University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2006 *Paper* 213

Targeted Maximum Likelihood Learning

Mark J. van der Laan* Daniel Rubin†

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

†Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

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 has available an estimate of the density of the data generating distribution such as a maximum likelihood estimator according to a given or data adaptively selected model. Suppose that one is concerned with estimation of a particular pathwise differentiable Euclidean parameter. A substitution estimator evaluating the parameter of the 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 this parameter 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 an efficient or locally efficient estimator of the parameter of interest under regularity conditions. We also 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. This tool provides us with a new class of targeted likelihood based estimators of pathwise differentiable parameters.

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. Finally, we show that targeted maximum likelihood estimation can be generalized to estimate any kind of parameter, such as infinite dimensional non-pathwise differentiable parameters, by restricting the likelihood and cross-validated log-likelihood to targeted candidate density estimators only. We illustrate the method with various worked out examples.

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 will 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 (2002)). 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 MLE's p_{ns}^* with likelihood based cross-validation, as explained in Section 2.5. 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. In addition, we discuss an important option for the targeted MLE which also allows it to update nuisance parameters which are needed to estimate the efficient influence curve. 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. 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 (2002). 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. Some

additional important benefits of targeted MLE relative to estimating function based estimation are provided in the discussion. In Section 6 we present important examples illustrating the power and computational simplicity of this new targeted maximum likelihood estimator: estimation of marginal means, marginal causal effects and the parametric component in a semiparametric regression model. We also provide a simulation to illustrate the targeted MLE of a marginal causal effect. 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), which provides a general template for searching among targeted MLE's indexed by the initial density. In Section 8 we present a potentially very promising alternative approach of targeted MLE learning relative to targeted MLE learning as presented in Section 7, which yields a particularly powerful approach for selecting nuisance parameter fits.

In Section 9 we present a study of the targeted MLE algorithm $P_n \rightarrow \Phi^*(P_n)$ (viewed as a mapping from the empirical probability distribution to the resulting density) in the case that P_n is replaced by the true distribution P_0, in particular, in relation to its starting density p^0. This will help us understand that the targeted MLE algorithm can be viewed as an algorithm providing bias reduction with respect to ψ_0 at each step, and that it is an algorithm which quickly converges to a solution of the equation setting the expectation under the true distribution of the efficient influence curve at a candidate density p equal to zero. In particular, in the case of convex models and linear parameters (typically a single step suffices) it always converges to a solution for which the parameter of interest is correctly identified. In Section 10 we show how the targeted MLE can be applied to estimate the risk function in unified loss function based learning of arbitrary (including non-pathwise differentiable) parameters, as presented in van der Laan and Dudoit (2003). In Section 11 we show how the targeted MLE principle can be used directly to provide a completely data driven targeted maximum likelihood learning methodology of any kind of parameter (finite dimensional or infinite dimensional). In Section 12 we illustrate this general targeted MLE methodology to data adaptively estimate W-adjusted variable importance $E_0(Y \mid A, W) - E_0(Y \mid A = 0, W)$ based on observing n i.i.d. copies of (W, A, Y). We end this article with a discussion in Section 13.

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 (1995b) 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 has recognized this problem such as in smoothing survival functions or smoothing the nonparametric components in a semiparametric regression model, noting that so called "under-smoothing" is needed 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. (1997). 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 related literature on estimation function based estimation of pathwise differentiable parameters based on censored data we refer to van der Laan and Robins (2002).

A general loss function-based approach for model selection and estimation, thereby including maximum likelihood estimation as a special case, is described in Barron et al. (1999). It uses sieve theory to define penalized empirical risk criteria. In particular, Barron (1989) and Barron (1991) develop this theory in the context of artificial neural networks. Connections with cross-validation methods are discussed in Birgé and Massart (1997). Barron et al. (1999) and Birgé and Massart (1997) have studied thoroughly the penalty functions to be used in adaptive estimation on sieves. They use powerful Talagrand concentration and deviation inequalities for empirical processes (Ledoux, 1996; Massart, 1998; Talagrand, 1996a,b) to obtain

so-called oracle inequalities for the theoretical risk of their estimators. The method of oracle inequalities was also used to prove minimax optimality properties of nonparametric estimators in Johnstone (1998). The Birgé-Massart penalties are based on the dimension of the classes of functions. This approach has been shown to perform well for sieves that frequently occur in nonparametric univariate regression and nonparametric univariate density estimation problems (e.g., nested families of Sobolev ellipsoids). In van der Laan et al. (2003) we prove an oracle inequality for likelihood based cross-validation. In van der Laan et al. (2006) and van der Vaart et al. (2006) we prove general results for loss based estimation based on ϵ-nets, and oracle inequalities for V-fold cross-validation, respectively. In the all of the above references the loss function is a known function of the experimental unit and a candidate parameter value whose expectation is minimized by the true parameter value, and, as a consequence, it only applies to a limited number of parameters.

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 (k-th Step) Targeted maximum likelihood estimators.

Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ 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 $p_{\delta=0} = p$

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

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 (2002), Chapter 1). As in van der Laan and Robins (2002), 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. (1993)), as we will also do in this article.

Let $(O, p) \to D(p)(O)$ 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

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

Example 1 (Cumulative Distribution function at a point in a non-parametric model) Let 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

$\Psi : \mathcal{M} \to \mathbb{R}$ 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 can 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 $\{D_h(p) : h \in \mathcal{H}\} \subset T^{\perp}_{nuis}(p)$ (or equality) is presented in van der Laan and Robins (2002). In van der Laan and Robins (2002) 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 $p \to \Psi(p)$ of the form $p \to D_h(\Psi(p), \Upsilon(p))$ with Υ representing a nuisance parameter. In van der Laan and Robins (2002), for a variety of general classes of models and censored data structures O, explicit representations of the orthogonal complement of the nuisance tangent space, $T^{\perp}_{nuis}(p)$, corresponding gradients, $T^{\perp *}_{nuis}(p)$, 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 $\epsilon \to p_n^0(\epsilon) \in \mathcal{M}$ 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 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 $\Psi : \mathcal{M} \to \mathbb{R}^d$ 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

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

assuming this limit exists.

2.1 Example: Targeted maximum likelihood estimation of the cumulative distribution function (CDF) in a nonparametric model.

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

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

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,

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

Its derivative is,

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

Its second derivative is easily seen to be,

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

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,

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

$$
\begin{aligned}
&= \int_{-\infty}^{t} f(o)do + \epsilon_n \int_{-\infty}^{t} I(o \leq t) f(o)do - \epsilon_1 F(t) \int_{-\infty}^{t} f(o)do \\
&= F(t) + \epsilon_n F(t) - \epsilon_n F(t)^2 = F(t) + \epsilon_1 F(t)(1 - F(t)) \\
&= F(t) + \frac{F_n(t) - F(t)}{F(t)(1 - F(t))} F(t)(1 - F(t)) \text{ from (7)} \\
&= F(t) + F_n(t) - F(t) = F_n(t).
\end{aligned}
$$

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.

2.2 Some remarks about the targeted maximum likelihood estimator

One particular (candidate) estimator of ψ_0 is to estimate ψ_0 with $\psi_n \equiv \lim_{k\to\infty} \Psi(p_n^k)$, assuming this limit exists as established in the next section if $\epsilon(P_n \mid p_n^k) \to 0$ for $k \to \infty$. The resulting estimator ψ_n is for all practical purposes equivalent with the k-th step targeted maximum likelihood estimator, for sufficiently large k.

We also note that if for a given $k \in \{0, 1, \ldots\}$, $P_n D(p_n^k) = 0$, then $p_n^m = p_n^k$ for $m = k + 1, \ldots$ and thus $\psi_n = \Psi(p_n^k)$, due to the fact that the MLE of ϵ equals zero at step k (since the value $\epsilon = 0$ solves the derivative $P_n D(p_n^k)$ of $\epsilon \to P_n \log p_n^k(\epsilon)$ and it is assumed that this corresponds with the maximum). For example, below we provide various examples in which p_n^1 already solves the optimal estimating equation $P_n D(p_n^1) = 0$ so that the targeted maximum likelihood estimator is achieved in the first step: $\psi_n = \Psi(p_n^1)$. Although, we will show that the consistency and asymptotic linearity (and sometimes even the efficiency) of the targeted MLE ψ_n in *convex models for linear parameter* Ψ does not depend on the asymptotic properties of p_n^0, in general, the targeted maximum likelihood estimator ψ_n of ψ_0 will typically depend on the (second order behavior of the) initial estimator p_n^0, so that it is typically important to obtain a likelihood based estimator p_n^0 of p_0 with reasonable consistency properties. Specific proposals for data adaptive searches among candidate initial density estimators will be provided in this article. In particular, we will argue that one can use likelihood based cross-validation to select among targeted MLE's indexed by different choices of p_n^0.

As a potentially useful practical modification of this targeted MLE algorithm $\Phi^*(P_n)$ for ψ_n we suggest that at each step one does not necessarily need to select the maximizer $\epsilon(P_n \mid p_n^k)$, but instead one might simply select an ϵ so that $P_n \log p_n^k(\epsilon) > P_n \log p_n^k$, thereby still guaranteeing that the likelihood increases at each step. The important property driving the asymptotics of the resulting estimator is that the algorithm is such that for k converging to infinity the likelihood increases at each step, and (as a consequence) the maximizer of $\epsilon \to P_n \log p_n^k(\epsilon)$ converges to zero so that in the limit $\lim_k P_n D(p_n^k) = 0$. A similar type of modification is provided below in the Section 2.4.

2.3 Targeting the nuisance parameter estimate in each step of the targeted MLE.

Since the efficient influence curve $D(p)$ at p is orthogonal to the nuisance tangent space $T_{nuis}(p)$ (i.e., the closure of the linear span of the scores of paths through p for which the pathwise derivative equals 0), the so called hardest submodel $p(\epsilon)$ through a $p \in \mathcal{M}$ with score $D(p)$ at $\epsilon = 0$ might not provide any updates for certain nuisance parameters $\Gamma(p)$ which are actually needed to evaluate $D(p)$: say $D(p) = D(Q(p), \Gamma(p))$. For example, the efficient influence curve of a survival function based on right censored data depends on the censoring mechanism, but the projection of the efficient influence curve on the space of scores of the censoring mechanism equals zero so that the hardest submodel does not need to update the censoring mechanism. As a consequence, in these situations the submodel $p(\epsilon)$ will not fluctuate the nuisance parameter Γ, so that the resulting targeted maximum likelihood estimator will solve an estimating equation $P_n D(Q(p), \gamma_n^0) = 0$ with $\gamma_n^0 = \Gamma(p_n^0)$ equal to the initial estimator of γ_0 corresponding with the initial density estimator p_n^0. Although the efficient influence curve's dependence on Γ is such that the properties of γ_n^0 only affects the second order terms in the resulting targeted ML estimator of ψ_0 (see e.g. van der Laan and Robins (2002)), it is still of practical interest to also make the estimator of γ_0 targeted towards estimation of $D(p_0)$. This can be achieved by selecting a $p(\epsilon)$ with scores in the tangent space for Γ at p so that the targeted MLE also involves iteratively updating $\Gamma(p^k)$. Specifically, one might add ϵ components to $p(\epsilon)$ (submodel through p at $\epsilon = 0$) with score at their null values equal to the efficient influence curve of a parameter $p_1 \to E_p D_1(Q(p), \Gamma(p_1))$

at $p_1 = p$, where D_1 is such that $D(p) = D_1(p) + D_2(p)$ for an appropriately chosen decomposition of $D(p)$ in terms $D_1(p)$ and $D_2(p)$. For example, in CAR censored data models considered in Section 5.1 we use the decomposition $D(p) = D_1(p) - \Pi(D_1(p) \mid T_\Gamma(p))$, where $T_\Gamma(p)$ is a nuisance space for the parameter $\Gamma(p)$ and Γ represents the censoring mechanism. The latter parameter $p_1 \rightarrow E_p D_1(Q(p), \Gamma(p_1))$ represents now a function of Γ relevant for identifying the mean of the efficient influence curve, and thereby can be used to indirectly target the estimator of γ_0 towards estimation of ψ_0. Selecting as score at $\epsilon = 0$ the efficient influence curve of the parameter $p_1 \rightarrow E_p D(Q(p), \Gamma(p_1))$ does not work, since the latter function has directional derivatives equal to zero, due to the fact that $D(Q(p), g(p))$ is orthogonal to the nuisance tangent space $T_{nuis}(p)$. We illustrate this approach for estimation of a parameter of the full data distribution based on a coarsening at random censored data observed data structures in the Section , and in Section 6 through specific examples.

2.4 Modified one-step targeted MLE.

In practice the following one-step approach applied to an initial density estimator p_n^0 of p_0 might result in estimators of ψ_0 with essentially the same practical performance as the targeted MLE $\psi_n = \lim_k \Psi(p_n^k)$ based on the same p_n^0, or it might simply be an alternative approach to finding the targeted MLE.

Consider an initial estimator p_n^0 of p_0. As above, let $\{p_n^0(\epsilon) : \epsilon \in \mathbb{R}^k\} \subset \mathcal{M}$ be a (hardest) submodel through p_n^0 at $\epsilon = 0$ with score $D^*(p_n^0)$ at $\epsilon = 0$ whose linear span includes the components of the canonical gradient $D(p_n^0)$ at p_n^0. Now, instead of letting the algorithm maximize the likelihood in ϵ and iterating the process till one converges to a solution p_n^∞ of $P_n D^*(p) = 0$, we suggest that one might also search for the solution ϵ_n of

$$P_n D^*(p_n^0(\epsilon)) = 0.$$

This corresponds with solving k equations in k unknowns. If this solution does not exist or requires an iterative algorithm to find, then this approach is not more practical than the iterative maximum likelihood algorithm defining the targeted MLE. If this solution ϵ_n exists it might not correspond with the actual maximum likelihood estimator in ϵ of $\epsilon \rightarrow P_n \log p_n^0(\epsilon)$, but we conjecture that it will typically still increase the likelihood relative to $\epsilon = 0$

(and thus p_n^0), and by solving $P_n D^*(p_n^0(\epsilon_n)) = 0$ the theorems in Section 4 for proving local efficiency of $\Psi(p_n^0(\epsilon_n))$ apply. In addition, we note that by setting $p_n^1 = p_n^0(\epsilon_n)$ the targeted MLE approach starting with p_n^1 will not provide any changes in the sense that $p_n^k = p_n^1$ for $k = 2, \ldots$, due to the fact that $P_n D^*(p_n^1) = 0$, and thus this estimator equals the targeted MLE starting at p_n^1. The following example illustrates this approach.

