 1). As in van der Laan and Robins (2003), we denote the set of gradients at p with $T_{nuis}^{\perp *}(p) \subset (T_{nuis}^{\perp}(p))^d$. Let $S^*(p)$ be the so called canonical gradient which is the unique gradient whose d components $S^*(p)_j$, $j = 1, \ldots, d$, are elements of the tangent space $T(P)$. A submodel $\{p_\epsilon : \epsilon\}$ with score $S^*(p)$ at $\epsilon = 0$ is often referred to as a hardest submodel (Bickel et al. (1993a)), as we will also do in this article.

Let $(O, p) \rightarrow D(p)(O)$ be a point-wise well defined class of functions on the Cartesian product of the support of O and the model $\mathcal{M}$, which satisfies

$$D(p) = S^*(p) \ P_0\text{-a.e. for all } p \in \mathcal{M}.$$

As an example, consider letting O be a Euclidean valued d-variate random variable with density p_0. Let $\mathcal{M}$ be the class of all continuous densities with respect to Lebesgue measure μ, and let $\Psi(p) = \int_0^t p(o) d\mu(o)$ be the cumulative distribution function at a point $t \in \mathbb{R}$ corresponding with density p. In this case $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ is pathwise differentiable parameter at p with efficient influence curve $S(p)(O) = I(O \leq t) - \Psi(p)$, and, because the model is locally saturated, it is also the only influence curve/gradient. So $D(p) = I(O \leq t) - \Psi(p)$. Similarly, given a set of user supplied points $\{t_1, \ldots, t_d\}$, we could define the d-dimensional Euclidean parameter $\Psi(p) = (\Psi(p)(t_j) \equiv \int_0^{t_j} p(o) d\mu(o) : j = 1, \ldots, d)$ representing the cumulative distribution function at d points. In this case, $D(p) = (I(O \leq t_j) - \Psi(p)(t_j) : j = 1, \ldots, d)$ has d components.

A general methodology for construction of functions $D_h(p)$ indexed by an $h \in \mathcal{H}$ so that $\{D_h(p) : h \in \mathcal{H}\} \subset T_{nuis}^{\perp}(p)$ (or equality) is presented in van der Laan and Robins (2003). In van der Laan and Robins (2003) the class of functions $\{D_h(p) : h \in \mathcal{H}\}$ is referred to as a representation of the orthogonal complement of the nuisance tangent space, which is then used to map into a class of corresponding estimating functions for the pathwise differentiable parameter $p \rightarrow \Psi(p)$ of the form $p \rightarrow D_h(\Psi(p), \Upsilon(p))$ with Υ representing a nuisance parameter. In van der Laan and Robins (2003), for a variety of general classes of models and censored data structures O, explicit representations of the orthogonal complement of the nuisance tangent space, $T_{nuis}^{\perp}(p)$, corresponding gradients, $T_{nuis}^{\perp *}(p)$, and canonical gradient $S^*(p)$, have been provided.

Let $p_n^0 = \Phi(P_n) \in \mathcal{M}$ be a density estimator of $p_0 = dP_0/d\mu$. Define now a parametric submodel $\{p_n^0(\epsilon) : \epsilon \in \mathbb{R}^k\} \subset \mathcal{M}$ through p_n^0 at $\epsilon = 0$ whose linear span of scores of ϵ at $\epsilon = 0$ includes all d components of $D(p_n)$. One possibility is to choose $\epsilon \in \mathbb{R}^d$ of the same dimension as $D(p)$ and arrange that the score of ϵ_j at $\epsilon = 0$ equals $D_j(p)$, $j = 1, \ldots, d$. For example, if the model $\mathcal{M}$ is convex then the following model typically applies

$$p_n^0(\epsilon) \equiv (1 + \epsilon^{\top} D(p_n^0)) p_n^0, \tag{1}$$

where $\epsilon \in \mathbb{R}^d$ denotes the parameter ranging over all values for which $p_n^0(\epsilon)$ is a proper density. Note that indeed $p_n^0(0) = p_n^0$, $p_n^0(\epsilon)$ is a density (positive valued and integrates till 1) for ϵ small enough, and $\frac{d}{d\epsilon} \log p_n^0(\epsilon)\Big|_{\epsilon=0} = D(p_n^0)$.

One can also use an exponential family

$$p_n^0(\epsilon) \equiv C(\epsilon, p_n^0)\exp(\epsilon^\top D(p_n^0))p_n^0$$

for $C(\epsilon, p_n^0)$ be a normalizing constant. In general, one can choose a parameterization $\epsilon \rightarrow p_n^0(\epsilon) \in \mathcal{M}$ which is smooth in ϵ at $\epsilon = 0$ and whose score at $\epsilon = 0$ equals $D(p_n^0)$. However, we will also consider submodels $p_n^0(\epsilon)$ with additional scores in order to arrange that the targeted MLE will be fully targeted towards estimation of $D(p_0)$.

Let

$$\epsilon_n = \epsilon(P_n \mid p_n^0) \equiv \arg \max_{\{\epsilon : p_n^0(\epsilon) \in \mathcal{M}\}} \sum_{i=1}^n \log p_n^0(\epsilon)(O_i)$$

be the maximum likelihood estimator of ϵ treating the density estimator p_n^0 as given and fixed. We will assume that the maximum is attained in the interior of $\mathcal{M}$ so that ϵ_n solves the estimating equation:

$$0 = P_n \frac{\frac{d}{d\epsilon} p_n^0(\epsilon)}{p_n^0(\epsilon)}.$$

Here we use the common notation $Pf \equiv \int f(o)dP(o)$. For example, if $p_n^0(\epsilon) = (1 + \epsilon^\top D(p_n^0))p_n^0$, as one might choose in convex models, then we have that ϵ_n is the solution of

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{D(p_n^0)(O_i)}{1 + \epsilon_n^\top D(p_n^0)(O_i)}.$$

This defines now an updated density estimator

$$p_n^1 \equiv p_n^0(\epsilon_n) = p_n^0(\epsilon(P_n \mid p_n^0)) \in \mathcal{M}.$$

Note that this simply defines a method for mapping an initial density estimator $p_n^0 \in \mathcal{M}$ in a new density estimator $p_n^1 \in \mathcal{M}$, which we call the first step targeted maximum likelihood estimator. By iterating this process one obtains the k-step targeted maximum likelihood estimator p_n^k, $k = 1, \ldots$

Definition 1 *Given an initial density estimator $p_n^0 = \hat{\Phi}^0(P_n)$ based on the empirical probability distribution P_n, a parametric fluctuation $\{p_n^0(\epsilon) : \epsilon\} \subset \mathcal{M}$ satisfying $p_n^0(0) = p_n^0$, and $\frac{d}{d\epsilon} \log p_n^0(\epsilon)\Big|_{\epsilon=0} = D^*(p_n^0)$, where the linear span of the components of $D^*(p_n^0)$ include all d components of a canonical gradient $D(p_n^0)$ of the parameter of interest $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ at p_n^0, a maximum likelihood estimator*

$$\epsilon(P_n \mid p_n^0) \equiv \arg\max_\epsilon \sum_{i=1}^n \log p_n^0(\epsilon)(O_i)$$

of ϵ, we define the first step targeted maximum likelihood density estimator as

$$p_n^1 = \hat{\Phi}^1(P_n) \equiv p_n^0(\epsilon(P_n \mid p_n^0)).$$

This process can be iterated to define the k-step targeted maximum likelihood density estimator as

$$p_n^{k+1} = \hat{\Phi}^{k+1}(P_n) \equiv p_n^k(\epsilon(P_n \mid p_n^k)),\ k = 0, 1, \ldots.$$

The corresponding k-step targeted maximum likelihood estimator of ψ_0 is defined as

$$\hat{\Psi}_k(P_n) = \Psi(p_n^k).$$

The targeted maximum likelihood estimator is defined as

$${}_n = \hat{\Phi}^*(P_n) \equiv \lim_{k\to\infty} \Psi(p_n^k),$$

assuming this limit exists.

2.1 Example: Estimating the CDF.

Consider an initial data generating density $p^0 = f$, let $F(t) = \int_{-\infty}^t f(o)do$ denote the associated CDF at some fixed point $t \in \mathbb{R}$, and consider the parametric model

$$\left\{ f_\epsilon(o) = (1 + \epsilon[I(o \leq t) - F(t)])f(o) : -\frac{1}{1 - F(t)} \leq \epsilon \leq \frac{1}{F(t)} \right\}, \qquad (2)$$

where one can check that the range restraint on ϵ serves merely to ensure that the family is indeed a proper class of densities. Consider estimating ϵ from maximum likelihood based on an i.i.d. sample $\{O_i\}_{i=1}^n$. The log likelihood is,

$$l(\epsilon) = \sum_{i=1}^n \log(1 + \epsilon[I(O_i \leq t) - F(t)]) + \sum_{i=1}^n \log f(O_i). \qquad (3)$$

Its derivative is,

$$l'(\epsilon) = \sum_{i=1}^n \frac{I(O_i \leq t) - F(t)}{1 + \epsilon[I(O_i \leq t) - F(t)]}. \qquad (4)$$

Its second derivative is easily seen to be,

$$l''(\epsilon) = -\sum_{i=1}^n \left\{ \frac{I(O_i \leq t) - F(t)}{1 + \epsilon[I(O_i \leq t) - F(t)]} \right\}^2. \qquad (5)$$

Because the log likelihood is concave, we know that the maximum is achieved if $l'(\epsilon) = 0$ has a solution. Letting $F_n(\cdot)$ denote the empirical distribution function, note that we can decompose the terms in $l'(\epsilon)$ into two parts (those for which $I(O_i \leq t)$ are 0 or 1), and the MLE of ϵ can be seen to solve,

$$\begin{aligned} 0 &= l'(\epsilon) \\ &= \sum_{i=1}^{n} \frac{I(O_i \leq t) - F(t)}{1 + \epsilon[I(O_i \leq t) - F(t)]} \\ &= nF_n(t)\frac{1 - F(t)}{1 + \epsilon[1 - F(t)]} + n(1 - F_n(t))\frac{-F(t)}{1 - \epsilon F(t)}. \end{aligned}$$

Moving the second term on the right to the other side of the equation, dividing both sides by n, and multiplying both sides by $(1 + \epsilon[1 - F(t)])(1 - \epsilon F(t))$, the equation reduces to,

$$F_n(t)(1 - F(t))(1 - \epsilon F(t)) = (1 - F_n(t))F(t)(1 + \epsilon(1 - F(t))). \tag{6}$$

This is linear in ϵ, and one can check that the solution is

$$\begin{aligned} \epsilon_n &= \frac{F_n(t)(1 - F(t)) - (1 - F_n(t))F(t)}{F(t)(1 - F(t))} \\ &= \frac{F_n(t) - F_n(t)F(t) - F(t) + F_n(t)F(t)}{F(t)(1 - F(t))} \\ &= \frac{F_n(t) - F(t)}{F(t)(1 - F(t))}. \end{aligned} \tag{7}$$