Example 2 (Smooth efficient estimation of a cumulative distribution function (CDF) at several points) In this case we have that the efficient influence curve of the cumulative distribution function $\Psi(p) = (\int_0^{t_j} p(o)d\mu(o) : j = 1, \ldots, d)$ at d points equals $D(p)(O) = \{I(O \leq t_j) - \Psi(p)(t_j) : j = 1, \ldots, d\}$. Let p_n^0 be an initial density estimator of p_0, and let $p_n^0(\epsilon) = (1+\epsilon^\top D(p_n^0))p_n^0$ be a d-dimensional parametric submodel with parameter $\epsilon = (\epsilon_1, \ldots, \epsilon_d)$ with score $D(p_n^0)$ at $\epsilon = 0$. Instead of solving for the MLE ϵ_{MLE} of $\epsilon \to P_n \log p_n^0(\epsilon)$, which corresponds with solving $P_n D(p_n^0)/(1+\epsilon^\top D(p_n^0)) = 0$, we can also simply solve

$$\begin{aligned} 0 &= P_n D(p_n^0(\epsilon)) = P_n D((1 + \epsilon^\top D(p_n^0))p_n^0) \\ &= P_n D(p_n^0) - E_{p_n^0} D(p_n^0) D(p_n^0)^\top \epsilon, \end{aligned}$$

where $\Sigma(p_n^0) \equiv E_{p_n^0} D(p_n^0) D(p_n^0)^\top$ equals the covariance matrix of $(I(O \leq t_j) : j = 1, \ldots, d)$ under p_n^0. Thus, the solution ϵ_n of this equation is given by:

$$\epsilon_n = \Sigma(p_n^0)^{-1} P_n D(p_n^0).$$

Thus, given a smooth distribution function $\int_0^{\cdot} p_n^0(o)d\mu(o)$ based on a density estimator p_n^0, we can map this smooth CDF into a CDF which agrees with the empirical distribution function at a user supplied set of points $t_1, \ldots, t_d$ given by:

$$\tilde{F}_n(\cdot) = \int_0^{\cdot} (1 + (\Sigma(p_n^0)^{-1} P_n D(p_n^0))^\top D(p_n^0)(o))p_n^0(o)d\mu(o).$$

This provides an interesting explicit methodology for construction of efficient smooth estimators of a cumulative distribution function. The explicitness of this estimator only relies on the linearity of the parameter $p \to \Psi(p)$, and the convexity of the model so that the used $p_n(\epsilon)$ is an appropriate submodel. We also note that the efficiency of the resulting estimator of the cumulative distribution function at these points does not depend on the initial density estimator p_n^0.

2.5 Philosophy of using the (cross-validated) log-likelihood to select among candidate targeted MLE's

Above we proposed that for the purpose of selecting among different targeted MLE's (e.g. indexed by different initial density estimators) of a pathwise differentiable parameter of interest, it is appropriate to use likelihood based cross-validation, while (seemingly contradictory) we also argued that likelihood based cross-validation would be inappropriate for selecting among general candidate density estimators. Firstly, one should note that this data adaptive targeted MLE approach differs in a crucial way from the current sieve based maximum likelihood methodology involving 1) proposing a sieve (say indexed by s as given previously) $\mathcal{M}_s \subset \mathcal{M}$ on the model $\mathcal{M}$, 2) computing the maximum likelihood estimator for each element of the sieve, 3) using likelihood based cross-validation to select among the resulting candidate maximum likelihood estimators, and 4) estimating the parameter of interest with the substitution estimator. Namely, we use the log-likelihood as criteria (both the empirical log-likelihood and the cross-validated log likelihood) to compare *targeted* MLE's instead of regular MLE's, where targeted MLE solve the efficient influence curve equation $P_n D(p) = 0$.

In order to understand the heuristic behind using the log-likelihood as criteria restricted to targeted MLE only, it helps to consider an infinite data set so that P_n is replaced by P_0 and the targeted MLE $p^*(p)$ starting at p is a solution of $P_0 D(p*) = 0$. For example, consider convex models and linear parameters so that $P_0 D(p) = \Psi(p_0) - \Psi(p)$ for any p with p_0/p bounded (van der Laan (1998)), and consider two candidate targeted MLEs p_1^*, p_2^*. Since they are targeted MLE's they solve $P_0 D(p_1) = 0$ and $P_0 D(p_2) = 0$. Thus $\Psi(p_1^*) = \Psi(p_2^*) = \psi_0$. Since they already perfectly nail down the parameter of interest ψ_0 a comparison of the log likelihoods $P_0 \log p_1^* - P_0 \log p_2^*$ now evaluates the performance of the nuisance part of the densities p_1^*, p_2^*. So our claim corresponds in this case with stating that the log likelihood loss function provides a sensible criteria for comparing densities which perform equally good with respect to the parameter of interest ψ_0. In general, we have that $P_0 D(p) \approx \Psi(p_0) - \Psi(p)$. In this case, solving $P_0 D(p) = 0$ does not guarantee that $\Psi(p) = \psi_0$. However, any deviation of $\Psi(p_0) - \Psi(p)$ from zero is now due to nuisance parameters needed to identify $D(p)$ (e.g. $D(p) = D(\Psi(p), \Upsilon(p))$ and $P_0 D(\psi, \upsilon_0) = 0$ is uniquely solved by ψ_0). Therefore, for two targeted densities p_1^*, p_2^* solving $P_0 D(p_1^*) = P_0 D(p_2^*) = 0$, it makes still sense to now just compare them by their log-likelihood risk.

For finite samples the principle idea behind is that if two density estimators, differing only in fits of parameters needed to identify $D(p_0)$, have been fully targeted towards fitting a certain pathwise differentiable parameter (by applying the targeted MLE algorithm using them as initial density estimator), then a difference in log-likelihood now reflects the performance of the two density estimators in estimating the needed (for the purpose of the pathwise differentiable parameter) nuisance parameters to nail down the parameter of interest, while if these two density estimators would be non-targeted, then a difference in log-likelihood can be due to a difference in how much each of them has been fully targeted to fit the parameter of interest as well as how well it fits the required nuisance parameters. For example, consider two possible increases in fit of the nuisance parameters needed to fit $D(p_0)$, but suppose that one of the fits results in a large gain of the log-likelihood during the targeted MLE algorithm, while for the other fit the targeted MLE algorithm yields only a small increase in log-likelihood. Then a comparison of the log likelihood for the two targeted fits will select the increase in nuisance parameter fit which results in the subsequent maximal increase in log-likelihood during the targeted MLE algorithm. That is, by using the log-likelihood to only compare targeted density estimators solving the efficient influence curve estimating equation, the criteria rewards increases in fits of the density which are directly relevant for estimation of the parameter of interest.

In particular, an increase in fit of nuisance parameters which are not needed to evaluate the efficient influence curve $D(p_0)$ will result in a zero increase during the targeted MLE algorithm. However, they will still increase the log-likelihood because the log-likelihood's starting value for the targeted MLE algorithm has increased. In order to avoid such irrelevant increases in fit for the purpose of fitting the parameter of interest, it is important to generate candidate initial density estimators (which will be inputted in the targeted MLE algorithm) which mainly differ in how they estimate the parameters needed to identify the efficient influence curve $D(p_0)$. This can be partly arranged by choosing the model $\mathcal{M}$ as small as possible without changing the efficient influence curve of the parameter of interest and how one would estimate the efficient influence curve (e.g., assuming normal error distributions if the parameter is a parameter of regressions and the efficient influence curve only depends on the first and second moment of the conditional distribution of the outcome given the covariates), *and* by using a sieve within the model $\mathcal{M}$ which only changes the candidate density estimators

with respect how they will fit the parameters needed to identity the efficient influence curve $D(p_0)$. In Section 8 we propose to use the actual increase of the log-likelihood during the targeted MLE algorithm as a way of evaluating among different nuisance parameter fits in Section 8.

3 The targeted maximum likelihood estimator solves the efficient influence curve 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 $k \to \infty$ (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 $P_n D(p_n^k) \to 0$.

Result 1 *Let P_n be given. Assume that*

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

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 $\limsup_{k \to \infty} \parallel A_k \parallel < \infty$, 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 $\epsilon(P_n \mid p_n^k) \to 0$ for $k \to \infty$, then we have

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

The condition (9) holds if the score of the one-dimensional submodel $p(\epsilon)$ at ϵ converges to the score at $\epsilon = 0$ for $\epsilon \to 0$ 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 $\epsilon(P_n \mid p_n^k) \to 0$. 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 the targeted maximum likelihood estimator.

In this section we provide templates for proving consistency, asymptotic linearity and efficiency of the targeted maximum likelihood estimator of a pathwise 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 Local efficiency of targeted maximum likelihood estimator of 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 $D_0(p_1) \equiv D(p_1) - P_0 D(p_1)$:

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

where we note that p_1 can be an arbitrary limit (i.e., $p_1 \neq p_0$ 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 $P_n D(p_n) = R(n, K(n)) = o_P(1/\sqrt{n})$ (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 $D(p_1) - P_0 D(p_1)$:

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

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 3 ((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 4 ((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 $\Psi : \mathcal{M} \to \mathbb{R}$ 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

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

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

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

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 (1995b)), 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 of the targeted maximum likelihood estimator for general smooth parameters under a consistency-rate condition on the initial density estimator.

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 $p_0/p < \infty$. Our bias reduction results for the targeted MLE algorithm in Section 9 show that the first step in the targeted MLE algorithm applied to a consistent initial density estimator reduces the rate at which the asymptotic bias of $\Psi(p_n^0) - \psi_0$ converges to zero to the rate at which second order terms in $p_n^0 - p_0$ converge to zero, suggesting that the first step might in many situations already yield an efficient estimator. However, it will be beyond the scope of this article to establish formal results for the targeted MLE in this article in more detail.

5 Fusion of targeted maximum likelihood estimation and estimating function based estimation.

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 (2002)), 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 (2002)) 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 $p \rightarrow T_{nuis}^{\perp}(p)$, 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 (2002)) 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) maxi-

mum 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.

We also note that the targeted MLE is more widely applicable 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. 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 Targeted maximum likelihood estimation of pathwise differentiable parameters in 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 (2002)). 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 (2002)). 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 $L_0^2(P)$ 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 ψ_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.

In particular, the one-step method to obtain a targeted MLE $p^0(\epsilon_n)$ solving the efficient influence curve estimating equation, presented in the previous subsection 2.4, corresponds now with solving

$$P_n D(g(p^0), Q^0(\epsilon)) = 0. \tag{11}$$

In particular, if $\mathcal{M}$ is convex and $\Psi(Q)$ is linear, then $Q \to D(g, Q)$ is linear (van der Laan and Robins (2002)). As a consequence, in this case, if one uses a model $Q^0(\epsilon) = (1 + \epsilon D(g^0, Q^0))Q^0$ for ϵ in an appropriate range, then the equation (11) is linear in ϵ and thus allows a closed form solution ϵ_n. One would need to check if this choice ϵ_n correspond with a valid density $p^0(\epsilon_n)$, and that it results in an increase of the likelihood relative to p^0 (as one certainly expects to be the case, since it solves the score equation at $\epsilon = 0$ of the next hardest parametric submodel).

This targeted maximum likelihood estimation of a pathwise differentiable parameter of the full data distribution of X relies on having an estimator $g(p^0)$ of the censoring mechanism $g_0 = g(p_0)$, and the targeted MLE provides no further updates of this estimator. Because of the factorization of the likelihood, estimation of g_0 can be based on the log-likelihood $g \to P_n \log g(O \mid X)$ and can thus be achieved with standard maximum likelihood estimation and likelihood based cross-validation to control the bias and variance trade-off. However, one might wish to also make a likelihood based estimator g_n^0 targeted to our goal of estimation of $\Psi(Q_0)$. As pointed out in Section 2.3, the definition of targeted maximum likelihood estimator allows one to also create extra scores for orthogonal nuisance parameters such as g for the purpose

of making the maximum likelihood for g targeted towards estimation of the efficient influence curve, which happens to depend on g_0.

5.2 Targeted maximum likelihood estimation in CAR censored data models: including 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 (2002) (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 $L_0^2(p)$. In order to relate these functions to estimating functions for ψ_0 (as in van der Laan and Robins (2002)) 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 $p_n^0 = (g_n^0, Q_n^0)$, in the censored data literature one defines the IPCW-estimator and DR-IPCW estimator as the solutions of the estimating equations $P_n D_{IPCW}(g_n^0, \psi) = 0$ and $P_n D(g_n^0, Q_n^0, \psi) = 0$, respectively, and $\Psi(Q_n^0)$ is called the likelihood based estimator (making the assumption that Q_n^0 is likelihood based).

We will now describe the targeted MLE algorithm also involving the updating of g_n^0. At step k it now involves also a parametric submodel $g(p_n^k)(\epsilon_2)$ through $g(p_n^k)$ with score $D_{CAR}(g_n^k, Q_n^k)$ 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 $k \to \infty$ (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 the targeted maximum likelihood estimator.

In this section we provide various important examples of the targeted MLE to illustrate its remarkable simplicity and good properties.

6.1 Targeted Maximum likelihood estimation of a mean in a nonparametric model.

Consider an initial data generating density p^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

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

Assume the density p^0 is such that the moment generating function of the random variable $w(O)$,

$$\phi^0(\epsilon) = E_{p^0}[\exp(\epsilon w(O))] = \int \exp(\epsilon w(x))p^0(x)d\mu(x),$$

its derivative

$$\phi^{0\prime}(\epsilon) = \int \frac{d}{d\epsilon}\exp(\epsilon w(x))p^0(x)d\mu(x) = \int w(x)\exp(\epsilon w(x))p^0(x)d\mu(x),$$

and its second derivative

$$\phi^{0\prime\prime}(\epsilon) = \int \frac{d}{d\epsilon}w(x)\exp(\epsilon w(x))p^0(x)d\mu(x) = \int w(x)^2\exp(\epsilon w(x))p^0(x)d\mu(x)$$

exist for all real-valued ϵ. For the exponential family

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

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

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

The first derivative is,

$$\begin{aligned} l'(\epsilon) &= \sum_{i=1}^n [w(O_i) - \psi^0 - \frac{\int (w(x)-\psi^0)\exp(\epsilon(w(x)-\psi^0))p^0(x)d\mu(x)}{\int \exp(\epsilon(w(x)-\psi^0))p^0(x)d\mu(x)}] \\ &= n\bar{W}_n - n\psi^0 \\ &\quad -n\frac{\exp(-\psi^0\epsilon)\int w(x)\exp(\epsilon w(x))p^0(x)d\mu(x) - \psi^0\exp(-\psi^0\epsilon)\int \exp(\epsilon w(x))p^0(x)d\mu(x)}{\exp(-\psi^0\epsilon)\int \exp(\epsilon w(x))p^0(x)d\mu(x)} \\ &= n\bar{W}_n - n\psi^0 - n\frac{\phi^{0\prime}(\epsilon)}{\phi^0(\epsilon)} + n\psi^0 = \bar{W}_n - n\frac{\phi^{0\prime}(\epsilon)}{\phi^0(\epsilon)}, \end{aligned} \tag{12}$$

where $\bar{W}_n$ denotes the sample mean $\frac{1}{n}\sum_{i=1}^n w(O_i)$. The second derivative is,

$$l''(\epsilon) = n\frac{[\phi^{0\prime}(\epsilon)]^2 - \phi^{0\prime\prime}(\epsilon)\phi^0(\epsilon)}{[\phi^0(\epsilon)]^2}.$$

Note that by Cauchy-Schwarz,

$$\begin{aligned}
|\phi^{0\prime}(\epsilon)| &= |\int w(x)\exp(\epsilon w(x))p^0(x)d\mu(x)| \\
&\leq \int |w(x)|(1)\exp(\epsilon w(x))p^0(x)d\mu(x) \\
&\leq \sqrt{\int w(x)^2\exp(\epsilon w(x))p^0(x)d\mu(x)\int \exp(\epsilon w(x))p^0(x)d\mu(x)} \\
&= \sqrt{\phi^{0\prime\prime}(\epsilon)\phi^0(\epsilon)},
\end{aligned}$$

implying that $l''(\epsilon) \leq 0$ for all ϵ. Because the log likelihood is concave, the maximum likelihood estimator of ϵ is any solution ϵ_1 of $l'(\epsilon_1) = 0$. Hence, by (12), if an mle ϵ_n exists then it solves,

$$\frac{\phi^{0\prime}}{\phi^0}(\epsilon_n) = \bar{W}_n.$$

Now, the mean of $w(X)$ under $p^0(\epsilon_n)$ is given by

$$\begin{aligned}
\int w(x)p^0(\epsilon_n)(x)d\mu(x) &= \frac{\int w(x)\exp(\epsilon_n(w(x)-\psi^0))p^0(x)d\mu(x)}{\int \exp(\epsilon_n(w(x)-\psi^0))p^0(x)d\mu(x)} \\
&= \frac{\exp(-\psi^0\epsilon_n)\int w(x)\exp(\epsilon_n w(x))p^0(x)d\mu(x)}{\exp(-\psi^0\epsilon_n)\int \exp(\epsilon_n w(x))p^0(x)d\mu(x)} \\
&= \frac{\phi^{0\prime}}{\phi^0}(\epsilon_n) = \bar{W}_n.
\end{aligned}$$

Therefore, given any initial density $p^0(\cdot)$ for O, the targeted one-step maximum likelihood estimator of the mean of $w(O)$ reduces to the sample mean $\bar{W}_n = \frac{1}{n}\sum_{i=1}^n w(O_i)$.

6.2 Targeted maximum likelihood estimation of a marginal causal effect in nonparametric model: submodel I

A locally optimal method for estimation of the causal effect of a time-independent treatment A in a semiparametric regression model has been

given in Robins and Mark (1992). 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), Robins et al. (2000) and Neugebauer and van der Laan (2002): see also van der Laan and Robins (2002).

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 $\Psi : \mathcal{M} \to \mathbb{R}$ 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, temperal ordering assumption, and randomzation assumption required for causal inference. In order to acknowledge that this parameter is of interest in general, van der Laan (2006b) 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 (2006b)). 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 will set $D(p) = S(p)$.

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

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

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

$$D_1(p)(O) \equiv D(p) - E_p(D(p) \mid A, W)$$

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

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 $P_n D_2(p_n^0) = 0$ (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 also define $D_0(p) = Y(I(A=1) - I(A=0))/g(p)(A \mid W)$ and $\theta(p)(A,W) = E_p(D_0(p) \mid A, W) = Q(p)(A,W)(I(A=1) - I(A=0))/g(p)(A \mid W)$, so that $D_1(p) = D_0(p) - \theta(p)$. As one dimensional submodel we consider the exponential family

$$
\left\{ p_n^0(\epsilon)(O) = p_n^0(W) g(p_n^0)(A \mid W) \frac{\exp(\epsilon(D_0(p_n^0)(O) - \theta(p_n^0)(A,W))) p_n^0(Y \mid A, W)}{E_{p_n^0}(\exp(\epsilon(D_0(p_n^0)(O) - \theta(p_n^0)(A,W)) \mid A, W)} : \epsilon \right\}. \tag{13}
$$

To compute the first step targeted MLE we need to estimate ϵ with maximum likelihood based on an i.i.d. sample $\{O_i\}_{i=1}^n$.

Given any density p, we define

$$
\phi_p(\epsilon \mid A, W) = E_p[\exp(\epsilon D_0(p)(O)) \mid A, W], \tag{14}
$$

its derivative

$$
\phi_p'(\epsilon) = E_p[D_0(p)(O) \exp(\epsilon D_0(p)(O)) \mid A, W), \tag{15}
$$

and its second derivative

$$
\phi_p''(\epsilon) = E_p[D_0(p)(O)^2 \exp(\epsilon D_0(p)(O)) \mid A, W], \tag{16}
$$

and we assume these integrals exist for all real-valued ϵ at $p = p_n^0 \in \mathcal{M}$. The log likelihood can be expressed in terms of these functions as follows:

$$
l(\epsilon) = \sum_{i=1}^n \{\log(p_n^0(O_i)) + \epsilon(D_0(p_n^0)(O_i) - \theta(p_n^0)(A,W))\}
$$

$$-\log\left(E_{p_n^0}(\exp(\epsilon(D_0(p_n^0)(O)-\theta(p_n^0)(A,W)))\mid A,W)\right). \quad (17)$$

Analogous to the algebraic derivation at (12) it follows that the first derivative and second derivative of $l(\epsilon)$ are given by

$$l'(\epsilon) = \frac{1}{n}\sum_{i=1}^{n}\left\{D_0(p_n^0)(O_i)-\frac{\phi'_{p_n^0}(\epsilon\mid A_i,W_i)}{\phi_{p_n^0}(\epsilon\mid A_i,W_i)}\right\},$$

and

$$l''(\epsilon)=P_n\frac{[\phi'_{p_n^0}(\epsilon)]^2-\phi''_{p_n^0}(\epsilon)\phi_{p_n^0}(\epsilon)}{[\phi_{p_n^0}(\epsilon)]^2},$$

respectively, where we treat $\phi'_{p_n^0}$ as a function of (A,W). By the Cauchy-Schwarz inequality (analogous to (13)), it follows that $l''(\epsilon)\leq 0$ for all ϵ. Because the log likelihood is concave, the maximum likelihood estimator of ϵ is any solution ϵ_1 of $l'(\epsilon_1)=0$. Hence, by (18), if an mle ϵ_n exists then it solves,

$$P_n\frac{\phi'_{p_n^0}}{\phi_{p_n^0}}(\epsilon_n)=P_nD_0(p_n^0). \quad (18)$$

Analogous to (13), it follows that the conditional mean of $D_0(p_n^0)(O)$, given A,W, under $p_n^0(\epsilon_n)$ is given by

$$E_{p_n^0(\epsilon_n)}(D_0(p_n^0)(O)\mid A,W)=\frac{\phi'_{p_n^0}}{\phi_{p_n^0}}(\epsilon_n\mid A,W),$$

which shows that ϵ_n satisfies the equality

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-Q(p_n^0(\epsilon_n))(A_i,W_i))\frac{I(A_i=1)-I(A_i=0)}{g(p_n^0)(A_i\mid W_i)}=0. \quad (19)$$

We also note that

$$\begin{aligned}\Psi(p_n^0(\epsilon_n)) &= E_{p_n^0(\epsilon_n)}\{Q(p_n^0(\epsilon_n))(1,W)-Q(p_n^0(\epsilon_n)(0,W)\}\\ &= P_n\{Q(p_n^0(\epsilon_n))(1,W)-Q(p_n^0(\epsilon_n))(0,W)\}, \quad (20)\end{aligned}$$

because the marginal distribution of W under $p_n^0(\epsilon_n)$ equals the marginal distribution of W under p_n^0 and the latter equals the empirical distribution

of W. Thus,

$$\begin{aligned} P_nD(p_n^0(\epsilon_n)) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - Q(p_n^0(\epsilon_n)(A_i,W_i))\frac{I(A_i=1)-I(A_i=0)}{g(p_n^0)(A_i \mid W_i)} \\ &\quad +Q(p_n^0(\epsilon_n)(1,W_i) - Q(p_n^0(\epsilon_n)(0,W_i) - \Psi(p_n^0(\epsilon_n) \\ &= 0 + \sum_{i=1}^{n} Q(p_n^0(\epsilon_n))(1,W_i) - Q(p_n^0(\epsilon_n))(0,W_i) - \Psi(p_n^0(\epsilon_n)) \text{ by (19)} \\ &= 0 \text{ by (20)}. \end{aligned}$$

This proves that the targeted maximum likelihood estimator is achieved in the first step of the algorithm and solves the efficient influence curve estimating equation $P_nD(p) = 0$. Note that we could also have solved directly $P_nD(p_n^0(\epsilon)) = 0$ and we would have found the same solution, but it would not have been an easier approach.

This result can be generalized to many other linear pathwise differentiable parameters in nonparametric models based on censored data structure under the coarsening at random assumption, where $D(p)$ represents the double robust estimating function/efficient influence curve, as presented in (van der Laan and Robins (2002)) in closed form for numerous censored data structures.

6.3 Targeted maximum likelihood estimation of a marginal causal effect in nonparametric model: submodel II

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 $p_n^0(\epsilon)$

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

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 $\epsilon_1 \to P_n \log p^0(\epsilon_1)(W)$ 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.

6.4 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

$$g(p_n^0)(\epsilon_2)(1 \mid W) = \frac{1}{1+\exp(-m_n^0(W) - \epsilon_2 h(p_n^0)(W))}. \tag{21}$$

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

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

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 ψ_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) of the corresponding estimating equations.

6.5 Simulation for targeted MLE of 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:

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

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 by:

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

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 $\sigma 2$ 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 (2002)).

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.

6.6 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 $\beta \to m(A, W \mid \beta)$ 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. 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:

$$T_{nuis}^{\perp}(p) = \{D_h(p) : h\} \subset L_0^2(P),$$

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:

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

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 (29). 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 and 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.

Hardest parametric submodel.

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 $h(p_n^0)$ satisfying $E_{p_n^0}(h(p_n^0)(A, W) \mid W) = 0$. 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,

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

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 $f_0'(x)/f_0(x) = 2x/\sigma^2$)

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

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

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

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

$$\begin{aligned} h(p_n^0)(A, W) \equiv\ & \frac{1}{\sigma^2(A, W)} \frac{d}{d\beta(Q_n^0)} m(A, W \mid \beta(Q_n^0)) \\ & - \frac{1}{\sigma^2(A, W)} \frac{E_{p_n^0} \left(\frac{d}{d\beta(Q_n^0)} m(A, W \mid \beta(Q_n^0))/\sigma^2(A, W) \mid W \right)}{E_{p_n^0}(1/\sigma^2(A, W) \mid W)}. \end{aligned} \tag{22}$$

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:

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

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

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} \frac{1}{\sigma^2(A_i, W_i)} \left(Y_i - m(A_i, W_i \mid \beta_n^0 + \epsilon) - \theta_n^0(W_i) - \epsilon r(p_n^0)(W_i)\right)^2 \quad (23)$$

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

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

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:

$$0 = P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n^\top r(p_n^0)\right)}{\sigma^2}. \quad (24)$$

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

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

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

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

but the latter equals zero by the fact that $P_n S(\epsilon_n) = 0$ (31). 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 $P_n D(p) = 0$. For nonlinear models $m(A, W \mid \beta)$ the targeted MLE algorithm will also need to be iterated till convergence.

6.7 Targeting the treatment mechanism as well.

We will now proceed with this example, but also use for $g_0(A \mid W)$ a targeted maximum likelihood estimator. We note that

$$D(p) = (h(p)(A \mid W) - E_{g(p)}(h(p) \mid W))(Y - m(A, W \mid \beta(p)) - D_{CAR}(p),$$

where

$$D_{CAR}(p) = (h(p)(A \mid W) - E_{g(p)}(h(p) \mid W))\theta(p)(W)$$

with $\theta(p) = E_p(Y \mid A = 0, W)$. $D_{CAR}(p)$ is a valid score for $g(p)$ since it is a function of A, W with conditional mean zero, given W. Let $g(p_n^0)(A \mid W)$ be the initial conditional distribution of A, given W, and consider a submodel $g(p_n^0)(\epsilon_2)$ with $g(p_n^0)(0) = g(p_n^0)$ and score $D_{CAR}(p_n^0)$ at $\epsilon_2 = 0$. If A is binary, we provided a logistic regression model with covariate $r(p_n^0)(W)$ which satisfies these restrictions (21). In general, we can use an exponential submodel $g(p_n^0)(\epsilon) = c(\epsilon_2, p_n^0)\exp(\epsilon_2 D_{CAR}(p_n^0))g(p_n^0)$, where $c(\cdot)$ is the normalizing constant. Let $\epsilon_{2n} = \arg\max P_n \log g(p_n^0)(\epsilon)$, or we can choose a model of conditional normal densities as we used for $Q(p_n^0)(Y \mid A, W)$ above. If A is binary and we use the logistic regression model (21) with additional covariate $r_2(p_n^0)(W)$, then we have

$$\frac{d}{d\epsilon_2} \log g(p_n^0)(\epsilon_2)\Big|_{\epsilon_2=0}(O) = r_2(p_n^0)(W)(A - g(p_n^0)(1 \mid W)).$$

So then one can solve for r_2 by setting

$$r_2(W)(A - g(p_n^0)(1 \mid W)) \;=\; D_{CAR}(p_n^0)(O).$$

This yields the solution

$$r_2(p_n^0)(W) = \frac{h(p_n^0)(1,W)\theta(p_n^0)(W)}{1 - g(p_n^0)(1 \mid W)}.$$

If A is continuous, one could use a normal density model updating the regression $E_{p_n^0}(A \mid W)$ and find the right updating function so that the score of $g(p_n^0)(\epsilon)$ at $\epsilon = 0$ equals $D_{CAR}(p_n^0)$, as we did above. We leave this excercise to the reader. 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 semiparametric regression model, including the targeting of the treatment mechanism. Thus the algorithm for targeted maximum likelihood estimation of $\psi_0 = \beta(p_0)$ can be described as follows. To be specific, we consider the case that A is binary and that we use the logistic regression model with the additional covariate r_2 described above. Let θ_n^0, β_n^0 and $g_n^0(A \mid W)$ be given. Set $k = 0$. Let $Q_n^k(A,W) = m(A,W \mid \beta_n^k) + \theta_n^k(W)$. The density p_n^k is identified by Q_n^k and g_n^k, where we treat $\sigma^2(A,W)$ as given, and the marginal distribution of W is fixed at the empirical probability distribution.