Because $0 \leq F_n(t) \leq 1$, one can check that indeed

$$-\frac{1}{1 - F(t)} = -\frac{F(t)}{F(t)(1 - F(t)} \leq \epsilon_n \leq \frac{1 - F(t)}{F(t)(1 - F(t))} = \frac{1}{F(t)}, \tag{8}$$

so the range restraint on ϵ for the family (2) always holds for the maximum likelihood estimator, meaning that $f_{\epsilon_n}(\cdot)$ is a proper density. Now, the resulting CDF at t for this density is then,

$$\begin{aligned} F_{\epsilon_n}(t) &= \int_{-\infty}^{t} f_{\epsilon_n}(o)do \\ &= \int_{-\infty}^{t} (1 + \epsilon_n[I(o \leq t) - F(t)])f(o)do \\ &= \int_{-\infty}^{t} f(o)do + \epsilon_n \int_{-\infty}^{t} I(o \leq t)f(o)do - \epsilon_1 F(t) \int_{-\infty}^{t} f(o)do \end{aligned}$$

$$\begin{aligned}
&= F(t)+\epsilon_n F(t)-\epsilon_n F(t)^2 = F(t)+\epsilon_1 F(t)(1-F(t)) \\
&= F(t)+\frac{F_n(t)-F(t)}{F(t)(1-F(t))}F(t)(1-F(t)) \text{ from (7)} \\
&= F(t)+F_n(t)-F(t)=F_n(t).
\end{aligned}$$

Therefore, for any initial density $f(\cdot)$ and any time point t, the targeted likelihood maximum likelihood estimator of the CDF reduces to the empirical distribution estimator in a single step. This result immediately generalizes to $\Psi(p)=\int_A p(o)d\mu(o)$ for any measurable set A.

3 Solving the efficient estimating equation.

We have the following trivial, but useful result. It states that if the MLE's $\epsilon(P_n \mid p_n^k)$ at step k of the targeted MLE algorithm converge to zero for $k \rightarrow \infty$ (as one expects to hold if the log likelihood of the data is uniformly bounded in the model $\mathcal{M}$), then the algorithm converges to a solution of the efficient influence curve equation $P_n D(p)=0$ in the sense that $P_n D(p_n^k) \rightarrow 0$.

Result 1 *Let P_n be given. Assume that*

$$\lim_{\epsilon\rightarrow 0}\limsup_{k\rightarrow\infty} \mid P_n \frac{\frac{d}{d\epsilon}p_n^k(\epsilon)}{p_n^k(\epsilon)} - P_n \frac{p_n^{k\prime}(0)}{p_n^k(0)} \mid \rightarrow 0, \qquad (9)$$

that for each k there exist a constant matrix A_k so that $A_k \frac{p_n^{k\prime}}{p_n^k} = D(p_n^k)$ with $\limsup_{k\rightarrow\infty} \parallel A_k \parallel < \infty$, where $\parallel A \parallel$ denotes a matrix norm.

If $\epsilon(P_n \mid p_n^k)$ solves $P_n \frac{\frac{d}{d\epsilon}p_n^k(\epsilon)}{p_n^k(\epsilon)} = 0$ for all k, and $\epsilon(P_n \mid p_n^k) \rightarrow 0$ for $k \rightarrow \infty$, then we have

$$P_n D(p_n^k) \rightarrow 0 \text{ for } k \rightarrow \infty.$$

The condition (9) holds if the score of the one-dimensional submodel $p(\epsilon)$ at ϵ converges to the score at $\epsilon=0$ for $\epsilon \rightarrow 0$ uniformly in a set containing the k-step targeted MLE's p_n^k, $k=1,2,\ldots$, and that for each $p \in \mathcal{M}$, the linear span of the components $\frac{p'(0)}{p(0)}$ includes the components of $D(p)$. Since the likelihood increases at each step one might indeed expect that typically the targeted MLE algorithm will converge and thereby that $\epsilon(P_n \mid p_n^k) \rightarrow 0$. That is, Result 1 essentially states that, if the targeted MLE algorithm converges, then the algorithm will converge to a solution of the efficient influence curve equation in the sense that by choosing k large enough $P_n D(p_n^k) \approx 0$ with arbitrary small deviation from 0.

Proof. Let $\epsilon_k = \epsilon(P_n \mid p_n^k)$, $k = 0, \ldots$. If $\epsilon_k \to 0$ for $k \to \infty$, then

$$P_n \frac{\frac{d}{d\epsilon_k} p_n^k(\epsilon_k)}{p_n^k(\epsilon_k)} - P_n \frac{p_n^{k\prime}(0)}{p_n^k(0)} \to 0$$

for $k \to \infty$. Let A_k be such that $A_k \frac{p_n^{k\prime}(0)}{p_n^k(0)} = D(p_n^k)$. By assumption, the matrix has a norm bounded uniformly in k. Thus, we also have

$$P_n A_k \frac{\frac{d}{d\epsilon_k} p_n^k(\epsilon_k)}{p_n^k(\epsilon_k)} - P_n D(p_n^k) \to 0$$

for $k \to \infty$. However, $P_n \frac{d}{d\epsilon_k} p_n^k(\epsilon_k)/p_n^k(\epsilon_k) = 0$ (and thus A_k applied to this equals 0 as well), which shows that $P_n D(p_n^k) \to 0$. □

4 Efficiency of targeted likelihood estimation.

In this section we provide templates for proving consistency, asymptotic linearity and efficiency of the targeted maximum likelihood estimator of a path-wise differentiable parameter. Since convexity of the model and linearity of the parameter allows a particular strong result, we separate this situation from the general case.

4.1 Linear parameters in convex models.

Let p_n^∞ denote the limit of our algorithm if it exists as a density with respect to μ in $\mathcal{M}$, and otherwise it represents a $p_n^k \in \mathcal{M}$ for a large enough k. If the condition of the above Result 1 holds, then $p_n^\infty \in \mathcal{M}$, and for all practical purposes, we have $P_n D(p_n^\infty) = 0$. If this is true, then this result can be used to establish efficiency of the substitution estimator $\Psi(p_n^\infty)$ as an estimator of ψ_0 under the assumption that the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ is linear and $\mathcal{M}$ is convex, under weak regularity conditions. Specifically, by the identity for convex models and linear parameters in van der Laan (1998) we have $\Psi(p) - \Psi(p_0) = -P_0 D(p)$ for any $p, p_0 \in \mathcal{M}$ for which $p_0/p < \infty$. Thus, if $p_n^\infty \in \mathcal{M}$ and it is bounded away from 0 on the support of p_0, then combining $P_n D(p_n^\infty) = 0$ with the latter identity gives us

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0) D(p_n^\infty). \tag{10}$$

Even if p_n^∞ does not satisfy $p_0/p_n^\infty < \infty$, then the identity $\Psi(p_n^\infty) - \Psi(p_0) = -P_0 D(p_n^\infty)$ can still be established under a continuity condition on $p \to P_0 D(p)$

(see van der Laan (1998)), so that (10) can even be established for density estimators not satisfying this support condition.

Applying empirical process theory (van der Vaart and Wellner (1996a)) now proves that $\Psi(p_n^\infty)$ is root-n consistent if $D(p_n^\infty)$ falls in a P_0 Donsker class with probability tending to 1. If one can now also establish that $P_0(D(p_n^\infty)-D(p_1))^2$ converges to zero in probability for a certain $p_1 \in \mathcal{M}$, then it follows that $\Psi(p_n^\infty)$ is asymptotically linear with influence curve $D_0(p_1) \equiv D(p_1)-P_0D(p_1)$:

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D_0(p_1) + o_P(1/\sqrt{n}),$$

where we note that p_1 can be an arbitrary limit (i.e., $p_1 \not= p_0$ is allowed). In particular, if the limit p_1 is such that $D(p_1) = D(p_0)$, then $\Psi(p_n^\infty)$ is asymptotically linear with influence curve $D(p_0)$. Thus, if $D(p_0)$ is the efficient influence curve, then $\Psi(p_n^\infty)$ is asymptotically efficient.

Theorem 1 *Suppose the conclusion of Result 1 holds, and $K = K(n)$ is chosen large enough so that the targeted MLE $p_n = p_n^K$ satisfies $P_nD(p_n) = R(n, K(n)) = o_P(1/\sqrt{n})$ (where $\lim_{K\to\infty} R(n, K) = 0$). Assume that $p_n \in \mathcal{M}$, $p_0/p_n < \infty$ uniformly over a support of p_0, $\mathcal{M}$ is convex, and $\Psi : \mathcal{M} \to \mathbb{R}^d$ is linear. Then*

$$\Psi(p_n) - \Psi(p_0) = (P_n - P_0)D(p_n) + R(n, K(n)).$$

If $D(p_n)$ falls in a P_0 Donsker class with probability tending to 1, then

$$\Psi(p_n) - \psi_0 = O_P(1/\sqrt{n}).$$

If it is also shown that $P_0(D(p_n) - D(p_1))^2 \to 0$ in probability for $n \to \infty$ for some $p_1 \in \mathcal{M}$, then it follows that $\Psi(p_n)$ is asymptotically linear with influence curve $D(p_1) - P_0D(p_1)$:

$$\Psi(p_n) - \Psi(p_0) = (P_n - P_0)D(p_1) + o_P(1/\sqrt{n}).$$

In particular, if $D(p_1) = D(p_0)$, and $D(p_0)$ is the efficient influence curve of Ψ at p_0, then $\Psi(p_n)$ is asymptotically efficient.

This shows that the targeted MLE of a linear parameter in a convex model is typically consistent and asymptotically linear for arbitrary starting density p_n^0, and if the targeted MLE p_n^∞ is consistent in the sense that $P_0(D(p_n^\infty) - D(p_0))^2 \to 0$ with probability tending to 1 for n converging to infinity (e.g., the initial starting density p_n^0 would already yield a consistent estimator $D(p_0^n)$ of $D(p_0)$), then the targeted MLE will also be efficient. We will now provide two examples illustrating this theorem. The first example represents a case in which the targeted MLE is efficient for arbitrary starting density p_n^0. The second example represents the case that the targeted MLE is consistent and asymptotically linear for arbitrary starting density p_n^0, and is efficient if the starting density consistently estimates $D(p_0)$.