Consider

$$r_1(p_n^k)(W) = \frac{E_{p_n^k}\left(\frac{d/d\beta_n^k m(A,W|\beta_n^k)}{\sigma^2(A,W)} \mid W\right)}{E_{p_n^k}\left(\frac{1}{\sigma^2(A,W)} \mid W\right)}$$

and

$$r_2(p_n^k)(W) = \frac{h(p_n^k)(1,W)\theta(p_n^k)(W)}{1 - g(p_n^k)(1 \mid W)}.$$

Let

$$\epsilon_{1n}^k = \arg\min_{\epsilon 1} \sum_{i=1}^n \frac{1}{\sigma^2(A_i,W_i)} \left(Y_i - m(A_i,W_i \mid \beta_n^k + \epsilon_1) - \theta_n^k(W_i) - \epsilon_1^\top r_1(p_n^k)(W_i)\right)^2 (25)$$

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

$$g_n^k(\epsilon_2)(1 \mid W) = \frac{1}{1 + \exp(-m_n^k(W) - \epsilon_2 r_2(p_n^k)(W))}.$$

Let $\epsilon_{2n}^k = \arg\max_{\epsilon_2} P_n \log g_n^k(\epsilon_2)$ be the maximum likelihood estimator of this univariate logistic regression model. In general, $\epsilon_{2n}^k = \arg\max_{\epsilon_2} P_n \log g_n^k(\epsilon_2)$, where $g_n^k(\epsilon_2)$ is a submodel so that $g_n^k(0) = g_n^k$ and it has score $D_{CAR}(p_n^k)$ at $\epsilon = 0$, where $D_{CAR}(p_n^k) = (h(p_n^k)(A,W) - E_{p_n^k}(h(p_n^k) \mid W))\theta(p_n^k)(W)$, and the dependence on p_n^k is only through g_n^k, Q_n^k. Now, update g_n^k and Q_n^k as follows:

$$\begin{aligned}
\beta_n^{k+1} &= \beta_n^k + \epsilon_{1n}^k \\
\theta_n^{k+1} &= \theta_n^k + \epsilon_{1n}^k r_1(p_n^k) \\
Q_n^{k+1}(A,W) &= m(A,W \mid \beta_n^{k+1}) + \theta_n^{k+1}(W) \\
m_n^{k+1}(W) &= m_n^k(W) + \epsilon_{2n}^k r_2(p_n^k)(W) \\
g_n^{k+1}(A \mid W) &= \frac{1}{1+\exp(-m_n^{k+1}(W))}
\end{aligned}$$

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 $D(g,Q)(O) = D_{PTW}(g,Q) - D_{CAR}(g,Q)$, where $D_{CAR}(g,Q) = (h(g,Q)(A,W) - E_g(h(g,Q) \mid W))\theta(Q)(W)$ and $D_{PTW}(g,Q) = (h(g,Q)(A,W) - E_g(h(g,Q) \mid W))(Y - m(A,W \mid \beta(Q))$, $h(g,Q)$ is a specified function, and PTW stands now for "Probability of Treatment Weighting". For k converging to infinity the targeted MLE yields a final estimator g_n of the treatment mechanism and a regression fit $Q_n(A,W) = m(A,W \mid \beta(Q_n)) + \theta(Q_n)(W)$ 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

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

Thus, we can conclude that the three estimators $\beta_{n,PTW}$ solving $P_n D_{PTW}(\beta, g_n) = 0$, $\beta_{n,DR-PTW}$ solving $P_n D(\beta, g_n, Q_n) = 0$, and the targeted MLE $\beta(Q_n)$ are identical: here $D_{PTW}(\beta, g)$ and $D(\beta, g, Q)$ are the natural representations so that $D_{PTW}(p) = D_{PTW}(\beta(p), g(p))$ and $D(p) = D(\beta(p), g(p), Q(p))$. That is, the targeted MLE $\beta(Q_n)$ equals the PTW and DR-PTW estimator based if one uses the targeted MLE (g_n, Q_n) as estimators of the nuisance parameters.

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:

$$p^{k+1} = arg \max_{p \in \{p^k(\epsilon):\epsilon\}} P_0 \log p. \tag{26}$$

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:

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

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 $\psi_n = \Psi(p_n)$.

8 Targeting the fitting of the nuisance parameters/initial density.

In the above sections we proposed a targeted MLE algorithm which takes an initial density estimator as starting point and maps it into a solution of the efficient influence curve equation while increasing the log-likelihood along the hardest submodel for the parameter of interest. In addition, as in Section 7, we noted that one can use the log likelihood and log-likelihood based cross-validation to select among candidate targeted MLE's indexed by different starting densities. In this loss based approach based on the loss function $L(p \mid P_0) = -\log p_0^*(p)$ a candidate fit of a nuisance parameter is now scored by how well it is fitting the nuisance parameter as well as how it helps to increase the likelihood along the hardest submodel during the targeted MLE algorithm. In this section we propose an approach which scores a candidate nuisance parameter fit only by how much it helps to increase the log-likelihood during the targeted MLE algorithm, since the latter is fully targeted at estimation of the parameter of interest. This provides a new criteria for selecting among candidate nuisance parameter fits and generates templates for powerful data adaptive algorithms for estimating the parameter of interest, as we point out in more detail below.

The resulting proposed algorithms are of the following general form: 1) we start with a simple initial targeted ML density estimator based on a low dimensional model (e.g., a singleton), 2) given the initial targeted MLE, we have a set of candidate nuisance parameter moves in which each move represents a mapping from the targeted ML density estimator into a new density estimator increasing the log-likelihood of the data, 3) we select the nuisance parameter move which results in a *maximal value of the log-likelihood at the targeted MLE starting at the nuisance parameter move minus the log-likelihood at the nuisance parameter move itself*, 4) we iterate this algorithm till convergence, and 5) certain constraints on the set of nuisance parameter moves will be selected with likelihood based cross-validation.

Formally, the algorithm can be described as follows. Given a $p \in \mathcal{M}$, we define $\mathcal{M}_{nuis,s}(p)$ as a collection of parametric nuisance submodels of $\mathcal{M}$ through p indexed by an index s ranging over an index set. Specifically, let $\mathcal{M}_{nuis,s}(p) = \{\{p(\delta \mid h) : \delta\} : h \in \mathcal{H}_s\}$ be a set of paths $\{p(\delta \mid h) : \delta\}$ with parameter δ through p at $\delta = 0$ and score $h \in T_{nuis}(p)$ at $\delta = 0$, where h ranges over a set of scores $\mathcal{H}_s \subset T_{nuis}(p)$ in the nuisance tangent space $T_{nuis}(p)$ at p of $\mathcal{M}$. Thus, each path $\{p(\delta \mid h) : \delta\}$ represents a parametric model in which the nuisance parameter fit can be improved satisfying $\frac{d}{d\delta}\Psi(p(\delta \mid h))\Big|_{\delta=0} = 0$. The proposed algorithm can now be formulated as follows.

Initial targeted MLE: For each s, start with an initial targeted MLE density p_{ns}^0 solving $P_n D(p_{ns}^0) = 0$.

Compute MLE's among all candidate nuisance moves increasing the log-likelihood. Let $p_{nsh} \equiv \arg\max_{p \in \{p_{ns}^0(\delta|h):\delta\}} P_n \log p$, $h \in \mathcal{H}_s$, for all s.

Select best nuisance parameter move with respect to the parameter of interest: For each s, let

$$h_{ns} \equiv \arg\max_{h \in \mathcal{H}_s} P_n \log \frac{p_n^*(p_{nsh})}{p_{nsh}},$$

where $p_n^*(p)$ denotes the targeted MLE starting at p.

Compute next targeted MLE: For each s, let $p_{ns}^1 \equiv p_n^*(p_{nh_{ns}})$ be the targeted MLE starting at $p_{nh_{ns}}$.

Iterate until convergence: Iterate this process mapping a targeted MLE p_{ns}^k into a new targeted MLE p_{ns}^{k+1}, $k = 0, 1, \ldots$, for each s. Note that, for each s, the likelihood is increasing in k. Denote the resulting limit density estimator with p_{ns}, for each s.

Likelihood based cross-validation to select s: Select among the candidate density estimators p_{ns} with likelihood based cross-validation, resulting in the targeted MLE $p_n \equiv p_{ns_n}$, where s_n is the index minimizing the cross-validated log-likelihood of the candidate estimators p_{ns}.

This algorithm provides for each s a sequence of solutions p_{ns}^k of the efficient influence curve estimating equation $P_n D(p_{ns}^k) = 0$, $k = 1, \ldots$, where the log-likelihood of the data increases in k. At step k of the algorithm it searches among candidate nuisance parameter fluctuations through p_{ns}^k

(i.e., MLE's of $p_n^k(\delta \mid h)$ in δ along directions h), each of them increasing the overall log-likelihood relative to the current density estimator p_{ns}^k, and it selects the one which results in a maximal *increase* of the log-likelihood along the hardest submodel through the improved fit with score being the efficient influence curve with respect to the parameter of interest.

The main deviation of this algorithm relative to the loss based estimation approach in Section 7 is that the criteria used to select among candidate improved nuisance parameter fits (say) p_h is given by $h \to P_n \log \frac{p_n^*(p_h)}{p_h}$ instead of $h \to P_n \log p_n^*(p_h)$. That is, this new criteria rewards improved nuisance parameter fits which yield a steep learning curve for the targeted MLE algorithm, but, on the other hand, an improved fit of a nuisance parameter which has no effect on the targeted MLE algorithm results in no award at all. In particular, this implies that only improved fits for nuisance parameters the efficient influence curve depends on can be selected, and among them, one selects the one which provides the steepest learning curve for the targeted MLE algorithm. As a consequence, this new criteria will select nuisance parameter fits directly affecting the canonical gradient (and specifically, moving the empirical mean of the canonical gradient away from 0), which itself is already a great property.

As an example, consider estimation of a parameter $\Psi(F_X)$ of the full data distribution F_X based on a censored data structure $O = \Phi(C, X)$ of $X \sim F_X$ under the assumption of CAR on the conditional distribution of C, given X. Represent a candidate density p of O as $p = g * Q$, where g represents the censoring mechanism and Q represents the factor indexed by the distribution of the full data. The efficient influence curve of the full data parameter of interest depends, in particular, on the censoring mechanism, while the parameter of interest is not a function of the censoring mechanism. Consider a hardest submodel only updating Q and let $Q^*(g, Q)$ represent the targeted MLE of Q starting at (g, Q). The log-likelihood $\log p$ consists of a sum of a log-likelihood of the censoring mechanism $\log g$ and a log likelihood indexed by the distribution of the full data $\log Q$. Consider two candidate densities $p_1 = g_1 Q$ and $p_2 = g_2 Q$. If one uses the targeted log-likelihood as criteria to select among p_1 and p_2 (and thus among the two censoring fits), we would calculate $P_n \log p^*(p_1) - \log p^*(p_2) = P_n \log g_1/g_2 + P_n \log Q^*(g_1, Q) - P_n \log Q^*(g_2, Q)$, and if it is larger than 0, then we would select p_1. As a consequence, $p_1 = g_1 Q$ affects the targeted log-likelihood in a direct manner through the censoring part of the log like-

lihood $P_n \log g_1$, and indirectly through an improved fit of the score/efficient influence curve of the hardest submodel the targeted MLE algorithm is based upon. However, an improved fit of the censoring mechanism itself is not necessarily of interest since the parameter of interest is only a function of the full data factor Q of the density of the data. On the other hand, in the above algorithm we score p_1 with $P_n \log p^*(p_1)/p_1 = P_n \log Q^*(g_1, Q)/Q$ and we score p_2 with $P_n \log p^*(p_2)/p_2 = P_n \log Q^*(g_2, Q)/Q$, so that we would select the p_j with $P_n \log Q^*(g_j, Q) = \max_l P_n \log Q^*(g_l, Q)$. In this manner one selects nuisance parameter fits g_j which give maximal log likelihood $P_n \log Q^*(g_j, Q)$, which makes sense since the corresponding targeted MLE will be defined as $\Psi(Q^*(g_j, Q))$.