Example 1 (**(Efficiency of a smooth cumulative distribution function)** In this example we have $D(p)(O) = I(O \leq t) - \int_0^t p(o)d\mu(o)$. A targeted MLE p_n solving $P_n D(p_n) = 0$ satisfies that $\Psi(p_n) = P_n I(\cdot \leq t)$ equals the empirical cumulative distribution function at t and is therefore asymptotically efficient, *for arbitrary starting density* p^0. Thus in this example the initial density does not need to be consistent in order to make the targeted MLE asymptotically efficient. Suppose that p_{nh}^0 is indexed by a bandwidth or model choice h, and let p_{nh}^* be the targeted MLE density estimator using as starting density p_{nh}^0. Each of the targeted MLE's p_{nh}^* results in the same estimator of the cumulative distribution function $\Psi(p_0)$ at time t. If one uses likelihood cross-validation to select h, then one selects among all of these targeted MLE's the one which is supposedly closest to the true density p_0 with respect to Kullback-Leibler divergence, which now provides a valid and reasonable criteria since all the candidates density estimators already map into efficient (and algebraically equivalent) estimators of ψ_0.

Example 2 (**(Local efficiency of targeted MLE based on censored data)** We consider a particular example of a censored data structure to illustrate that Theorem 1 yields local efficiency of the targeted MLE based on CAR censored data structures based on any starting density p_n^0, under very weak conditions.

Suppose that the full data structure $X = (W, Y(a) : a \in \{0,1\})$ on the experimental unit consists of a set of baseline covariates W, and treatment specific outcomes $Y(a)$, indexed by treatment values $a \in \{0,1\}$. Suppose that the observed data structure $O = (W, A, Y = Y(A)) \sim p_0$, and it is assumed that the conditional probability distribution $g_0(\cdot \mid X)$ of A, given X, satisfies $g_0(A \mid X) = g_0(A \mid W)$: that is, A is independent of X, given W. Suppose that this conditional probability distribution of $g_0(A \mid W)$ of A, given W, is known, and satisfies $0 < g_0(1 \mid W) < 1$, as it would be in a randomized trial aiming to establish the causal effect of A on Y. Let $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure. We have

$$\mathcal{M} = \{p(O) = Q_{XA}(W,Y)g_0(A \mid X) : Q_{X0}, Q_{X1}\},$$

where the full data sub-distributions $Q_{Xa}(w,y) = P_{W,Y(a)}(w,y)$ are joint densities of $(W, Y(a))$, $a \in \{0,1\}$, and are unspecified. As a consequence, $\mathcal{M}$ is a convex model. Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined as $\Psi(p) = E_p(Y(1) - Y(0)) = E_p(E_p(Y \mid A = 1, W) - E_p(Y \mid A = 0, W))$, which is often called the marginal causal effect of treatment A on the outcome Y. In this case, $\Psi(p)$ is pathwise differentiable at p with efficient influence curve $S(p)$ defined by

$$S(p) = \frac{(Y - Q(p)(A, W))(A - (1 - A))}{g(p)(A \mid W)} + Q(p)(1, W) - Q(p)(0, W) - \Psi(p),$$

where $g(p)(\cdot \mid W) = Pr_p(A = \cdot \mid W) = g_0(\cdot \mid W)$, and $Q(p)(A, W) = E_p(Y \mid A, W)$. Note that $\Psi(p)$ depends on p through $Q(p)$ and its marginal distribution p_W of W. Due to the factorization of the density of O in a Q_X-factor and g_0 factor, this is also the efficient influence curve if g_0 is unknown or modelled. The class of all gradients at $p \in \mathcal{M}$ is given by:

$$\left\{ \frac{(Y - Q(A, W))(I(A = 1) - I(A = 0))}{g_0(A \mid W)} + Q(1, W) - Q(0, W) - \Psi(p) : Q \right\},$$

where Q can be an arbitrary function of A, W.

So we could define

$$D_Q(p)(O) \equiv \frac{(Y - Q(A, W))(A - (1 - A))}{g_0(A \mid W)} + Q(1, W) - Q(0, W) - \Psi(p),$$

and $D(p) = D_{Q(p)}(p)$ represents the efficient influence curve. We are now ready to define the targeted MLE of p_0 with respect to the parameter ψ_0.

Let p_n^0 be an initial density estimator of p_0. For example, p_n^0 could correspond with the empirical distribution of W, and a normal distribution for the conditional density of Y, given A, W, with mean $Q_n^0(A, W)$ and variance $\sigma_n^2(A, W)$, where Q_n^0 is an estimate of $Q(p_0)(A, W) = E_0(Y \mid A, W)$. Let p_n^* be a targeted MLE, as we explicitly define in the later Section 6 in detail, solving $P_n D(p_n^*) = 0$. In Section 6, we show for a particular hardest submodel $p_n^k(\epsilon)$ consisting of normal densities of Y, conditional on A, W, with ϵ corresponding with a fluctuation of current regression $Q_n^k(A, W)$, that the targeted MLE is achieved in the first step (i.e., $p_n^* = p_n^1$), and indeed solves the score equation $P_n D(p_n^1) = 0$. Let's consider this particular targeted MLE for illustration, but the following arguments apply to any targeted MLE solving $P_n D(p_n^*) = 0$.

Application of the theorem teaches us that

$$\Psi(p_n^*) - \psi_0 = (P_n - P_0) D_{Q(p_n^*)}.$$

Since g_0 is bounded away from zero, if Q_n^1 is a nice smooth function (e.g., with a uniformly bounded uniform sectional variation norm, van der Laan

(1995)), it follows that $D_{Q(p_n^*)}$ falls in a P_0-Donsker class, and thus that $\Psi(p_n^*) - \psi_0 = O_P(1/\sqrt{n})$. If the initial regression estimator $Q_n^0 = Q(p_n^0)$ converges to a possibly misspecified $Q_1 = Q(p_1)$, then it follows that $\Psi(p_n^*)$ is asymptotically linear with influence curve $D_{Q(p_1)}(O)$, where p_1 is the possibly misspecified limit of p_n^1. Finally, if Q_n^0 is actually consistent for $Q(p_0)$, then the targeted MLE of ψ_0 is asymptotically efficient. We can use likelihood based cross-validation to select among targeted MLE's indexed by different candidate initial estimators Q_n^0, thereby improving the efficiency relative to a targeted MLE with a fixed initial Q_n^0. Thus this example teaches us that the targeted MLE $\Psi(p_n^*)$ of ψ_0, which typically equals the first step targeted MLE, is consistent and asymptotically linear for arbitrary initial regression estimator Q_n^0, and it is efficient if Q_n^0 happens to be consistent, where the latter can potentially be achieved by using a machine learning type algorithm and selecting the fine tuning parameters with likelihood based cross-validation. These results still carry through if g_0 is unknown but is known to belong to a parametric model.

4.2 Local efficiency for general smooth parameters.

The remarkable robustness with respect to the starting density p_n^0 as observed in the previous subsection is a consequence of the convexity of the model and linearity of the parameter Ψ. In general, such results cannot be expected to hold. In this subsection we present a more general approach for establishing the wished asymptotic linearity and efficiency of the targeted MLE of any pathwise differentiable parameter.

Let $p_n^\infty \in \mathcal{M}$ denote the limit of the targeted MLE algorithm if it exists and otherwise it represents a p_n^k for a large k. If the targeted MLE solves the efficient influence curve equation, then for all practical purposes, we have $P_n D(p_n^\infty) = 0$. Let $R(p, p_0)$ be defined by

$$\Psi(p) - \Psi(p_0) = -P_0 D(p) + R(p, p_0)$$

for any $p \in \mathcal{M}$. We note that by pathwise differentiability of Ψ at p, $R(p, p_0)$ represents a second order term in the difference $p - p_0$. Combining $P_n D(p_n^\infty) = 0$ with the latter identity gives us

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0) D(p_n^\infty) + R(p_n^\infty, p_0).$$

Applying empirical process theory now proves that $\Psi(p_n^\infty)$ is root-n consistent if $D(p_n^\infty)$ falls in a P_0 Donsker class with probability tending to 1, and $R(p_n^\infty, p_0) = o_P(1/\sqrt{n})$. If one can now also establish that $P_0(D(p_n^\infty) - D(p_1))^2$

converges to zero in probability for a possibly misspecified $p_1 \in \mathcal{M}$, then it follows that $\Psi(p_n^\infty)$ is asymptotically linear with influence curve $D(p_1) - P_0 D(p_1)$:

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_1) + o_P(1/\sqrt{n}).$$

In particular, if $D(p_1) = D(p_0)$, then the targeted MLE is asymptotically efficient. Note that the asymptotic linearity requires that $R(p_n^\infty, p_0) = o_P(1/\sqrt{n})$, while the convexity of the model and linearity of the parameter as assumed in the previous subsection allowed us to avoid such a condition: i.e. in that case we had $R(p, p_0) = 0$ for arbitrary $p \in \mathcal{M}$ with $p_0/p < \infty$.

5 Fusion of MLE and estimating equations

In this section we show that the targeted MLE can be viewed as a solution of an optimal estimating equation for the parameter of interest, if one estimates the nuisance parameters with the targeted MLE itself. This comparison can only be made by making the assumption that the efficient influence curve can be viewed as an estimating function of the parameter of interest, which is needed for the estimating function methodology (van der Laan and Robins (2003)), but not for targeted MLE.

As previously argued, a sieve-based maximum likelihood estimator of a pathwise differentiable parameter is based on choices such as the sieve and the criteria for trading off variance and bias, which is completely unrelated to the actual parameter Ψ. As a consequence, such likelihood based estimators suffer, in principle, from serious bias for the parameter of interest ψ_0. Let p_n^0 be such a likelihood based estimator of p_0 and $\Psi(p_n^0)$ be the corresponding substitution estimator of ψ_0.

On the other hand, estimating function methodology (van der Laan and Robins (2003)) constructs estimating functions $D_h(\psi, \upsilon)(O)$ for the parameter of interest ψ indexed by a choice h, based on a representation of the orthogonal complement of the nuisance tangent space $p \rightarrow T_{nuis}^{\perp}(p)$ (i.e., $D_h(\Psi(p), \Upsilon(p)) \in T_{nuis}^{\perp}(p)$ for all h), which typically also depend on an unknown nuisance parameter Υ satisfying $E_p D_h(\Psi(p), \Upsilon(p)) = 0$ for all $p \in \mathcal{M}$. The current recommendation in estimating function methodology (see e.g., van der Laan and Robins (2003)) proposes to use an external estimator υ_n of nuisance parameters and estimate ψ_0 with the solution of $0 = P_n D_{h_n}(\psi, \upsilon_n) = 0$ in ψ. For example, one could use the maximum likelihood estimator p_n^0 and estimate ψ_0 with the solution ψ_{n0} of $0 = P_n D_{h(p_n^0)}(\psi, \Upsilon(p_n^0))$. This estimator ψ_{n0} is not necessarily, and in fact, will typically not be equal to $\Psi(p_n^0)$. Thus, even if the nuisance parameters are based on a maximum likelihood estimator p_n^0, the resulting estimating function based estimators of ψ_0 are intrinsically different from (and less biased than) the likelihood based estimator $\Psi(p_n^0)$.

However, let p_n be the targeted maximum likelihood estimator based on hardest submodels at p with efficient influence curve $D(p) = D_{h(p)}(\Psi(p), \Upsilon(p))$ and starting with the initial density estimator p_n^0, so that p_n solves $P_n D(p_n) = D_{h(p_n)}(\Psi(p_n), \Upsilon(p_n)) = 0$. Again, we consider the (now targeted) maximum likelihood estimator $\Psi(p_n)$ versus the estimating function based estimator described in the previous paragraph. The estimating function based estimator ψ_n of ψ_0 is defined as the solution of the estimating equation $0 = P_n D_{h(p_n)}(\psi, \Upsilon(p_n))$, which differs from above by now using the targeted MLE p_n (based on p_n^0) to estimate the index and nuisance parameters (instead of likelihood based p_n^0). Because $P_n D_{h(p_n)}(\Psi(p_n), \Upsilon(p_n)) = 0$, it follows that the estimating function based estimator ψ_n now equals $\Psi(p_n)$, assuming that this solution is unique. That is, if one estimates the nuisance parameters and index in the estimating function methodology with a targeted maximum likelihood estimator p_n, then the (or, at least, one of the) estimating function based estimator ψ_n and the targeted maximum likelihood estimator $\Psi(p_n)$ are identical.

Note that the targeted MLE is more general than the estimating function based methodology since it does not require the representation of an estimating function as a function of the parameter of interest and a variation independent nuisance parameter, thereby making it more widely applicable. Another advantage of targeted MLE relative to estimating function based estimation that it is invariant to monotone transformations of the parameter of interest.

5.1 CAR-censored data models

This targeted MLE approach has a particular nice application in estimation of pathwise differentiable parameters based on censored data under the coarsening at random assumption (Heitjan and Rubin (1991), Jacobsen and Keiding (1995), Gill et al. (1997), van der Laan and Robins (2003)). That is, let $O = \Phi(C, X) \sim p_0$ for some known many to one mapping Φ, $X \sim F_{X0}$ is the full data structure one wishes to observe on a randomly sampled experimental unit, and assume that the conditional distribution of the censoring variable C, given X, i.e., the censoring mechanism, satisfies coarsening at random (CAR). In this case it is known that the density of O factorizes as: $p_0(0) = g(p_0)(O \mid X)Q(p_0)(O)$, where $g(p_0)(O \mid X)$ (which is only a function of O by CAR) is the conditional density of O, given X, which thus only depends on the conditional distribution of C, given X. The $Q(p_0)$ factor only depends on the distribution F_{X0} of the full data structure X (van der Laan and Robins (2003)). Thus given a model $\mathcal{M}$ for O obtained by modelling

F_{X0} and or the censoring mechanism $g_0(O \mid X)$, each $p \in \mathcal{M}$ is identified by $(g(p), Q(p))$. Let $\Psi(p) = \Psi(Q(p))$ be a pathwise differentiable parameter of the $Q(p)$-part of the density p of O: i.e., it represents an identifiable parameter of F_X. In this case, it is known that the efficient influence curve $D(p) = D(g(p), Q(p))$ at $p \in \mathcal{M}$ is orthogonal to the tangent space $T_{CAR}(p)$ of the censoring mechanism g at p only assuming CAR (i.e., the Hilbert space in $L^2_0(P)$ spanned by all scores of parametric submodels through $g(p)$ at p), where $T_{CAR}(p) = \{h(O) : E_p(h(O) \mid X) = 0\}$ consists of all functions of O with conditional mean, given X, equal to zero. As a consequence, given an initial estimator Q^0 of $Q(p_0)$ and g^0 of $g(p_0)$, a hardest parametric model for $_0$ can be chosen to be of the form $p^0(\epsilon) \approx (1 + \epsilon D(p^0))p^0 = g^0 Q^0(\epsilon)$, where $Q^0(\epsilon) \approx (1 + \epsilon D(Q^0, g^0))Q^0$. That is, the hardest parametric model only corresponds with changing Q^0, but it leaves g^0 untouched. The targeted MLE approach proceeds now as defined above.

5.2 Targeting the censoring mechanism.

In this subsection we propose a targeted maximum likelihood methodology for estimation of ψ_0 which involves updating of estimators of both g_0 and Q_0. As shown in van der Laan and Robins (2003) (Theorem 1.3), we have that any gradient $D(p)$ can be decomposed as $D(p) = D_{IPCW}(p) - D_{CAR}(p)$ with D_{IPCW} being a so called Inverse Probability of Censoring Weighted (IPCW) function, and $D_{CAR}(p) = \Pi(D_{IPCW}(p) \mid T_{CAR}(p))$ is the projection of the IPCW function $D_{IPCW}(p)$ onto $T_{CAR}(p)$ in the Hilbert space $L^2_0(p)$. In order to relate these functions to estimating functions for ψ_0 (as in van der Laan and Robins (2003)) we will also sometimes use $D_{IPCW}(p) = D_{IPCW}(g(p), \Psi(p))$ and $D(p) = D(g(p), Q(p), \Psi(p))$ in the case that these functions can be represented as an estimating function in ψ indexed by nuisance parameters being functions of $g(p)$ and $Q(p)$: we note that the IPCW estimating function typically only depends on p through $g(p)$ and $\Psi(p)$. Given an initial estimator $p^0_n = (g^0_n, Q^0_n)$, in the censored data literature one defines the IPCW-estimator and DR-IPCW estimator as the solutions of the estimating equations $P_n D_{IPCW}(g^0_n, \psi) = 0$ and $P_n D(g^0_n, Q^0_n, \psi) = 0$, respectively, and $\Psi(Q^0_n)$ is called the likelihood based estimator (making the assumption that Q^0_n is likelihood based).

We will now describe the targeted MLE algorithm also involving the updating of g^0_n. At step k it now involves also a parametric submodel $g(p^k_n)(\epsilon_2)$ through $g(p^k_n)$ with score $D_{CAR}(g^k_n, Q^k_n)$ at $\epsilon_2 = 0$. It can be shown that $D_{CAR}(g(p), Q(p))$ corresponds with the efficient influence curve of the parameter $\Phi(g) = E_p D_{IPCW}(g, Q(p))$ at $g = g(p)$, so that this parametric submodel makes the estimator of g_0 targeted for estimation of the mean of the $IPCW$-component of the efficient influence curve. In particular, it is also the parametric submodel which makes the IPCW estimator $\psi_{n,IPCW}$, defined as the solution of the IPCW estimating equation $0 = P_n D_{IPCW}(g_n, \psi)$, efficient if the submodel is correctly specified, under regularity conditions. As above, let $Q_n^k(\epsilon_1)$ be a parametric submodel through Q_n^k with score $D(g_n^k, Q_n^k)$ at $\epsilon_1 = 0$.

Targeted MLE algorithm:

- Set $k = 0$.
- Let $p_n^k = (g_n^k, Q_n^k)$.
- Let $\epsilon_{1nk} = \arg\max_{\epsilon_1} P_n \log Q_n^k(\epsilon_1)$, and $\epsilon_{2nk} = \arg\max_{\epsilon_2} P_n \log g_n^k(\epsilon_2)$.
- Set $g_n^{k+1} = g_n^k(\epsilon_{2n})$ and $Q_n^{k+1} = Q_n^k(\epsilon_{1n})$. Set $p_n^{k+1} = (g_n^{k+1}, Q_n^{k+1})$.
- Set $k = k + 1$, and iterate this process utill convergence.

If ϵ_{1nk} and ϵ_{2nk} converge to zero for $k \rightarrow \infty$ (which can be expected because both factors g and Q of the likelihood are increasing at each step), then the targeted MLE algorithm will converge to a simultaneous solution of

$$\lim_k P_n D_{CAR}(g^k, Q^k) = 0 \text{ and } \lim_k P_n D(g^k, Q^k) = 0.$$

Equivalence of IPCW, DR-IPCW, and targeted MLE: As a consequence of the decomposition $D(p) = D_{IPCW}(p) - D_{CAR}(p)$, this implies also $\lim_k D_{IPCW}(g^k, \Psi(Q^k)) = 0$. Note that the double robust IPCW estimator defined as the solution in ψ of $P_n D(g_n^k, Q_n^k, \psi) = 0$, the targeted maximum likelihood estimator $\Psi(Q_n^k)$, and the IPCW estimator defined as the solution of $P_n D(g_n^k, \psi) = 0$, all based on these targeted MLE's g_n^k, Q_n^k are identical up to an arbitrarily small error decreasing in k (assuming uniqueness of the DR-IPCW and IPCW solution).

6 Examples of targeted maximum likelihood.

In this section we provide some important examples of the targeted MLE to illustrate its remarkable simplicity and good properties. For additional examples we refer to our accompanying technical report.

6.1 Estimation of a mean in a nonparametric model.

Consider an initial data generating density p_n^0 (with respect to a dominating measure μ) of a possibly multivariate random variable O, a given function $w(\cdot)$, and define the parameter of interest as

$$\Psi(p) = E_p[w(O)] = \int w(o)p(o)d\mu(o).$$

For the exponential family

$$\left\{ p_n^0(\epsilon)(x) = \frac{\exp(\epsilon(w(x) - \psi_n^0))p_n^0(x)}{\int \exp(\epsilon(w(x) - \psi_n^0))p_n^0(x)d\mu(x)} : \epsilon \right\},$$

consider attempting to estimate ϵ with maximum likelihood based on an i.i.d. sample $\{O_i\}_{i=1}^n$. Here $\psi_n^0 = \Psi(p_n^0)$. The log likelihood is then,

$$l(\epsilon) = \sum_{i=1}^n [\log(p_n^0(O_i)) + \epsilon(w(O_i) - \psi_n^0) - \log\left(\int \exp(\epsilon(w(x) - \psi_n^0))p_n^0(x)d\mu(x) \right)].$$

In our accompanying technical report we show that (for each initial p_n^0) the one-step targeted maximum likelihood estimator $\Psi(p_n^1) = \Psi(p_n^0(\epsilon_n)$ of the mean of $w(O)$ equals the sample mean $\bar{W}_n = \frac{1}{n}\sum_{i=1}^n w(O_i)$. For the detailed proof we refer to our technical report.

6.2 Estimation of a marginal causal effect.

Double robust locally efficient estimation of the causal effect of a point treatment assuming a marginal structural model has been provided in Robins (2000), Robins and Rotnitzky (2001), and Robins et al. (2000): see also van der Laan and Robins (2003).