As another example, consider estimation of $E(Y \mid A, W) - E(Y \mid A = 0, W) = m(A, W \mid \beta)$ according to a model $m(\cdot \mid \beta)$ based on i.i.d. sampling (W, A, Y). The efficient influence curve is unknown up to the parameter of interest $\beta(p)$, and the variation independent nuisance parameters $\theta(p) = E_p(Y \mid A = 0, W)$ and $g(p)(A \mid W)$. In Section 6 we presented the targeted MLE $p_n^*(p)$ of β based on a normal regression model taking as starting density p. Nuisance parameter moves represent now proposed MLE updates of a current fit of $\theta(p^k)$ and $g(p^k)$. Such moves could, for example, be represented by adding or substituting variables to fit these nuisance parameters. These moves could be restricted by a constraint s constraining the size of the models for θ and g, the choice of variables, the Euclidan norm of the coefficient vector, among others. The proposed algorithm would now score these MLE updates $p^k(\delta)$ of the nuisance parameters $\theta(p^k)$ and $g(p^k)$ by evaluating the increase of the targeted log-likelihood $P_n \log p_n^*(p^k(\delta))/p^k(\delta)$, and selecting the update $p^k(\delta)$ which maximizes this increase of the log-likelihood during the targeted MLE algorithm. As a consequence, our algorithms above would aim to select variables in the treatment mechanism and $E(Y \mid A = 0, W)$ which yield a maximal change in the parameter of interest, as measured by the increase of the targeted log-likelihood, while always mapping into a new improved solution of the efficient influence curve equation. Again, this means that the nuisance parameter fits are selected to be increase the likelihood *and* be very relevant and important for fitting the parameter of interest.

An interesting variation of this class of algorithms is obtained by replacing the targeted MLE step by an alternative maximization w.r.t. to the parameter of interest value only, assuming an appropriate way of parameterizing the density estimator in terms of nuisance terms and a term directly affecting the parameter of interest. We would still recommend to run the targeted

MLE algorithm at the very end to guarantee that one ends up with a density estimator solving the efficient influence curve estimating equation.

9 The bias correction the targeted maximum likelihood algorithm achieves at each step.

The purpose of this section is to provide an understanding of the bias correction the targeted MLE algorithm provides relative to the bias of the initial density estimator p_n^0.

9.1 The first step bias correction in the targeted MLE algorithm.

Consider a collection of candidate density estimators $p_{nh} \equiv \Phi_h(P_n)$ of p_0 indexed by an index h, where h corresponds with the bias of p_{nh} with respect to p_0: e.g. h is the bandwidth of a kernel density estimator p_{nh} or h indexes a model choice and p_{nh} is the corresponding maximum likelihood estimator. Let $p_h \equiv \Phi_h(P_0)$ represent the asymptotic target of these candidate density estimators for fixed index h. Let $p_{nh}^1 = p_{nh}(\epsilon_n) \approx (1 + \epsilon_n D(p_{nh}))p_{nh}$ be the first step targeted maximum likelihood density estimator, and let $p_h^1 \equiv p_h(\epsilon_0) \approx (1 + \epsilon_0 D(p_h))p_h$ denote its asymptotic target with ϵ_0 being the asymptotic limit of ϵ_n for fixed h as defined below. In this subsection we are concerned with comparing, for each fixed h, the asymptotic bias $\Psi(p_h) - \Psi(p_0)$ of the substitution estimator $\Psi(p_{nh})$ relative to the asymptotic bias $\Psi(p_h^1) - \Psi(p_0)$ of the first step targeted MLE $\Psi(p_{nh}^1)$. The theorem below shows that $\{\Psi(p_h^1) - \Psi(p_0)\}/\{\Psi(p_h) - \Psi(p_0)\}$ will generally converge to zero at a certain rate if h converges to zero: in the special case that the model is convex and the parameter Ψ is linear the result is particular strong. In this theorem we suppress the dependence on h and thus denote p_h with p and $p_h^1 = p(\epsilon_0)$, but the remainders are studied in the context of p approximating p_0. For simpicity, we present the result for univariate parameters $\Psi(p) : \mathcal{M} \to \mathbb{R}$, but it can be easily generalized to Euclidean valued parameters.

In order to establish a result we first prove the following expression for the relative asymptotic bias of the first step targeted MLE relative to the asymptotic bias of the initial estimator.

Lemma 1 *Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a pathwise differentiable parameter at any $p \in \mathcal{M}$, and let $D(p)$ be the canonical gradient of this pathwise derivative at p. Let p_0, p be given. Let $p(\epsilon) = (1 + \epsilon D(p))p + o(\epsilon) \in \mathcal{M}$ be a smooth submodel through p with score $D(p)$ and information $E_p D(p)^2$ at $\epsilon = 0$. Let*

$$\epsilon_0 = \epsilon(p_0 \mid p) \equiv \arg\max_{\epsilon} E_{P_0} \log p(\epsilon).$$

We assume that ϵ_0 solves $P_0 \frac{\frac{d}{d\epsilon}p(\epsilon)}{p(\epsilon)} = 0$.

We note

$$\frac{d}{d\epsilon}\Psi(p(\epsilon))\bigg|_{\epsilon=0} = PD(p)^2.$$

Define the following remainders:

$$\begin{aligned} R_1(\epsilon_0, p) &\equiv \Psi(p(\epsilon_0)) - \Psi(p) - PD(p)^2\epsilon_0 \\ R_2(p, p_0) &\equiv \Psi(p_0) - \Psi(p) - P_0 D(p) \\ R_3(p, p_0) &\equiv \epsilon(p_0 \mid p) - \left\{P_0 D(p_0)^2\right\}^{-1} P_0 D(p) \end{aligned}$$

We have

$$\begin{aligned} \frac{\Psi(p(\epsilon_0)) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} &= R(p(\epsilon_0), p, p_0) \\ &\equiv \frac{PD(p)^2 - P_0 D(p_0)^2}{PD(p)^2} - \frac{R_2(p, p_0) + PD(p)^2 R_3(p, p_0) + R_1(\epsilon_0, p)}{\Psi(p) - \Psi(p_0)}. \end{aligned}$$

Proof: We have

$$\begin{aligned} \frac{\Psi(p(\epsilon_0)) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} &= \frac{\Psi(p(\epsilon_0)) - \Psi(p)}{\Psi(p) - \Psi(p_0)} + 1 \\ &= \frac{\epsilon_0 PD(p)^2 + R_1(\epsilon_0, p)}{\Psi(p) - \Psi(p_0)} + 1 \\ &= \frac{\epsilon_0 PD(p)^2}{\Psi(p) - \Psi(p_0)} + 1 + \frac{R_1(\epsilon_0, p)}{\Psi(p) - \Psi(p_0)} \\ &= \frac{\frac{P_0 D(p)}{P_0 D(p_0)^2} PD(p)^2}{\Psi(p) - \Psi(p_0)} + 1 + \frac{R_3(p, p_0) PD(p)^2}{\Psi(p) - \Psi(p_0)} + \frac{R_1(\epsilon_0, p)}{\Psi(p) - \Psi(p_0)} \end{aligned}$$

Let $c_0 \equiv PD(p)^2/P_0D(p_0)^2$. Note

$$\begin{aligned} c_0\frac{P_0D(p)}{\Psi(p)-\Psi(p_0)} &= c_0\frac{\Psi(p_0)-\Psi(p)-R_2(p,p_0)}{\Psi(p)-\Psi(p_0)} \\ &= -c_0 - \frac{R_2(p,p_0)}{\Psi(p)-\Psi(p_0)} \\ &= -1+(1-c_0) - \frac{R_2(p,p_0)}{\Psi(p)-\Psi(p_0)}. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\Psi(p(\epsilon_0))-\Psi(p_0)}{\Psi(p)-\Psi(p_0)} &= (1-c_0) - \frac{R_2(p,p_0)}{\Psi(p)-\Psi(p_0)} + \frac{R_3(p,p_0)PD(p)^2}{\Psi(p)-\Psi(p_0)} + \frac{R_1(\epsilon_0,p)}{\Psi(p)-\Psi(p_0)} \\ &= \frac{PD(p)^2-P_0D(p_0)^2}{PD(p)^2} - \frac{R_2(p,p_0)+PD(p)^2R_3(p,p_0)+R_1(\epsilon_0,p)}{\Psi(p)-\Psi(p_0)} \\ &= R(p(\epsilon_0),p,p_0).\square \end{aligned}$$

This theorem can now be applied to $p = p_h$, and $p_h^1 = p_h(\epsilon_0)$ with $h = h(n)$ be a sequence converging to zero representing the asymptotic bias of the initial density estimators p_{nh}. One would now need to establish that $R(p_{h(n)}(\epsilon_0), p_{h(n)}, p_0) = O(r(n))$ for some sequence $r(n)$ converging to zero. Such a result would then establish that the first step targeted MLE has a smaller asymptotic bias than the original estimator by an order of magnitude. We will now formalize this approach by carefully investigating the remainders R_1, R_2 and R_3.

Understanding the relative bias correction term: The first and second remainder terms R_1 and R_2 will typically involve second order differences in $p - p_0$ because of pathwise differentiability of Ψ at p and p_0, respectively, and are therefore easy to understand/express in terms of $p - p_0$ and ϵ_0. We will now carefully derive an expression for R_3 which will allow us to establish the theorem. Define

$$U(\epsilon,p) = P_0\frac{\frac{d}{d\epsilon}p(\epsilon)}{p(\epsilon)}$$

as the expectation under P_0 of the score at ϵ of $p(\epsilon)$. For example, if $p(\epsilon) = (1+\epsilon D(p))p$, then

$$U(\epsilon,p) = P_0\frac{D(p)}{1+\epsilon D(p)}.$$

It is assumed that $\frac{d}{d\epsilon}U(\epsilon, p_0)$ at $\epsilon = 0$ equals $-P_0D(p_0)^2$. We have $U(0, p_0) = 0$ and $U(\epsilon_0, p)$=0. Thus we can start with the identity

$$U(\epsilon_0, p_0) - U(0, p_0) = -\left\{U(\epsilon_0, p) - U(\epsilon_0, p_0)\right\}.$$

By continuous differentiability of $\epsilon \to U(\epsilon, p_0)$ at $\epsilon = 0$ with invertible derivative, it follows that

$$\epsilon_0 = -\left.\frac{d}{d\epsilon}U(\epsilon, p_0)\right|^{-1}_{\epsilon=0}\left\{U(\epsilon_0, p) - U(\epsilon_0, p_0)\right\} + R_{31}(p, p_0, \epsilon_0),$$

where $R_{31} = o(\mid \epsilon_0 \mid)$. By assumption, we have $\left.\frac{d}{d\epsilon}U(\epsilon, p_0)\right|_{\epsilon=0} = -P_0D(p_0)^2$. So

$$\begin{aligned}
\epsilon_0 &= \frac{1}{P_0D(p_0)^2}\left\{U(\epsilon_0, p) - U(\epsilon_0, p_0)\right\} + R_{31}(p, p_0, \epsilon_0) \\
&= \frac{1}{P_0D(p_0)^2}\left\{U(0, p) - U(0, p_0)\right\} + R_{32}(p, p_0, \epsilon_0) + R_{31}(p, p_0, \epsilon_0) \\
&= \frac{1}{P_0D(p_0)^2}P_0D(p) + R_{32}(p, p_0, \epsilon_0) + R_{31}(p, p_0, \epsilon_0),
\end{aligned}$$

where

$$R_{32} \equiv \frac{1}{P_0D(p_0)^2}\left\{U(\epsilon_0, p) - U(\epsilon_0, p_0) - U(0, p) + U(0, p_0)\right\}.$$

Now, note that $R_3 = R_{31}/P_0D(p_0)^2 + R_{32}$ and that typically R_{31}, R_{32} involve terms with ϵ_0 squared and ϵ_0 times $p - p_0$.

Suppose now that $p = p(n)$ is a sequence approximating p_0 for $n \to \infty$. Assume that for this sequence we have $R_{31}(p, p_0, \epsilon_0) = o(\mid \epsilon_0 \mid)$ and $R_{32}(p, p_0, \epsilon_0) = o(\mid \epsilon_0 \mid)$. Then it follows that $\epsilon_0 = O(\mid P_0D(p) \mid)$. Define $d(p, p_0) \equiv \mid P_0D(p) \mid$. Then it follows that $\epsilon_0 = O(d(p, p_0))$.

Now, assume that for the sequence p approximating p_0, and using that $\epsilon_0 = O(d(p, p_0))$, there exist a polynomial function $r_3(x) = x^{a_3}$ for some $a_3 > 1$ so that $R_{31}(p, p_0, \epsilon_0) = O(r_3(d(p, p_0)))$ and $R_{32}(p, p_0, \epsilon_0) = O(r_3(d(p, p_0)))$. As a consequence, this shows that $R_3(p, p_0) = O(r_3(d(p, p_0)))$. Assume that, if $\epsilon_0 = O(d(p, p_0))$, then there exists a polynomial function $r_1(x) = x^{a_1}$ for some $a_1 > 1$ so that $R_1(\epsilon_0, p) = O(r_1(d(p, p_0)))$. One might even expect that $R_1(\epsilon, p) = O(\epsilon^2)$ so that this might hold for $a_1 = 2$ in many applications. Assume that there exists a polynomial function $r_2(x) = x^{a_2}$ for some $a_2 > 1$

so that $R_2(p, p_0) = O(r_2(d(p, p_0)))$. Again, this is reasonable condition since $R_2(p, p_0)$ is the second order term in $\Psi(p_0) - \Psi(p) = P_0 D(p) + R_2(p, p_0)$, which should thus be of smaller order than the first order approximation $P_0 D(p)$.

Let $a \equiv \min(a_1, a_2, a_3)$. Assume that $P_0 D(p)/\{\Psi(p_0) - \Psi(p)\} = O(1)$. Then an application of the Lemma teaches us that

$$\frac{\Psi(p(\epsilon_0)) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} \quad = \quad O(\mid PD(p)^2 - P_0 D(p_0)^2 \mid) + O(d(p, p_0)^{a-1}).$$

This proves the following Theorem.

Theorem 2 *Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a pathwise differentiable parameter at any $p \in \mathcal{M}$, and let $D(p)$ be the canonical gradient of this pathwise derivative at p. Let p_0 be given and let $p = p(n)$ be sequence of densities in $\mathcal{M}$ (possibly approximating p_0). We will suppress the dependence on n of all quantities. Let $p(\epsilon) = (1 + \epsilon D(p))p + o(\epsilon) \in \mathcal{M}$ be a smooth submodel through p with score $D(p)$ and information $E_p D(p)^2$ at $\epsilon = 0$. Let*

$$\epsilon_0 = \epsilon(p_0 \mid p) \equiv \arg\max_{\epsilon} E_{P_0} \log p(\epsilon).$$

We assume that ϵ_0 solves $P_0 \frac{\frac{d}{d\epsilon} p(\epsilon)}{p(\epsilon)} = 0$. We refer to the definitions of R_1, R_2, R_3 in Lemma 1.