Let $O = (W, A, Y)$, W be a vector of baseline covariates, A be a binary treatment variable, and Y an outcome of interest. Let $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure: so $\mathcal{M}$ is nonparametric up to possible smoothness conditions. Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined as $\Psi(p) = E_p(E_p(Y \mid A = 1, W) - E_p(Y \mid A = 0, W))$, where it is assumed $0 < P(A = 1 \mid W) < 1$ with probability one so that this parameter is well defined. This parameter corresponds with the marginal causal effect of A on Y if one assumes the usual consistency assumption, temporal ordering assumption, and randomization assumption required for causal inference. In order to acknowledge that this parameter is of interest in general, van der Laan (2006) refers to this parameter as the variable importance of variable

A. This parameter $\Psi(p)$ is pathwise differentiable at p with efficient influence curve $S(p)$ defined by

$$\begin{aligned} S(p) &= \frac{(Y - Q(p)(A,W))(I(A=1) - I(A=0))}{g(p)(A \mid W)} \\ &\quad + Q(p)(1,W) - Q(p)(0,W) - \Psi(p), \end{aligned}$$

where $g(p)(\cdot \mid W) = Pr_p(A = \cdot \mid W)$, and $Q(p)(A,W) = E_p(Y \mid A,W)$ (see e.g., Robins (2000), van der Laan (2006)). Note that $\Psi(p)$ depends on p through $Q(p)$ and its marginal distribution p_W of W. Because the model is locally saturated, it is also the *only* influence curve/gradient (Gill et al. (1997)). So we set $D(p) = S(p)$.

We can decompose this efficient score $D(p)$ into three subcomponents as follows:

$$\begin{aligned} D(p) &= D(p) - E_p(D(p) \mid A,W) + E_p(D(p) \mid A,W) - E_p(D(p) \mid W) \\ &\quad + E_p(D(p) \mid W) - E_pD(p), \end{aligned}$$

which corresponds with scores for $p(Y \mid A,W)$, $p(A|W)$ and $p(W)$, respectively. We have

$$\begin{aligned} D_1(p)(O) &\equiv D(p) - E_p(D(p) \mid A,W) \\ &= (Y - Q(p)(A,W))\frac{A - (1-A)}{g(p)(A \mid W)} \\ E_p(D(p) \mid A,W) - E_p(D(p) \mid W) &= 0 \\ D_2(p) &\equiv E_p(D(p) \mid W) - E_p(D(p)) \\ &= Q(p)(1,W) - Q(p)(0,W) - \Psi(p). \end{aligned}$$

Consider an initial density estimator p_n^0 of the density p_0 of (W,A,Y) with marginal distribution of W being the empirical probability distribution of $W_1,\ldots,W_n$. We have that $D(p_n^0) = D_1(p_n^0) + D_2(p_n^0)$ and thus that a one-dimensional $p_n^0(\epsilon)$ with score $D(p_n^0)$ at $\epsilon = 0$ corresponds with a zero score for $g(p_n^0)$. In addition, we have that $P_nD_2(p_n^0) = 0$ (i.e., the empirical distribution of W is a nonparametric maximum likelihood estimator) so that $p_n^0(\epsilon)$ can be selected to only vary $p_n^0(Y \mid A,W)$ with a score $D_1(p_n)$ at $\epsilon = 0$.

We now propose an easily implemented targeted maximum likelihood estimator of the marginal causal effect by using a normal regression model as hardest submodel. Specifically, consider an initial density estimator p_n^0 with marginal distribution of W equal to the empirical probability distribution of $W_1,\ldots,W_n$, and let the conditional probability density $p_n^0(Y \mid A,W) =$

$\frac{1}{\sigma(Q_n^0)(A,W)} f_0(\{Y - Q_n^0(A,W)\}/\sigma(Q_n^0)(A,W))$ be a normal density with mean $Q_n^0(A,W)$ and variance $\sigma(Q_n^0)^2(A,W)$. Here f_0 denotes the $N(0,1)$ density. In addition, $g(p_0^n)(A \mid W)$ is a particular fit of the conditional density of A, given W. We now consider as possible submodels $p_n^0(\epsilon)$

$$p_n^0(\epsilon)(Y \mid A, W) = \frac{1}{\sigma(Q_n^0(A,W)} f_0 \quad \frac{Y - Q_n^0(A,W) - \epsilon h(p_n^0)(A,W)}{\sigma(Q_n^0)(A,W)}\Bigg),$$

where the function h will be specified so that the score of p_n^0 at $\epsilon = 0$ equals the efficient influence curve at p_n^0. The maximum likelihood estimator of ϵ is simply given by the weighted least squares estimator for a univariate linear regression model:

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} (Y_i - Q_n^0(A_i, W_i) - \epsilon h(p_n^0)(A_i, W_i))^2 \frac{1}{\sigma(Q_n^0)^2(A_i, W_i)}.$$

The score of $p_n^0(\epsilon)(Y \mid A, W)$ at a value ϵ is given by:

$$S(\epsilon) = -\frac{Y - Q_n^0(A,W) - \epsilon h(p_n^0)(A,W)}{\sigma(Q_n^0)^2(A,W)} h(p_n^0)(A,W),$$

and ϵ_n solves indeed $P_n S(\epsilon_n) = 0$. If we set

$$h(p_n^0)(A,W) \equiv \left(\frac{I(A=1)}{g_n^0(1 \mid W)} - \frac{I(A=0)}{g_n^0(0 \mid W)} \right) \sigma(Q_n^0)^2(A,W),$$

then the score $S(0) = D_1(p_n^0) = (Y - Q_n^0(A,W))(I(A=1)/g_n^0(1 \mid W) - I(A=0)/g_n^0(0 \mid W))$ of $p_n^0(\epsilon)(Y \mid A, W)$ at $\epsilon = 0$ corresponds with the efficient influence curve at p_n^0. As in our previous subsection, since $p_n^0(W)$ equals the empirical distribution of W the MLE of $\epsilon_1 \rightarrow P_n \log p^0(\epsilon_1)(W)$ equals $\epsilon = 0$, and $g_n^0(A \mid W)$ will not be varied by $p_n^0(\epsilon)$: that is, the marginal distribution of W and the treatment mechanism $g^0(A \mid W)$ will not be updated in the algorithm for calculating the targeted maximum likelihood estimator.

Let $p_n^1 = p_n^0(\epsilon_n)$ whose conditional distribution of Y, given A, W, is a normal density with mean $Q_n^1(A,W)$ and variance $\sigma^2(Q_n^1)(A,W)$, where

$$Q_n^1(A,W) = Q(p_n^1)(A,W) = Q_n^0(A,W) + \epsilon_n h(p_n^0)(A,W).$$

The corresponding estimate of ψ_0 is given by

$$\Psi(p_n^1) = \frac{1}{n} \sum_{i=1}^{n} Q_n^1(1, W_i) - Q_n^1(0, W_i).$$

It is straightforward to show that $P_n D(p_n^1) = 0$ in the case that $\sigma_n^0(A, W)$ is constant in the model $\{p_n^0(\epsilon) : \epsilon\}$, but is simply set at an initial estimate. Thus in this case the targeted maximum likelihood is achieved at the first step. For arbitrary fixed values of $\sigma(A, W)$, the targeted MLE is locally efficient in the sense that if $g(p_n^0)$ is consistent at some rate, then it is consistent and asymptotically linear for arbitrary Q_n^0, and it is efficient if Q_n^0 is consistent for $Q_0(A, W)$. Likewise, a consistent $Q_n^1(A, W)$ will lead to a consistent estimator of the parameter of interest ψ_0, even with an arbitrary fit of the treatment mechanism $g(A|W)$. Iterative estimation of σ provides no (asymptotic) reward, and could simply be omitted by setting (e.g.) σ at an initial estimate, so that the targeted MLE is achieved in a single step.

6.3 Targeting the treatment mechanism as well.

We will now proceed with this example, but also use for g_0 a targeted maximum likelihood estimator. Our goal is to make the IPTW estimator $\psi_{n,IPTW} = \frac{1}{n}\sum_{i=1}^n Y_i \frac{I(A_i=1)-I(A_i=0)}{g_n(A_i|W_i)}$ corresponding wiht the targeted MLE g_n an efficient estimator. Let $g(p_n^0)(A \mid W)$ be an initial estimator and represent it as a logistic function:

$$g(p_n^0)(1 \mid W) = \frac{1}{1 + \exp(-m_n^0(W))}.$$

Consider as parametric submodel

$$g(p_n^0)(\epsilon_2)(1 \mid W) = \frac{1}{1 + \exp(-m_n^0(W) - \epsilon_2 h(p_n^0)(W))}. \tag{11}$$

Let $\epsilon_{2n} = \arg\max P_n \log g(p_n^0)(\epsilon)$. In practice this can be done by fitting a logistic regression in the covariates $m_n^0(W)$ and $h(p_n^0)(W)$, setting the intercept equal to zero, and setting the coefficient in front of $m_n^0(W)$ equal to 1, and set ϵ_{2n} equal to fitted coefficient in front of $h(p_n^0)(W)$. It is also fine to refit the intercept and coefficient in front of $m_n^0(W)$, since choosing additional parameters still guarantees that the linear span of scores includes the score of $h(p_n^0)(W)$. We have

$$\frac{d}{d\epsilon_2} \log g(p_n^0)(\epsilon_2)\Big|_{\epsilon_2=0}(O) = h(p_n^0)(W)(A - g(p_n^0)(1 \mid W)).$$

Solving for h so that

$$h(W)(A - g(p_n^0)(1 \mid W)) \;=\; D_{CAR}(p_n^0)(O)$$

DOI: 10.2202/1557-4679.1043

$$\begin{aligned} &= \frac{Q(p_n^0)(A,W)}{g_n^0(A \mid W)}\{I(A=1)-I(A=0)\} \\ &\quad -\{Q(p_n^0)(1,W)-Q(p_n^0)(0,W)\} \end{aligned}$$

yields the solution

$$h(p_n^0)(W) = \frac{Q(p_n^0)(1,W)}{g(p_n^0)(1 \mid W)} + \frac{Q(p_n^0)(0,W)}{g(p_n^0)(0 \mid W)}.$$

We are now ready to present the proposed targeted MLE which also targets the treatment mechanism fit.

The algorithm for targeted maximum likelihood estimation of a marginal causal effect, including the targeting of the treatment mechanism. Thus the algorithm for targeted maximum likelihood estimation of $_0$ can be described as follows. Let $k = 0$, and let $g^0(A \mid W)$ and the regression fit $Q^0(A,W)$ of $E_0(Y \mid A,W)$ be given. Let

$$h_1^k = h_1(g^k, Q^k)(A,W) \equiv \left(\frac{I(A=1)}{g^k(1 \mid W)} - \frac{I(A=0)}{g^k(0 \mid W)}\right)\sigma(Q^k)^2(A,W)$$

and

$$h_2^k = h_2(g^k, Q^k)(W) = \frac{Q^k(1,W)}{g^k(1 \mid W)} + \frac{Q^k(0,W)}{g^k(0 \mid W)}.$$

Let $m^k(W) = \log(g^k(1 \mid W)/g^k(0 \mid W))$ so that $g^k(1 \mid W) = 1/(1 + \exp(-m^k(W))$. Consider the logistic regression model

$$g^k(\epsilon_2)(1 \mid W) = \frac{1}{1+\exp(-m^k(W) - \epsilon_2 h_2^k(W))}.$$

Let $\epsilon_{2n}(k) = \arg\max_{\epsilon_2} P_n \log g^k(\epsilon_2)$ be the maximum likelihood estimator of this univariate logistic regression model, and let

$$\epsilon_{1n}(k) = \arg\min_{\epsilon_1} \sum_{i=1}^n (Y_i - Q^k(A_i,W_i) - \epsilon_1 h_1^k(A_i,W_i))^2 \frac{1}{\sigma(Q^k)^2(A_i,W_i)},$$

the univariate least squares estimator of ϵ_1.

Now, update g^k and Q^k as follows:

$$\begin{aligned} Q^{k+1}(A,W) &= Q^k(A,W) + \epsilon_{1n}(k)h_1^k(A,W) \\ m^{k+1}(A,W) &= m^k(W) + \epsilon_{2n}(k)h_2^k(W) \\ g^{k+1}(A \mid W) &= \frac{1}{1+\exp(-m^{k+1}(W))} \end{aligned}$$

Set $k = k+1$ and iterate this algorithm.

Equivalence of IPTW, DR-IPTW, and targeted maximum likelihood estimators. Recall that the efficient influence curve function is decomposed as $D(g,Q)(O) = D_{IPTW}(g,Q) - D_{CAR}(g,Q)$, where $D_{IPTW}(g,Q) = \frac{Y}{g(A|W)}(I(A=1) - I(A=0)) - \Psi(Q)$, and $D_{CAR}(g,Q) = \frac{Q(A,W)}{g(A|W)}(I(A = 1) - I(A=0)) - (Q(1,W) - Q(0,W))$. For k converging to infinity the targeted MLE yields a final estimator g_n of the treatment mechanism and a regression fit $Q_n(A,W)$ so that the score equations of the two submodels in ϵ_1 and ϵ_2 are solved at $\epsilon_1 = \epsilon_2 = 0$:

$$P_n D(g_n, Q_n) = 0 \text{ and } P_n D_{CAR}(g_n, Q_n) = 0.$$

This implies also that

$$P_n D_{IPTW}(g_n, Q_n) = 0.$$

Thus, we can conclude that the three estimators

$$\begin{aligned}
\Psi_{n,IPTW} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{g_n(A_i \mid W_i)}(I(A_i = 1) - I(A_i = 0)) \\
\Psi_{n,DR-IPTW} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{g_n(A_i \mid W_i)}(I(A_i = 1) - I(A_i = 0)) \\
&- D_{CAR}(g_n, Q_n)(A_i, W_i) \\
\Psi_{n,MLE} &= \frac{1}{n}\sum_{i=1}^{n} Q_n(1, W_i) - Q_n(0, W_i)
\end{aligned}$$

are algebraically identical: $\Psi_{n,IPTW} = \Psi_{n,DR-IPTW} = \Psi_{n,MLE}$. That is, the targeted MLE $\Psi(Q_n)$ equals the IPTW and DR-IPTW estimator based on the targeted MLE (g_n, Q_n) as estimators of the nuisance parameters (g_0, Q_0) in the corresponding estimating equations. Preliminary results suggest that consistency of the resulting targeted likelihood algorithm depends on the consistency of either the g_0 or Q_0 component of the initial density estimator.

6.4 Simulation for marginal variable importance.

Simulated data can be used to illustrate the benefits of the targeted likelihood procedure. We simulated replicates of the data structure $O = (W, A, Y) \sim p_0$ representing baseline covariates, a binary treatment, and a response measurement on a subject, and attempted to estimate the causal effect of treatment A on response Y. We generated 1000 datasets of size $n = 200$ according to the following mechanism:

$$W \sim U(0,1)$$

$$\begin{aligned}
A &\in \{0,1\} \\
g(1|W) &= P(A=1|W) = \frac{1}{1+\exp(-8W2+8W-1)} \\
\epsilon &\sim N(0,1), \quad \epsilon \perp (W,A) \\
Y &= AQ(1,W) + (1-A)Q(0,W) + \epsilon \\
Q(0,W) &= -\frac{2}{3}, \, Q(1,W) = -(8W^2 - 8W + 1)
\end{aligned}$$

Here O represented a censored data structure. The unavailable *counterfactual* data was given by,

$$X = (W, Y_0, Y_1) = (W, Q(0,W) + \epsilon, Q(1,W) + \epsilon).$$

It could be be verified that the coarsening at random assumption held, or that,

$$\{A \perp X|W\},$$

as well as the experimental treatment assignment assumption, implied by,

$$0 < 0.26 < g(1|W) < .74 < 1 \text{ with probability one.}$$

Together these assumptions made it possible to estimate the parameter,

$$\Psi(p_0) = E[Y_1] - E[Y_0] = 1,$$

representing the counterfactual mean difference between the treatment group ($A = 1$) and the control group ($A = 0$).

The standard estimators for this problem are the inverse probability of treatment (IPTW), maximum likelihood (G-computation), and doubly robust (efficient) estimators. These respectively depend on fitting either the censoring mechanism g or the nuisance parameter $Q(A,W) = E[Y|W]$, and are given as follows, where $h_g(A,W) = \frac{A}{g(1|W)} - \frac{1-A}{g(0|W)}$:

$$\begin{aligned}
\Psi_{n,\text{IPTW}}(g) &= \frac{1}{n}\sum_{i=1}^{n} Y_i h_g(A_i, W_i) \\
\Psi_{n,\text{MLE}}(Q) &= \frac{1}{n}\sum_{i=1}^{n} [Q(1,W_i) - Q(0,W_i)] \\
\Psi_{n,\text{DR-IPTW}}(g,Q) &= \Psi_{n,\text{IPTW}} + \Psi_{n,\text{MLE}} - \frac{1}{n}\sum_{i=1}^{n} h_g(A_i,W_i)Q(A_i,W_i)
\end{aligned}$$

Typically estimation is based on forming external estimates of at least one of the two nuisance parameters g or Q, and then applying one of the IPTW, maximum likelihood, or double robust estimators. The three estimators can potentially be very different from one another, leading to difficulties when interpreting the data. Targeted likelihood resolves this problem, by estimating both nuisance parameters g and Q accurately with maximum likelihood, but in a way so that the IPTW, maximum likelihood, and doubly robust estimators are algebraically equivalent.

As our initial fit to p_0 prescribed that $\{Y|A,W\}$ followed a Gaussian distribution with fixed variance, the hardest one-dimensional submodel $\epsilon \to p_\epsilon$ for estimation of $\Psi(p_0)$ could be given by,

$$\{Y|A,W\} \sim N(Q_n^{(0)}(A,W) + \epsilon h_g(A,W), \sigma^2),$$

while the laws of $\{W\}$ and $\{A|W\}$ were left unchanged. The maximum likelihood estimator of ϵ became,

$$\epsilon_n = \frac{\sum_{i=1}^n h_g(A_i,W_i)(Y_i - Q_n^{(0)}(A_i,W_i))}{\sum_{i=1}^n h_g(A_i,W_i)},$$

leading to the updated estimate of $Q(A,W) = E[Y|A,W]$,

$$Q_n^{(1)}(A,W) = Q_n^{(0)}(A,W) + \epsilon_n h_g(A,W).$$

When the treatment mechanism g was not updated, the targeted likelihood algorithm converged in a single iteration. Note that the update did not depend in any way on the choice of variance σ^2 for the law of $\{Y|A,W\}$, so long as it was a constant. The parameter $\Psi(p_0)$ was then estimated with $\Psi(p(\epsilon_n))$, which was equal to $\Psi_{n,\text{MLE}}(Q_n^{(1)})$ and $\Psi_{n,\text{DR-IPTW}}(g, Q_n^{(1)})$. The treatment mechanism g could also be updated with targeted likelihood, to make the IPTW estimator equivalent with the maximum likelihood and double robust estimators. This was done by making a one-dimensional model $g_\epsilon(1|W)$ through $g(1|W)$ at $\epsilon = 0$, whose score at $\epsilon = 0$ was the projection of the IPTW estimator's influence curve on T_{CAR}. Such a submodel could be formed by taking,

$$\text{logit}(g_\epsilon(1|W)) = g(1|W) + \epsilon[\frac{Q(1,W)}{g(1|W)} + \frac{Q(0,W)}{g(0|W)}].$$

Because this was simply a logistic model for $\{A|W\}$, we could estimate ϵ through logistic regression. After iterating the targeted likelihood procedure to update both of the Q and g nuisance parameters until convergence, the IPTW, maximum likelihood, and double robust estimators of $\Psi(p_0)$ became equivalent.

For this data structure, $\Psi_{n,\text{DR-IPTW}}(g,Q)$ was asymptotically efficient, meaning that its asymptotic performance was superior to any other regular estimator. This efficient estimator could not be used directly on observed data, due to its dependence on the unknown nuisance paramters g and Q. We assessed the quality of an estimator Ψ_n through the ratio

$$R(\Psi_n) = \frac{E_{p_0}[n|\Psi_n - \Psi(p_0)|^2]}{E_{p_0}[n|\Psi_{n,\text{DR-IPTW}}(g,Q) - \Psi(p_0)|^2]}$$

For large enough sample size n, and consistent and asymptotically linear Ψ_n, this approximated the asymptotic relative efficiency of Ψ_n to the efficient estimator, and necessarily exceeded one. We approximated $R(\Psi_n)$ after forming Ψ_n on 1000 simulated datasets of size $n = 200$.

In our simulations, we considered known censoring mechanism g, as could occur in a randomized clinical trial. We misspecified the nuisance parameter Q, by estimating $E[Y|W]$ in the $A = 0$ and $A = 1$ strata with linear regression, while quadratic regression would have been appropriate. This first-order approximation to Q lead to an inaccurate maximum likelihood estimator, having $R(\Psi_n) = 2.63$. Confidence intervals for $R(\Psi_n)$ were negligible, due to the number of simulations. The misspecified nuisance parameter Q did not affect the performance of the IPTW estimator, or the consistency of the double robust estimator, which respectively had asymptotic relative efficiencies $R(\Psi_n)$ of 1.18 and 1.15. Note that the IPTW estimator was unbiased, but was less accurate than the double robust estimator with misspecified Q. After updating Q with a single targeted likelihood iteration, $R(\Psi_n)$ decreased to 1.10. The resulting estimator was then a maximum likelihood estimator (and double robust estimator) with updated Q, and the update greatly increased of the accuracy of the parameter estimate. When also updating the censoring mechanism g, the asymptotic relative efficiency dropped even further to 1.07, making the estimator almost equivalent with the efficient estimator. In spite of the fact that the censoring mechanism g was already known, estimating it from the data was nevertheless beneficial, as could be surmised from Chapter 2.3.7 of (van der Laan and Robins (2003)).