Define

$$U(\epsilon, p) = P_0 \frac{\frac{d}{d\epsilon} p(\epsilon)}{p(\epsilon)}.$$

It is assumed that $\frac{d}{d\epsilon} U(\epsilon, p_0)$ at $\epsilon = 0$ equals $-P_0 D(p_0)^2$, $U(0, p_0) = 0$ and $U(\epsilon_0, p)$=0. Define

$$\begin{aligned} R_{31}(p, p_0, \epsilon_0) &\equiv U(\epsilon, p_0) - U(0, p_0) - \frac{d}{d\epsilon} U(\epsilon, p_0)\Big|_{\epsilon=0} \epsilon_0 \\ R_{32}(p, p_0, \epsilon_0) &\equiv \frac{1}{P_0 D(p_0)^2} \{U(\epsilon_0, p) - U(\epsilon_0, p_0) - U(0, p) + U(0, p_0)\}. \end{aligned}$$

(In relation to Lemma 1, we have $R_3 = R_{31}/P_0 D(p_0)^2 + R_{32}$.) Suppose now that for the sequence $p = p(n)$ approximating p_0 we have $R_{31}(p, p_0, \epsilon_0) = o(\mid \epsilon_0 \mid)$ and $R_{32}(p, p_0, \epsilon_0) = o(\mid \epsilon_0 \mid)$. Then $\epsilon_0 = O(\mid P_0 D(p) \mid)$.

Let $d(p, p_0) \equiv \mid P_0 D(p) \mid$ so that $\epsilon_0 = O(d(p, p_0))$.

Assume that, if $\epsilon_0 = O(d(p, p_0))$, *then there exist a polynomial function* $r_3(x) = x^{a_3}$ *for some* $a_3 > 1$ *so that* $R_{31}(p, p_0, \epsilon_0) = O(r_3(d(p, p_0)))$ *and* $R_{32}(p, p_0, \epsilon_0) = O(r_3(d(p, p_0))$. *Then, in relation to the Lemma 1,* $R_3(p, p_0, \epsilon_0) = O(r_3(d(p, p_0)))$.

Assume that, if $\epsilon_0 = O(d(p, p_0))$, *then there exists a polynomial function* $r_1(x) = x^{a_1}$ *for some* $a_1 > 1$ *so that* $R_1(\epsilon_0, p) = O(r_1(d(p, p_0)))$. *Assume that there exists a polynomial function* $r_2(x) = x^{a_2}$ *for some* $a_2 > 1$ *so that* $R_2(p, p_0) = O(r_2(d(p, p_0)))$.

Let $a \equiv \min(a_1, a_2, a_3)$. *Assume that* $P_0D(p)/\{\Psi(p_0) - \Psi(p)\} = O(1)$. *Then*

$$\frac{\Psi(p(\epsilon_0)) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} \quad = \quad O(\mid PD(p)^2 - P_0D(p_0)^2 \mid) + O(d(p, p_0)^{a-1}).$$

9.2 The asymptotic bias of the k-th Step targeted MLE.

Iterative application of this theorem provides us also with a result for the bias of the k-th step targeted MLE.

Result 2 *We have*

$$\frac{\Psi(p^2) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} = \frac{\Psi(p^2) - \Psi(p_0)}{\Psi(p^1) - \Psi(p_0)} \frac{\Psi(p^1) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} \equiv R(p^2, p^1, p_0)R(p^1, p, p_0).$$

In general, we obtain:

$$\frac{\Psi(p^k) - \Psi(p_0)}{\Psi(p) - \Psi(p_0)} = \prod_{j=1}^{k} R(p^j, p^{j-1}, p_0),$$

where $p^0 \equiv p$.

It should be observed that, for a sequence $p = p(n)$ approximating p_0, $R(p^j, p^{j-1}, p_0)$ for $j \geq 2$ might not converge to zero for $n \to \infty$ anymore because for j large enough, the denominator $\Psi(p^j) - \Psi(p_0)$ in $R(p^j, p^{j-1}, p_0)$ might be of smaller order than the second order remainders R_1, R_2, R_3 in the numerator of $R(p^j, p^{j-1}, p_0)$. Because of this observation we expect that in many examples (with nonlinear parameter and/or nonconvex model) the rate of convergence of the second order terms R_l, R_2, R_3 at the initial p will typically dominate the rate of convergence of $\Psi(p^j) - \Psi(p_0)$ for $j = 1, 2, \ldots$. That is, the main bias reduction seems to occur at the first step of the targeted MLE algorithm.

9.3 Bias correction term for linear parameter in convex model.

We will apply Lemma 1. Let p_h denote the limit of a density estimator based on a sample of random variables $O_1, \ldots, O_n$ from $p_0 \in \mathcal{M}$ with $\mathcal{M}$ a convex model. Let $\Psi(p)$ be linear in p, and let $p_h(\epsilon) = (1 + \epsilon D(p))p_h \in \mathcal{M}$ for a set of ϵ values including 0. Thus for each such ϵ we have

$$\frac{\Psi(p_h(\epsilon)) - \Psi(p_h)}{\epsilon} = \int D(p_h)^2 p_h(o) d\mu(o),$$

which shows that $R_1(\epsilon, p) = 0$ for all p and allowed ϵ values. In addition, by the identity for convex models and linear parameters (van der Laan (1998, 1995a)) we also have

$$P_0 D(p_h) = \int D(p_h) \frac{p_0 - p_h}{p_h} p_h d\mu = \Psi(p_0) - \Psi(p_h),$$

under the assumption that $p_0/p_h < \infty$, which proves that $R_2(p, p_0) = 0$ for all p, p_0 with $p_0/p < \infty$.

It remains to establish an explicit expression or bound on $R_3(p_h, p_0)$. We have $U(\epsilon, p) = P_0 D(p)/(1 + \epsilon D(p))$ with $U(\epsilon_0, p) = 0 = U(0, p_0)$, and it is assumed that ϵ_0 is such that $p(\epsilon_0) \in \mathcal{M}$. Thus

$$U(\epsilon_0, p) - U(0, p) = -\{U(0, p) - U(0, p_0)\} = -P_0 D(p).$$

By the continuous differentiability of $\epsilon \to U(\epsilon, p)$ at 0, it follows that

$$U(\epsilon_0, p) - U(0, p) = -\epsilon_0 P_0 \frac{D(p)^2}{1 + \epsilon_1 D(p)}$$

for some $\epsilon_1 \in (0, \epsilon_0)$. Thus

$$\epsilon_0 = \left\{ P_0 \frac{D(p)^2}{1 + \epsilon_1 D(p)} \right\}^{-1} P_0 D(p).$$

Thus,

$$\epsilon_0 = O(\mid P_0 D(p) \mid) = O(\mid \Psi(p) - \Psi(p_0) \mid).$$

Define now $g(\epsilon_1) \equiv \left\{ P_0 \frac{D(p)^2}{1+\epsilon_1 D(p)} \right\}^{-1}$ and note that $g(0) = 1/P_0 D(p_0)^2$. Thus

$$\epsilon_0 = g(\epsilon_1) P_0 D(p) = \frac{P_0 D(p)}{P_0 D(p_0)^2} + (g(\epsilon_1) - g(0)) P_0 D(p).$$

Thus,

$$\begin{aligned} R_3(p, p_0) &= (g(\epsilon_1) - g(0))P_0D(p) \\ &= O(\mid \epsilon_0 \mid)P_0D(p) \\ &= O(\mid P_0D(p) \mid^2) \\ &= O(\mid \Psi(p) - \Psi(p_0) \mid^2). \end{aligned}$$

Thus, we have now proved that

$$\frac{\Psi(p_h(\epsilon_0)) - \Psi(p_0)}{\Psi(p_h) - \Psi(p_0)} = \frac{P_hD(p_h)^2 - P_0D(p_0)^2}{P_hD(p_h)^2} - P_hD(p_h)^2O(\mid \Psi(p_h) - \Psi(p_0) \mid).$$

Finally, we note that

$$PD(p)^2 - P_0D(p_0)^2 = \Psi(p)(1-\Psi(p)) - \Psi(p_0)(1-\Psi(p_0)) = O(\mid \Psi(p) - \Psi(p_0) \mid).$$

Application of Lemma 1 now yields the following result.

Result 3 *Let p_h denote the limit of a density estimator based on a sample of random variables $O_1, \ldots, O_n$ from p_0 satisfying that p_0/p_h is uniformly bounded in O on a support of p_0 for each h. Let $\mathcal{M}$ be convex and let $\Psi : \mathcal{M} \to \mathbb{R}$ be linear. Let $D(p)$ be a gradient of Ψ at p. Let $p_h(\epsilon) = (1 + \epsilon D(p))p_h \in \mathcal{M}$ for ϵ in an interval I including 0. Let ϵ_0 be defined as the solution of $P_0D(p_h)/(1 + \epsilon D(p_h)) = 0$ and be such that $p_h(\epsilon_0) \in \mathcal{M}$ (i.e., $\epsilon_0 \in I$). We have*

$$\frac{\Psi(p_h(\epsilon_0)) - \Psi(p_0)}{\Psi(p_h) - \Psi(p_0)} = O(\mid \Psi(p_h) - \Psi(p_0) \mid).$$

This result can be generalized to other parameterizations $p(\epsilon) = (1 + \epsilon D(p))p + o(\epsilon)$. It shows that in convex models and for linear parameters Ψ the targeted MLE algorithm applied to an arbitrary initial p and P_n replaced by the true P_0 always converges to ψ_0. In other words, in the case of a convex model and linear parameter, the targeted MLE algorithm can be viewed as a very fast algorithm mapping an arbitrarily biased initial p (in particular, $\Psi(p)$ far away from ψ_0) into a density p^* with $\Psi(p^*) = \Psi(p_0)$. In fact, by carrying out the infinite sample versions (replace P_n by P_0) of our examples in which we showed that the one-step targeted MLE solved the efficient influence curve equation $P_nS(p) = 0$, it follows that for many choices of hardest submodels $\{p(\epsilon) : \epsilon\}$ the algorithm converges in the first step.

9.4 Targeted MLE algorithm converges to solution of efficient influence curve equation.

Analogous to result 1 one can show that if the targeted MLE algorithm with P_n replaced by P_0 converges, then it will converge to a solution of the equation $P_0D(p) = 0$.

Result 4 *Let P_0 be given. Assume that*

$$\lim_{\epsilon \to 0} \limsup_{k \to \infty} \mid P_0 \frac{\frac{d}{d\epsilon} p^k(\epsilon)}{p^k(\epsilon)} - P_0 \frac{p^{k\prime}(0)}{p^k(0)} \mid \to 0, \tag{27}$$

that for each k there exist a matrix A_k so that $A_k \frac{p^{k\prime}(0)}{p^k(0)} = D(p^k)$ with $\limsup_{k \to \infty} \parallel A_k \parallel < \infty$, where $\parallel \cdot \parallel$ denotes a matrix norm. If $\epsilon(P_0 \mid p^k)$ solves $P_0 \frac{\frac{d}{d\epsilon} p^k(\epsilon)}{p^k(\epsilon)} = 0$ for all k, and $\epsilon(P_0 \mid p^k) \to 0$ for $k \to \infty$ then we have

$$P_0 D(p^k) \to 0 \text{ for } k \to \infty.$$

In convex models and linear parameters we have that, if $p_0/p^k < \infty$, then $P_0D(p^k) = \Psi(p_0) - \Psi(p^k)$, so that in this case the convergence of the algorithm $P_0D(p^k) \to 0$ implies $\Psi(p^k) \to \Psi(p_0)$.

For example, this convergence of the targeted MLE algorithm at P_0 to a solution $P_0D(p) = 0$, for an arbitrary initial starting value $p \in \mathcal{M}$, for linear parameters in convex models, can now be applied to the targeted MLE algorithm of a full data parameter of a distribution of the full data structure X based on censored data $O = \Phi(C, X)$ in the case that the conditional distribution of C, given X, is known.

9.5 Example: Marginal causal effect in nonparametric model.

We revisit this example to illustrate these convergence results of the targeted MLE algorithm at P_0 for arbitrary starting density p^0. In the case that $g_0(A \mid W)$ is known, then the efficient influence curve at $p \in \mathcal{M}$ is still given by $D(p) = \frac{(Y - Q(p)(A,W))}{g_0(A|W)} \{I(A = 1) - I(A = 0)\} - (Q(p)(1, W) - Q(p)(0, W))$. We have that $D(p) = D_{IPTW}(p) - D_{CAR}(p)$, where $D_{IPTW}(p) = Y/g_0(A \mid W)(I(A = 1) - I(A = 0))$. Indeed, we have $P_0D(p) = \Psi(Q(p_0)) - \Psi(Q(p))$ so that solving the equation $P_0D(p) = 0$ fully identifies ψ_0. That is, the

targeted MLE algorithm at P_0 will for an arbitrary starting density p^0 still converges to ψ_0 in the sense that $\Psi(p^k) - \Psi(p_0) \to 0$ for $k \to \infty$.

Consider now the model in which g_0 is unknown. Suppose that we start the algorithm with a p^0 for which $g(p^0)$ is misspecified: i.e., $g(p^0) \neq g_0$. In addition, assume that our hardest submodels $p(\epsilon)$ have score $D(p)$ at $\epsilon = 0$ and only fluctuate the conditional distribution of Y, given A, W, and marginal distribution of W. Since $g(p)$ is now not updated, the algorithm will converge to a solution $P_0 D(g(p^0), Q) = 0$. One can show that

$$\begin{aligned} P_0 D(g(p^0), Q) &= \Psi(Q) - \Psi(Q_0) + E_0(Q - Q_0)(1, W)\frac{g(p^0) - g_0}{g(p^0)}(1 \mid W) \\ &\quad + E_0(Q - Q_0)(0, W)\frac{g_0 - g(p^0)}{1 - g(p^0)}(1 \mid W). \end{aligned}$$

Thus solving the equation $P_0 D(g(p^0), Q) = 0$ in Q does not provide a guarantee that $\Psi(Q) = \psi_0$, though it is certainly a possible solution. It is possible to prove a result that if g_m^0 converges for $m \to \infty$ at a certain rate to the the true g_0, then the limits or k-step versions of the targeted MLE algorithm with initial (g_m^0, Q_m^0) will give a ψ_{0m} converging much faster to ψ_0 than the original $\Psi(Q_m^0)$ converges to ψ_0, but where the gain in rate will be bounded from above by the rate at which second order term involving $g_m^0 - g_0$ and $Q_m^0 - Q_0$ will converge to zero.