Thus, the targeted likelihood algorithm allowed us to estimate the nuisance parameters g and Q with maximum likelihood in a manner such that three standard estimators become identical, and led to better performance than was achieved by the initial IPTW, maximum likelihood, and double robust estimators.

6.5 Semiparametric regression example.

Let $O = (W, A, Y) \sim p_0$ and consider the semiparametric regression model $\mathcal{M} = \{p : E_p(Y \mid A, W) - E_p(Y \mid A = 0, W) = m(A, W \mid \beta(p))\}$ for some parametrization $\beta \rightarrow m(A, W \mid \beta)$ satisfying $m(0, W \mid \beta) = 0$ for all $\beta \in \mathbb{R}^d$. This is equivalent with assuming $E_0(Y \mid A, W) = m(A, W \mid \beta_0) + \theta_0(W)$ with θ_0 unspecified and $m(0, W \mid \beta) = 0$, and can therefore also be viewed as a semiparametric regression model. It has been recognized that a maximum likelihood fit (e.g., generalized additive models) of the semiparametric regression suffers from bias for the parametric part, so that one needs to undersmooth the nonparametric components in the semiparametric regression model. However, the literature does not provide practical guidance about how to undersmooth. Therefore, the targeted MLE approach presented here provides an importance practical improvement. Let $\Psi(p) = \beta(p) \in \mathbb{R}^d$ be the parameter of interest.

This type of semiparametric regression models has been considered by various authors (e.g., Newey (1995); Rosenbaum and Rubin (1983); Robins et al. (1992); Robins and Rotnitzky; Yu and van der Laan (2003)). The latter three articles derive the orthogonal complement of the nuisance tangent space (i.e., the set of all gradients of the pathwise derivative), the efficient influence curve/canonical gradient, and establish the wished double robustness of the corresponding estimating functions. In particular, for our purpose we refer to Theorem 2.1 and 2.2 in Yu and van der Laan (2003) for the following statements.

The orthogonal complement of the nuisance tangent space is given by:

$$T^{\perp}_{nuis}(p) = \{D_h(p) : h\} \subset L^2_0(P),$$

where $D_h(p)(O) \equiv (h(A, W) - E_p(h(A, W) \mid W))(Y - m(A, W \mid \beta(p)) - E_p(Y \mid A = 0, W))$. The orthogonal complement of the nuisance tangent space corresponds with the set of gradients for Ψ at p given by:

$$T^{\perp}_{nuis}(p)^* = \left\{-c(p)(h)^{-1} D_h(p)(O) : h = (h_1, \ldots, h_d)\right\},$$

where $c(p)(h) = \frac{d}{d\beta} E_p D_h(p, \beta)\Big|_{\beta=\beta(p)}$, and D_h now represents a vector function $(D_{h_1}, \ldots, D_{h_d})$. The efficient influence curve is identified by a closed form index $h(p)$ (see e.g., Yu and van der Laan (2003)), which is provided below (12). Let $D(p) = D_{h(p)}(p)$ be this efficient influence curve at p as identified by this index $h(p)$.

Let $g(p)$ be the conditional density of A, given W, under p, let $Q(p)$ be the conditional distribution of Y, given A, W, under p. We note that the parameter $\Psi(p)$ is only a function of $Q(p)$, and the density factorizes as $p(O) = p(W)g(p)(A \mid W)Q(p)(Y \mid A, W)$. As a consequence, the elements $D_h(p)$ are orthogonal to the tangent spaces of the nuisance parameter $g(p)$ and the nuisance parameter $p(W)$. That is, we can decompose the efficient score $D(p)$ into three subcomponents as follows:

$$\begin{aligned} D(p) &= D(p) - E_p(D(p) \mid A, W) + E_p(D(p) \mid A, W) - E_p(D(p) \mid W) \\ &\quad + E_p(D(p) \mid W) - E_p D(p), \end{aligned}$$

which corresponds with scores for $p(Y \mid A, W)$, $p(A|W)$ and $p(W)$ at p, respectively, but $E_p(D(p) \mid A, W) - E_p(D(p) \mid W) = 0$ and $E_p(D(p) \mid W) - E(D(p)) = 0$. Thus the efficient influence curve $D(p)$ represents only a score for $Q(p)(Y \mid A, W)$, and indeed satisfies $E_p(D(p)(O) \mid A, W) = 0$.

Consider an initial density estimator $p_n^0 = (p_{nW}^0, g(p_n^0), Q(p_n^0))$ of (W, A, Y) with marginal distribution of W being the empirical probability distribution of $W_1, \ldots, W_n$. Above we showed that a submodel $p_n^0(\epsilon)$ through p_n^0 with score $D(p_n^0)$ at $\epsilon = 0$ can be selected to only vary the conditional density $Q(p_n^0)$ of Y, given A, W, with a score $D(p_n^0)$ at $\epsilon = 0$. Such a submodel will now be presented.

Let $p_n^0 \in \mathcal{M}$. Suppose that $Q(p_n^0)$ is a normal distribution with mean $\theta(p_n^0)(A, W) = E_{p_n^0}(Y \mid A, W)$ and variance $\sigma^2(A, W) = \sigma^2(Q_n^0)(A, W)$. Recall that $D(p_n^0) = (h(p_n^0)(A, W) - E_{p_n^0}(h(p_n^0) \mid W))(Y - m(A, W \mid \beta(p^0)) - E_{p_n^0}(Y \mid A = 0, W))$. For notational convenience, we will represent this function as $h(p_n^0)(A, W)(Y - E_{p_n^0}(Y \mid A, W))$ with now $h(p_n^0)$ so that $E_{p_n^0}(h(p_n^0)(A, W) \mid W) = 0$. Consider the parametric submodel of $\mathcal{M}$ defined as the normal density with conditional variance $\sigma^2(A, W)$ and conditional mean $m(A, W \mid \beta_n^0(\epsilon)) + \theta_n^0(\epsilon)$. That is,

$$Q_n^0(\epsilon)(Y \mid A, W) = \frac{1}{\sigma(A, W)} f_0 \left. \quad \frac{Y - m(A, W \mid \beta_n^0(\epsilon)) - \theta_n^0(\epsilon)(W)}{\sigma(A, W)} \right),$$

where $\beta_n^0(0) = \beta(Q_n^0)$, $\theta_n^0(0) = \theta(Q_n^0) = E_{Q_n^0}(Y \mid A = 0, W)$, and f_0 is the standard normal density. We note that this is a valid submodel through Q_n^0 at $\epsilon = 0$. Let $\beta(\epsilon) \equiv \beta(Q_n^0) + \epsilon$ and $\theta_n^0(\epsilon) = \theta(Q_n^0) + \epsilon^\top r$. It remains to find a function $r(W)$ so that the score of $Q_n^0(\epsilon)$ at $\epsilon = 0$ equals the efficient influence curve $D(p_n^0)$.

We have that the score $S(\epsilon)$ at ϵ is given by (note that $f_0'(x)/f_0(x) = 2x/\sigma^2$)

$$\begin{aligned} & S(\epsilon)\sigma^2(A, W) \\ = \quad & (Y - m(A, W \mid \beta_n^0(\epsilon)) - \theta_n^0(\epsilon)(W)) \left\{ \frac{d}{d\epsilon} m(A, W \mid \beta_n^0(\epsilon)) - \frac{d}{d\epsilon} \theta_n^0(\epsilon)(W) \right\} \end{aligned}$$

$$= \left\{ \frac{d}{d\beta_n^0(\epsilon)} m(A, W \mid \beta_n^0(\epsilon)) - r(W) \right\} (Y - m(A, W \mid \beta_n^0(\epsilon)) - \theta_n^0(\epsilon)(W)).$$

Solving for r so that $S(0) = D(p^0)$ yields the equation

$$h(p_n^0)(A, W)(Y - E_{Q^0}(Y \mid A, W)) =$$
$$\frac{1}{\sigma^2(A,W)} \left\{ \frac{d}{d\beta(Q_n^0)} m(A, W \mid \beta(Q_n^0)) - r(W) \right\} (Y - E_{Q_n^0}(Y \mid A, W)).$$

In order to have that the score equals D_h for a particular $h(A, W)$ with $E_{p_n^0}(h(A, W) \mid W) = 0$, we need

$$r(p_n^0)(W) = \frac{E_{p_n^0}\left(\frac{d/d\beta_n^0 m(A,W|\beta_n^0)}{\sigma^2(A,W)} \mid W \right)}{E_{p_n^0}\left(\frac{1}{\sigma^2(A,W)} \mid W \right)}.$$

This yields the following score for our submodel $p_n^0(\epsilon)$ at $\epsilon = 0$:

$$S(0) = h(p_n^0)(A, W)(Y - m(A, W \mid \beta(Q_n^0)) - \theta(Q_n^0)(W)),$$

where

$$\begin{aligned} h(p_n^0)(A, W) \equiv & \frac{1}{\sigma^2(A, W)} \frac{d}{d\beta(Q_n^0)} m(A, W \mid \beta(Q_n^0)) \\ & - \frac{1}{\sigma^2(A, W)} \frac{E_{p_n^0}\left(\frac{d}{d\beta(Q_n^0)} m(A, W \mid \beta(Q_n^0))/\sigma^2(A, W) \mid W \right)}{E_{p_n^0}(1/\sigma^2(A, W) \mid W)}. \end{aligned}$$

This choice $h(p_n^0)$ gives a score $S(0)$ equal to the efficient influence curve (see e.g., Yu and van der Laan (2003)). So we succeeded in finding a submodel $p_n^0(\epsilon)$ with a score at $\epsilon = 0$ equal to the efficient influence curve at p_n^0. Thus we are now ready to define the targeted MLE.