Consider now the targeted MLE algorithm in which the hardest submodel $p(\epsilon)$ fluctuates both factors $Q(p)$ and $g(p)$ of p with scores $D(p) = D_{IPTW}(p) - D_{CAR}(p)$ for $Q(p)$ and $D_{CAR}(p) = \frac{Q(p)(A,W)}{g(p)(A|W)}(I(A = 1) - I(A = 0)) - (Q(p)(1, W) - Q(p)(0, W))$ for $g(p)$ at $\epsilon = 0$. Again, consider the targeted MLE algorithm at P_0 with starting density p^0. Under minor conditions the algorithm will converge to a solution of $P_0 D(g, Q) = 0$ and $P_0 D_{CAR}(g, Q) = 0$, and thus also of $P_0 D_{IPTW}(g, Q) = 0$. In addition, in this case g will be closer to g_0 than $g(p^0)$ since the algorithm involves at each step k maximization of the log-likelihood $g \to P_0 \log g$ along submodels through the current $g(p^k)$.

10 Application of targeted maximum likelihood estimation in unified loss based learning.

In van der Laan and Dudoit (2003), van der Laan et al. (2006), van der Vaart et al. (2006), van der Laan and Rubin (2005)) we have developed the theoretical underpinnings of a unified approach to statistical/machine learning, termed targeted empirical learning, which generalizes existing methods such as maximum likelihood estimation, estimating function based estimation, and nonparametric regression/classification. Targeted empirical learning is based on defining the parameter of interest, modelling this parameter while leaving nuisance parameters as unspecified as possible, and then developing targeted, robust, and highly efficient estimators of the parameter of interest. The random variables/experimental units considered in this context are typically longitudinal data structures observed on a randomly sampled subject, which may be subject to censoring, missingness and time-dependent confounding of the treatment variables of interest. The methodology we propose is applicable to such longitudinal data structure regardless of whether they arise in randomized trials or observational studies. Targeted empirical learning is also the approach to follow for testing a null hypothesis about the parameter of interest since it avoids bias due to modelling assumptions about nuisance parameters, a common problem with likelihood-based approaches that aim to estimate the entire distribution of the data instead of targeting the parameter of interest.

If the parameter of interest is pathwise differentiable, so that it is a relatively smooth function of the data generating distribution, then such targeted, robust estimators can be constructed using estimating functions (which are orthogonal to all the nuisance scores) for the parameter of interest (Robins and Rotnitzky (1992), van der Laan and Robins (2002)), and, as we now know, with the targeted MLE.

We have developed two general approaches for targeted empirical learning of general parameters (including non-smooth parameters), not only generalizing any of the currently available statistical learning methods such as maximum likelihood estimation, machine learning involving nonparametric estimation of regression functions or densities, estimating function methodology, but also providing a framework that is flexible enough to address any question of interest about a data generating experiment. These meth-

ods, namely unified loss-based learning and unified estimating function based learning, rely on defining the parameter of interest as a minimizer of the expectation of a loss function or as a minimizer of the norm of the expectation of a set (possibly infinite dimensional) of estimating functions (van der Laan and Dudoit (2003), van der Laan and Rubin (2005)), generating candidate estimators by minimizing the empirical counterparts of these loss-based or estimating function based criteria, and selecting the estimator that achieves the best bias-variance trade-off through cross-validation. In this discussion, we will focus on the loss based estimation approach.

Nuisance Parameter Estimation: Typically, the loss function is indexed by a possibly high dimensional nuisance parameter. It is often not hard to construct candidate estimators for this nuisance parameter, ranging from highly biased estimators with small variance to estimators with small bias but high variance. The main challenge instead lies in selecting among these candidate estimators of the nuisance parameters for the purpose of obtaining a good corresponding estimator of the parameter of interest. One of the main methodological challenges is thus to develop new methods for data adaptively selecting among candidate estimators of the nuisance parameter using a method that is targeted at the parameter of interest itself rather than at the nuisance parameter. The targeted maximum likelihood estimator is well suited for this purpose.

Unified loss based learning: We will now describe a short abstract summary of unified loss based learning and how targeted MLE can be applied to estimate the nuisance parameters of the loss function. Subsequently, we provide a concrete illustration. Suppose we observe n i.i.d. observations $O_1, \ldots, O_n$ of a random variable $O \sim p_0$. Let $\mathcal{M}$ be the model for p_0 (which is often nonparametric), and let $\Theta : \mathcal{M} \to D$ be the parameter of interest, which can be function valued so that D might represent a function space. Let $\mathbf{\Theta} = \{\Theta(p) : p \in \mathcal{M}\} \subset D$ denote the parameter space. Let $(O, \theta, p) \to L(O, \theta \mid p)$ be a so called loss (real valued) function defined on the cartesian product of the support of O, $\mathbf{\Theta}$, and $\mathcal{M}$, satisfying

$$\theta_0 = \arg\min_{\theta \in \mathbf{\Theta}} P_0 L(\theta \mid p_0),$$

where $P_0 L(\theta \mid p_0) = \int L(o, \theta \mid p_0) dP_0(o)$.

Consider now the parameter $\Psi(P)(\theta) \equiv PL(\theta \mid p)$, where $p = dP/d\mu$. That is, $\Psi(P)$ denotes the so called risk function at P in machine learning, and $\Psi(P)(\theta)$ denotes the risk at θ. Given θ, we wish to construct a good estimator of $\Psi(P)(\theta)$. Clearly, the challenge is how to estimate the unknown loss

function $L(\theta \mid p_0)$ for the purpose of obtaining a good estimate of $\Psi(P_0)(\theta)$. We note that $\Psi(P)(\theta)$ is a real valued and typically pathwise differentiable parameter. Let $p_{n\theta} \equiv \hat{\Phi}_\theta(P_n)$ be a targeted MLE density estimator of p_0, which is targeted towards the parameter $\Psi(P_0)(\theta)$. This results now in a targeted MLE estimator

$$\hat{\Psi}(P_n)(\theta) \equiv \Psi(p_{n\theta})(\theta)$$

of the risk function $\psi_0(\theta)$. Suppose that the model is nonparametric and the loss function is chosen so that $P_n L(\theta \mid p_0)$ is an efficient estimator of $\Psi(p_0)(\theta)$. In that case the efficient influence curve of $p \to \Psi(p)(\theta)$ at $p = p_0$ is $D_\theta(p_0) = L(\theta \mid p_0) - \Psi(p_0)(\theta)$. As a consequence, since the targeted MLE $p_{n,\theta}$ solves the efficient influence curve estimating equation, we have $P_n D_\theta(p_{n,\theta}) = 0$ and thus that the targeted MLE of the risk function $\Psi(p_{n,\theta})(\theta)$ equals the empirical mean $P_n L(\theta \mid p_{n,\theta})$ of the loss function with the nuisance parameter estimated with the targeted MLE:

$$\hat{\Psi}(P_n)(\theta) = P_n L(\theta \mid p_{n,\theta}).$$

Given this targeted MLE estimator of the risk, we can now proceed in a standard sieve based manner to construct candidate estimators of θ_0. Let $\mathbf{\Theta}_s$ be a subspace of the parameter space $\mathbf{\Theta}$ indexed by s ranging over some index set. We can now define s-specific estimators of θ_0 given by:

$$\hat{\Theta}_s(P_n) \equiv \arg\min_{\theta \in \mathbf{\Theta}_s} \hat{\Psi}(P_n)(\theta),$$

and, in a nonparametric model this can also be represented as

$$\hat{\Theta}_s(P_n) \equiv \arg\min_{\theta \in \mathbf{\Theta}_s} P_n L(\theta \mid p_{n,\theta}).$$

In order to select among these candidate estimators $\hat{\Theta}_s(P_n)$ of θ_0 we use cross-validation. That is, let $B_n \in \{0,1\}^n$ denote a random split in a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, and let P^0_{n,B_n}, P^1_{n,B_n} denote the corresponding empirical probability distributions, respectively. We will now estimate the loss function based on the training sample: $L(\theta \mid \hat{\Phi}_\theta(P^0_{n,B_n}))$, where we recall that $p_{n,\theta} = \hat{\Phi}_\theta(P_n)$. Subsequently, we proceed as usual by evaluating the empirical mean of the loss function over the validation sample at candidate estimators of θ_0 over the training sample. This gives the following cross-validation selector:

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

The final estimator is defined as:

$$\hat{\Theta}(P_n) \equiv \hat{\Theta}_{S(P_n)}(P_n).$$

In order to be specific we will illustrate this general application of targeted maximum likelihood estimation to the unified loss function based learning by using a concrete example.

10.1 Data adaptive estimation of the adjusted causal effect of a binary treatment.

Let $X = (W, (Y(a) : a)) \sim F_{X0}$ be the full data random variable of interest consisting of a set of baseline covariates W, and treatment specific outcomes $Y(a)$ indexed by a finite set of treatment values. For simplicity, we will assume that treatment is binary. Let $V \subset W$ be a subset of the covariates, and suppose we wish to estimate the causal effect of treatment adjusted by V. That is, our parameter of interest is $\theta_0(a, V) \equiv E_0(Y(a) \mid V)$. In addition, suppose we wish to estimate this parameter without assuming any particular functional form for $\theta_0(a, V)$. That is, our model for F_{X0} is nonparametric. A valid loss function for the full data structure for this causal effect is given by:

$$L(X, \theta) = \sum_a (Y(a) - \theta(a, V))^2 h(a, V),$$

where h can be an arbitrary function. This loss function satisfies that $\theta_0 = \arg\min_\theta E_0 L(X, \theta)$. Let $\mathbf{\Theta}$ denote the parameter space consisting of all functions θ of a, V. Consider the risk function $\psi_0(\theta) \equiv E_0 L(X, \theta)$. Given an estimate of this risk function at each value of θ, one can construct cross-validated sieve based estimators of θ_0 as in van der Laan and Dudoit (2003) (see above). In the following, $\psi_0(\theta)$ will represent our parameter of interest we wish to estimate with a targeted MLE $p_{n,\theta}$ based on observing n i.i.d. copies of the observed (missing) data structure $O = (W, A, Y(A)) \sim p_0$. Subsequently, we will show how this estimate is used to construct such cross-validated sieve based estimators, exactly analogous to the general presentation already given.

The observed data structure is $O = (W, A, Y = Y(A)) \sim p_0$, which is thus a missing data structure on X with missingness variable A. We assume that A is conditionally independent of X, given W, and that $0 < P(A = 1 \mid X) < 1$ F_{X0}-a.e. Let $g_0 = g(p_0)$ denote the conditional probability distribution

of A, given X, which is often called the treatment mechanism. This so called randomization assumption corresponds with the coarsening at random assumption for censored data, which teaches us that the data generating density factorizes as $p_0(O) = Q(p_0)(Y, A, W)g(p_0)(A \mid W)$, where $Q(p_0)(Y, A, W) = p_0(Y \mid A, W)p_0(W)$ is identified by the full data distribution F_{X0} through the relation $P_{Y(a),W}(y, w) = Q(p_0)(y, a, w)$. We make no further assumption about the observed data distribution p_0 so that the model $\mathcal{M}$ is nonparametric.

The double robust IPTW loss function, as used in Wang et al. (2006) and van der Laan (2006a), is given by:

$$\begin{aligned} L(O, \theta \mid g(p_0), Q(p_0)) \equiv & \frac{(Y - \theta(A, V))^2 h(A, V)}{g(p_0)(A \mid W)} \\ & - \frac{E_{Q(p_0)}((Y - \theta(A, V))^2 \mid A, W)h(A, V)}{g(p_0)(A \mid W)} \\ & + \sum_a E_{Q(p_0)}((Y - \theta(A, V))^2 \mid A = a, W)h(a, V). \end{aligned}$$

This loss function satisfies $E_0 L(O, \theta \mid g, Q) = \psi_0(\theta)$ if $g = g(p_0)$ and $0 < g(p_0)(a \mid W) < 1$ a.e., or $Q = Q(p_0)$. We have that $D_\theta(p_0)(O) = L(O, \theta \mid g(p_0), Q(p_0)) - \Psi(p_0)(\theta)$ is the efficient influence curve of $p \to \Psi(p)(\theta)$ at $p = p_0$. Because the model is nonparametric, it is also the only influence curve/gradient.

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

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

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

$$\begin{aligned} D_{1\theta}(p)(O) \equiv & D_\theta(p) - E_p(D_\theta(p) \mid A, W) \\ = & \frac{(Y - \theta(A, V))^2 h(A, V)}{g(p_0)(A \mid W)} \\ & - \frac{E_{Q(p_0)}((Y - \theta(A, V))^2 \mid A, W)h(A, V)}{g(p_0)(A \mid W)} \\ D_{2\theta}(p) \equiv & E_p(D_\theta(p) \mid W) - E_p(D_\theta(p)) \end{aligned}$$

$$
\begin{aligned}
&= \sum_a E_{Q(p)}((Y-\theta(A,V))^2 \mid A=a,W)h(a,V) - \Psi(p)(\theta) \\
D_{3\theta}(p)(O) &\equiv E_p(D_\theta(p) \mid A,W) - E_p(D_\theta(p) \mid W) \\
&= 0.
\end{aligned}
$$

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 the treatment mechanism $g(p_n^0)$. In addition, we have that $P_n D_2(p_n^0) = 0$ (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 also define $D_0(p) = \frac{(Y-\theta(A,V))^2 h(A,V)}{g(p)(A|W)}$ so that we have $D_1(p) = D_0(p) - \rho(p)$ with $\rho(p) = E(D_0(p) \mid A,W)$. As one dimensional submodel we consider the exponential family

$$
\left\{ p_n^0(\epsilon)(O) = p_n^0(W) g(p_n^0)(A \mid W) \frac{\exp(\epsilon(D_0(p_n^0)(O) - \rho(p_n^0)(A,W)))p_n^0(Y \mid A,W)}{E_{p_n^0}(\exp(\epsilon(D_0(p_n^0)(O) - \rho(p_n^0)(A,W)) \mid A,W)} : \epsilon \right\}, \tag{28}
$$

but we could also apply the normal regression model as in Section 6. To compute the first step targeted MLE we need to estimate ϵ with maximum likelihood based on an i.i.d. sample $\{O_i\}_{i=1}^n$. Exactly analogous to Section 6, it is shown that the maximum likelihood estimator ϵ_n satisfies $P_n D(p_n^0(\epsilon_n)) = 0$. This proves that the targeted maximum likelihood estimator is achieved in the first step of the algorithm and solves the efficient influence curve estimating equation $P_n D(p) = 0$.