Consider the log-likelihood for $p_n^0(\epsilon)$ in ϵ:

$$l(\epsilon) \equiv \frac{1}{n} \sum_{i=1}^{n} \log f_0 \left. \quad \frac{Y_i - m(A_i, W_i \mid \beta_n^0 + \epsilon) - (\theta_n^0(W) + \epsilon^\top r(p_n^0)(W))}{\sigma(A, W)} \right).$$

Let ϵ_n be the maximizer, which can thus be computed with standard weighted least squares regression:

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} \frac{1}{\sigma^2(A_i, W_i)} \left(Y_i - m(A_i, W_i \mid \beta_n^0 + \epsilon) - \theta_n^0(W_i) - \epsilon r(p_n^0)(W_i) \right)^2.$$

The score equation $0 = d/d\epsilon l(\epsilon) = P_n S(\epsilon)$ for ϵ_n is given by

$$0 = P_n \frac{\left\{ \frac{d}{d\beta_n^0(\epsilon)} m(\beta_n^0(\epsilon)) - r(p_n^0)) \right\} (Y - m(\beta_n^0(\epsilon)) - \theta_n^0 - \epsilon^\top r(p_n^0))}{\sigma^2}.$$

DOI: 10.2202/1557-4679.1043

In the sequel we consider the case that $m(A,W \mid \beta) = \beta^\top m_1(A,W)$ is linear in β for some specified covariate vector $m_1(A,W)$. In this case we have $d/d\beta m(A,W \mid \beta) = m_1(A,W)$ so that the score equation $P_n S(\epsilon) = 0$ reduces to:

$$0 = P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n^\top r(p_n^0)\right)}{\sigma^2}. \tag{12}$$

Firstly, we note that ϵ_n exist in closed form:

$$\epsilon_n = A_n^{-1} P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - \beta_n^{0\top} m_1 - \theta_n^0\right)}{\sigma^2},$$

where the $d \times d$ matrix A_n is given by

$$A_n \equiv \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma^2(A_i,W_i)}\left\{m_1(A_i,W_i) - r(p_n^0)(W_i)\right\}\left(m_1(A_i,W_i) + r(p_n^0)(W_i)\right)^\top.$$

Let $p_n^0(\epsilon_n)$ be the new density estimator. Recall that the distribution of (A,W) under $p_n^0(\epsilon_n)$ is still the same as under p_n^0, because $p_n^0(\epsilon)$ only updates the conditional distribution of Y, given A,W. We now wish to investigate if the first step targeted MLE $p_n^1 \equiv p_n^0(\epsilon_n)$ already solves the efficient score equation: $P_n D(p_n^1) = P_n D(p_n^0(\epsilon_n)) = 0$. We have that $P_n D(p_n^0(\epsilon_n))$ is given by

$$P_n \frac{\{m_1 - r(p_n^0(\epsilon_n))\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n r(p^0(\epsilon_n))\right)}{\sigma^2}.$$

Because $r(p_n^0(\epsilon)) = r(p_n^0)$, it follows that $P_n D(p^0(\epsilon_n))$ is given by

$$P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n r(p_n^0)\right)}{\sigma^2},$$

but the latter equals zero by the fact that $P_n S(\epsilon_n) = 0$ (12). This proves that, if $m(A,W \mid \beta)$ is linear in β, then the targeted maximum likelihood estimator is achieved in the first step of the algorithm and solves the efficient influence curve estimating equation $P_n D(p) = 0$. If one would also update $\sigma^2(A,W)$ in the submodel $p_n^0(\epsilon)$, then the algorithm would have to be iterated in order to converge to a targeted MLE solving $P_n D(p) = 0$. Similarly, for nonlinear models $m(A,W \mid \beta)$ the targeted MLE algorithm will also need to be iterated till convergence.

7 Targeted MLE as loss based estimation.

In the previous sections we defined a targeted MLE in terms of an initial density estimator and the targeted MLE algorithm applied to this initial density estimator. In order to provide a general data adaptive likelihood based approach for construction of targeted MLE's (also allowing for an integrated data adaptive approach for searching over the initial densities, just as in sieve based MLE), we now note that the targeted MLE approach corresponds with a particular modified log-likelihood loss function. Specifically, let

$$L(p \mid P_0) \equiv -\log p^*(p),$$

where $p^*(p)$ is defined as the limit for $k \to \infty$ of the targeted MLE applied to P_0 and starting at p:

$$p^{k+1} = arg \max_{p \in \{p^k(\epsilon):\epsilon\}} P_0 \log p. \qquad (13)$$

Note that $L(p \mid P_0)$ is a loss function for densities p of the data indexed by unknown nuisance parameters, since the $\epsilon_0^k \equiv \arg\max_\epsilon P_0 \log p^k(\epsilon)$ are unknown. However, estimation of the unknown nuisance parameter corresponds simply with applying the targeted MLE algorithm to the data starting at p. The loss function satisfies

$$p_0 = \arg\min_{p \in \mathcal{M}} P_0 L(p \mid P_0),$$

because $p^*(p_0) = p_0$ and $p_0 = \arg\min_{p \in \mathcal{M}} -P_0 \log p$. Therefore, we can apply the unified loss based learning approach presented in van der Laan and Dudoit (2003) based on this new loss function $L(p \mid P_0)$ for a candidate density p. Succinctly, this loss based learning approach works as follows. Let $\mathcal{M}_s \subset \mathcal{M}$ be a sieve of $\mathcal{M}$ indexed by fine tuning parameters s. Let

$$p_{sn} = \hat{\Phi}_s(P_n) \equiv \arg\min_{p \in \mathcal{M}_s} P_n L(p \mid P_n) = \arg\max_{p \in \mathcal{M}_s} P_n \log p_n^*(p),$$

where $p_n^*(p)$ represents the limit density of the targeted MLE algorithm starting at p applied to the data P_n. Note that this maximization corresponds with maximizing the log likelihood over solutions of $P_n D(p^*) = 0$, where the $p^* = p^*(p)$ is restricted by the constraints on the initial p. We can select s with likelihood based cross-validation:

$$s_n = \hat{S}(P_n) \equiv \arg\min_s E_{B_n} P^1_{n,B_n} L(\hat{\Phi}_s(P^0_{n,B_n}) \mid P^0_{n,B_n}),$$

resulting in the targeted ML density estimator

$$p_n \equiv p_{s_n n} = \hat{\Phi}_{\hat{S}(P_n)}(P_n)$$

and targeted ML estimator of ψ_0 given by $\psi_n = \Psi(p_n)$.

DOI: 10.2202/1557-4679.1043

8 Discussion.

In this article we assumed a model in terms of densities with respect to a known dominating measure, and our targeted MLE density estimators are assumed to be dominated by this dominating measure. This allowed us to simplify the presentation of the method. However, we also wish to stress that the presented targeted maximum likelihood estimation methodology can easily be generalized to targeted maximum likelihood estimation in models in terms of probability distributions including (say) discrete as well as continuous distributions, just as this is common practice in maximum likelihood estimation in semiparametric models. The targeted MLE algorithm takes as input an initial density with respect to a specified dominating measure, and is based on a hardest submodel in terms of densities with respect to this same dominating measure. Thus, the targeted MLE algorithm can be applied to discrete distributions as well as continuous distributions, and as a consequence, the (loss based) targeted MLE learning as presented in Section 7 applies to models that are not necessarily dominated by a single dominating measure.

As a further generalization, the iterative principle underlying this work can be applied to loss functions other than the negative log likelihood. Given a loss function defined on the data and parameter space (and possibly a nuisance parameter η), we can make a one-dimensional ϵ-extension through a space containing both the parameter Ψ and nuisance parameter η, initialize the parameter estimate at $\Psi(0)$, and then update the parameter estimate by choosing ϵ to minimize the empirical risk $\frac{1}{n}\sum_{i=1}^{n} L(O_i, \Psi(\epsilon)|\eta(\epsilon))$. The requirement underlying the procedure is that $\frac{d}{d\epsilon}L(O, \Psi(\epsilon)|\eta(\epsilon))|_{\epsilon=0}$ is equal to an estimating equation for the parameter Ψ. If this condition is met, then solving this estimating equation should correspond to convergence of the iterative empirical risk minimization algorithm. Hence, applying the algorithm with such a loss function $L(O, \Psi|\eta)$ leads to a fusion of general loss based estimation and estimating function methodology.

Given a density estimator we defined a targeted density estimator through an iterative maximum likelihood algorithm along hardest submodels with a score equal to the efficient influence curve of the parameter of interest. This tool allows us to map any candidate density p into its targeted version $p_n^*(p)$. We now showed that by using the minus log density as loss function and thereby use the log-likelihood criteria in combination with the cross-validated log-likelihood criteria, *but restricted to targeted density estimators only*, we can build data adaptive sieve based algorithms for generating a final targeted ML density estimator and corresponding substitution estimator of the parameter of interest.

By restricting the log-likelihood criteria and cross-validated log-likelihood criteria to targeted densities only, targeted maximum likelihood estimation provides now a purely likelihood based methodology for estimation of any kind of parameter such as pathwise differentiable parameters and infinite dimensional parameters: see our accompanying technical report.

In particular, we showed that targeted maximum likelihood estimation completely unifies maximum likelihood estimation and estimating function based estimation, and results in important improvements in both. Targeted MLE also deals naturally with the issue of multiple solutions of estimating equations by using the log-likelihood as the criteria to be maximized. Another nice feature of targeted MLE is that it always improves on the initial density estimator by increasing the log-likelihood fit. As a consequence, when targeted MLE is applied to estimate pathwise differentiable parameters of a full data distribution F_X in CAR censored data models as in (van der Laan and Robins (2003)), if one applies the targeted MLE to an initial $p_n^0 = (g_n^0, Q_n^0)$ with g_n^0 and Q_n^0 being fits of the censoring mechanism g_0 and the F_X-factor Q_0 of the density p_0, then it provides an estimator which is guaranteed to be more efficient than the double robust IPCW estimator based on estimating the nuisance parameters (g_0, Q_0) with p_n^0. So the targeted MLE algorithm provides a natural way to always improve on any initial double robust IPCW locally efficient estimator as presented in van der Laan and Robins (2003).

References

P.J. Bickel, A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive inference in semiparametric models.* Johns Hopkins university press, Baltimore, 1993a.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer-Verlag, 1993b.

S.R. Cosslett. Efficient semiparametric estimation of censored and truncated regressions via smooth self-consistency equation. *Econometrica*, 72(4):1277–1284, 2004.

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

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of statistics*, 19(4):2244–2253, December 1991.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23:774–86, 1995.

C.A.J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15:1548–1562, 1987.

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

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

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

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

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

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser, 1992.

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

M.J. van der Laan. *Efficient and Inefficient Estimation in Semiparametric Models.* Centre of Mathematics and Computer Science (CWI), Amsterdam, 1995.

M.J. van der Laan. Identity for npmle in censored data models. *Lifetime Data Models*, 4(0):83–102, 1998.

M.J. van der Laan. Statistical inference for variable importance. *International Journal of Biostatistics*, 2(1), 2006.

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

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality.* Springer, New York, 2003.

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

M.J. van der Laan and D. Rubin. Estimating function based cross-validation. In J. Fan and H.L. Koul, editors, *Frontiers of Statistics*, pages 87–108. Imperial College Press, 2006.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, 1996a.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes.* Springer-Verlag New York, 1996b.

Z. Yu and M.J. van der Laan. Measuring treatment effects using semiparametric models. Technical report, Division of Biostatistics, University of California, Berkeley, 2003.

DOI: 10.2202/1557-4679.1043