We will denote this targeted maximum likelihood density estimator of p_0 with $p_{n,\theta}$, and its corresponding mapping from P_n to the density with $\hat{\Phi}_\theta(P_n)$: i.e. $p_{n,\theta} = \hat{\Phi}_\theta(P_n)$. Because the targeted MLE solves the efficient influence curve estimating equation, we have that the targeted MLE of $\psi_0(\theta)$ can be expressed as the empirical mean of the loss function $L(\theta \mid p_{n,\theta})$ with the nuisance parameters in the loss function estimated with the targeted MLE:

$$
\hat{\Psi}(P_n)(\theta) = \Psi(p_{n,\theta})(\theta) = P_n L(\theta \mid p_{n,\theta}).
$$

Let $\mathbf{\Theta}_s$ be a subspace of the parameter space $\mathbf{\Theta}$ indexed by s ranging over some index set. We can now define s-specific estimators of θ_0 given by:

$$
\hat{\Theta}_s(P_n) \equiv \arg\min_{\theta \in \mathbf{\Theta}_s} \hat{\Psi}(P_n)(\theta) = P_n L(\theta \mid \hat{\Phi}_\theta(P_n)).
$$

In order to select among these candidate estimators $P_n \rightarrow \hat{\Theta}_s(P_n)$ of θ_0 we use cross-validation:

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

The final estimator is defined as:

$$\hat{\Theta}(P_n) \equiv \hat{\Theta}_{S(P_n)}(P_n).$$

Finally, we note that the targeted MLE $p_{n,\theta,}$ of p_0 can also be modified so that the estimator of g_0 is also updated at each step, by using submodels $g(p^k_n)(\epsilon)$ through $g(p^k_n)$ at $\epsilon = 0$ with score (at $\epsilon = 0$)

$$\begin{aligned} D_{\theta,CAR} \equiv & -\frac{E_{Q(p^k_n))}((Y-\theta(A,V))^2 \mid A,W)h(A,V)}{g(p^k_n)(A \mid W)} \\ & + \sum_a E_{Q(p^k_n)}((Y-\theta(A,V))^2 \mid A=a,W)h(a,V). \end{aligned}$$

If one would use the latter targeted MLE $p_{n,\theta} = (g_{n,\theta}, Q_{n,\theta})$ of $p_0 = (g_0, Q_0)$, then one would have

$$P_n L_{IPTW}(\theta \mid g_{n,\theta}) = P_n L(\theta \mid (g_{n,\theta}, Q_{n,\theta}) = \Psi(p_{n,\theta})(\theta),$$

where $L_{IPTW}(\theta \mid g) = (Y-\theta(a,V))^2 h(a,V)/g(A \mid W)$ is the so called IPTW loss function, and $\Psi(p_{n,\theta})(\theta)$ is the likelihood based estimator of risk. So in this case the three types of estimators (DR-IPTW, IPTW, targeted likelihood) of the risk of a candidate θ are all identical.

11 Targeted maximum likelihood learning.

Loss based learning provides a general approach for the nonparametric/semiparametric estimation of nonregular (and often infinite dimensional) parameters, by naturally incorporating loss based cross-validation as a tool to trade off bias and variance of candidate estimators of the parameter of interest. In the previous section we showed that the targeted MLE could now be used to obtain a targeted estimate of the risk (mean of loss) function. In this section, we propose a (direct) targeted MLE approach of such infinite dimensional non-regular parameters, thereby still using the minus log-likelihood/minus log density as loss function.

Let $O_1, \ldots, O_n$ be i.i.d. copies of a random variable $O \sim p_0 \in \mathcal{M}$. Let $\Psi : \mathcal{M} \to D$ be the parameter we wish to estimate, where D can be a function space. In this case, we do *not* assume that Ψ is pathwise differentiable, but Ψ could be an infinite dimensional parameter. Let $\mathbf{\Psi} \equiv \{\Psi(p) : p \in \mathcal{M}\} \subset D$ be the parameter space. We wish to estimate $\psi_0 = \Psi(p_0)$.

Example 5 (**Nonparametric estimation of adjusted causal effect**) Let $O = (W, A, Y) \sim p_0$, W co-variates, A a binary treatment, and Y an outcome of interest. Let the parameter of interest be the V-adjusted variable importance $\Psi(p)(V) = E_p(E_p(Y \mid A = 1, W) - E_p(Y \mid A = 0, W) \mid V)$, which equals the causal effect $E(Y(1) - Y(0) \mid V)$ if one assumes the time ordering W, A, Y, the consistency assumption stating that $O = (W, A, Y = Y(A))$ is a missing data structure on $X = (W, Y(0), Y(1))$, and the randomization assumption $P(A = 1 \mid X) = P(A = 1 \mid W)$. Suppose that we are not willing to make any assumptions on the functional form of $\psi_0(V) = \Psi(p_0)(V)$. That is, our model for p_0 is nonparametric.

11.1 Outline of approach.

The approach we propose involves a number of steps. Firstly, one proposes a parametrization of the parameter space $\mathbf{\Psi}$ in terms of subsets I of basis functions and corresponding euclidean vector of coefficients $\beta_I = (\beta(j) : j \in I)$. It is assumed that each of these I-specific subspaces $\mathbf{\Psi}_I$ defines a pathwise differentiable parameter in our model by defining it as the "projection" of $\Psi(p)$ onto the I-specific finite dimensional subspace. Secondly, one selects a sieve $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ indexed by constraints s ranging over a set of possible constraints on the subsets I of basis functions, and thereby of the parameter space $\mathbf{\Psi}$. For each s and each I-specific subspace in $\mathbf{\Psi}_s$ one computes the targeted MLE density estimator targeted towards the I-specific pathwise differentiable parameter, and one selects the targeted MLE density estimator with maximal value of the log likelihood. This now provides a collection of s-specific targeted MLE density estimators. Finally, one selects s with likelihood based cross-validation. The proposed estimator of ψ_0 is the substitution estimator by plugging in the by likelihood based cross-validation selected targeted MLE density estimator into the parameter mapping Ψ. Since the targeted MLE's are also indexed by initial density estimators which itself might be indexed by constraints measuring how data adaptive they are, one might also wish to

select these constraints with likelihood based cross-validation. Therefore, in our approach outlined in the next subsection we include this option as well.

11.2 Stepwise presentation of targeted maximum likelihood learning.

Below, we provide the step by step implementation of this data adaptively selected targeted MLE of an infinite dimensional parameter.

Parametrization: Consider a parametrization of $\mathbf{\Psi}$ in terms of an index set I and a possibly infinite dimensional vector β:

$$\mathbf{\Psi} = \{\psi_{I,\beta} : I, \beta\}.$$

For example, $\psi_{I,\beta} = \sum_{j \in I} \beta(j)\phi_j$, where $\{\phi_j : j \in J\}$ is a basis for the function space D in which $\mathbf{\Psi}$ is embedded.

***I*-specific pathwise differentiable parameter:** Consider now the parameter $\Psi_I : \mathcal{M} \rightarrow \mathbf{\Psi}$ defined as

$$\Psi_I(p) = \inf_{\psi \in \mathbf{\Psi}_I} d(\psi, \Psi(p)).$$

That is, we simply define an I-specific parameter as the projection of the true $\Psi(p)$ onto the (finite dimensional) sub-space $\mathbf{\Psi}_I$. We assume that $\Psi_I : \mathcal{M} \rightarrow \mathbf{\Psi}_I$ is now a pathwise differentiable parameter.

***I*-specific targeted MLE(s):** Let $D_I(p)$ denote the efficient influence curve/canonical gradient at p. Let p^0_{nIl}, $l = 1, \ldots, L$ be a collection of possible initial density estimators of p_0, we will use to define the targeted MLE of ψ_{I0}, where l indicates a measure of how data adaptive (e.g., nonparametric) one selects these density estimators. Let now $p^*_{nIl} = \hat{\Phi}_{Il}(P_n)$ be the targeted MLE with respect to $\psi_{I0} = \Psi_I(P_0)$ based on the initial density estimator p^0_{nIl} and a hardest submodel $p^k_{nIl}(\epsilon)$ with score $D_I(p^k_{nIl})$ at $\epsilon = 0$ at step k of the targeted MLE algorithm. Because the starting density can affect the performance of the targeted MLE we decided to respect the fact that also this starting density might be indexed by a choice l (for each I), which will need to be selected with likelihood based cross-validation: see below.

Collection of targeted MLE(s): This provides us with a collection of candidate targeted MLE density estimators p^*_{nIl} indexed by I and starting density choices l, and corresponding targeted MLE $\hat{\Psi}_{Il}(P_n) = \Psi(p^*_{nIl})$ of ψ_{I0}. It remains to select (I, l).

Sieve on the parameter space: Consider now a sieve $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ indexed by vector valued s, where the index set is rich enough so that $\inf_s \inf_{\psi \in \mathbf{\Psi}_s} d(\psi_0, \psi) = 0$ with respect to some metric d. To be specific, given a vector function m measuring various measures of complexity of a candidate $\psi_{I,\beta}$, we define

$$\mathbf{\Psi}_s = \{\psi_{I,\beta} : m(I) \leq s\}.$$

For example, $m_1(I) =\mid I \mid$ might represent the number of non-zero coefficients, and $m_2(I)$ might represent a maximal complexity of the basis functions ϕ_j with $j \in I$.

(s,l)-specific targeted MLE(s): We now define the s-specific targeted MLE by

$$p^*_{nsl} = \hat{\Phi}_{s,l}(P_n) \equiv \arg \max_{\{I:m(I)\leq s\}} P_n \log p^*_{nIl}.$$

That is, to compute this estimator requires choosing the I-specific targeted MLE among all I with $m(I) \leq s$ with the maximal value of the log-likelihood. In practice, to approximate this maximization problem aggressive algorithms searching among subsets I might be required such as the DSA algorithm for variable selection in regression (Sinisi and van der Laan (2004)).

Likelihood based cross-validation to select s, l: We now propose to select l and s with likelihood based cross-validation. Thus we select s, l data adaptively with

$$(s_n, l_n) \equiv \arg \min_{s,l} E_{B_n} P^1_{n,B_n} \log \hat{\Phi}_{sl}(P^0_{n,B_n}),$$

to obtain a cross-validated targeted MLE density estimator $p^*_n \equiv p^*_{ns_nl_n}$, and corresponding cross-validated targeted MLE

$$\psi_n \equiv \Psi_{s_n}(p^*_{ns_nl_n}).$$

12 Example: Targeted maximum likelihood learning of W-adjusted variable importance

We now apply the targeted maximum likelihood learning template as presented in the previous section to a particular problem.

Let $O = (W, A, Y) \sim p_0$ and assume that the parameter of interest is $\Psi(p)(A, W) = E_p(Y \mid A, W) - E_p(Y \mid A = 0, W)$. Suppose that we are

not willing to make any assumptions on the functional form of $\psi_0(A,W) = \Psi(p_0)(A,W)$, beyond the known fact that $\psi_0(0,W) = 0$ for all W.

Let $\mathcal{M}$ only assume that the conditional distribution of Y, given A, W, is normally distributed, where $\theta(p)(A,W) = E_p(Y \mid A,W)$ denotes the mean of this conditional normal distribution, and $\sigma^2(p)(A,W)$ denotes its variance. Given a collection of basis functions $\{\phi_j : j\}$ satisfying $\phi_j(0,W) = 0$ for all W, we define a parametric subspace $\mathbf{\Psi}_I = \{\sum_{j\in I}\beta(j)\phi_j : \beta\}$ of the parameter space $\mathbf{\Psi}$. Let $\psi_{I,\beta} \equiv \sum_{j\in I}\beta(j)\phi_j$. We now define the pathwise differentiable parameter $\Psi_I : \mathcal{M} \to \mathbf{\Psi}$ by letting $\Psi_I(p)$ denote a "projection" of $\Psi(p)$ onto $\mathbf{\Psi}_I$. The precise definition involves the choice of an estimating function $h_I(p)(A,W)$ presented below, and defining $\Psi_I(p) = \sum_{j\in I}\beta(p)(j)\phi_j$ with $\beta(p)(j)$, $j \in I$ identified by the solution in β of

$$0 = E_p(h_I(p)(A,W) - E_p(h_I(p)(A,W) \mid W))(Y - \psi_{I,\beta}(A,W) - \theta(p)(0,W)).$$

I-specific targeted MLE

Let $O = (W,A,Y) \sim p_0$ and consider the semiparametric regression model $\mathcal{M}_I = \{p : E_p(Y \mid A,W) - E_p(Y \mid A = 0,W) = \psi_{I,\beta(p)}(A,W)\}$ for some parametrization $\beta \to \psi_{I,\beta}(A,W)$ satisfying $\psi_{I,\beta}(0,W) = 0$ for all $\beta \in \mathbb{R}^d$. Let $\beta_I(p)) = (\beta(p)(j) : j \in I)$ be the parameter of interest in this semiparametric regression model.

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

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

where

$$D_h(p)(O) \equiv (h(A,W) - E_p(h(A,W) \mid W))(Y - \psi_I(A,W \mid \beta(p)) - E_p(Y \mid A = 0,W)).$$

The orthogonal complement of the nuisance tangent space corresponds with the set of gradients for β_I at p given by:

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

where $c(p)(h) = \frac{d}{d\beta}E_pD_h(p,\beta)\Big|_{\beta=\beta(p)}$, 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 $h_I(p)$, which is provided below (29). Let $D_I(p) = D_{h_I(p)}(p)$ be this efficient influence curve at p as identified by this index $h_I(p)$.

Let $g(p)$ be the conditional density of A, given W, under p, and let $Q(p)$ be the conditional distribution of Y, given A, W, under p. We note that the parameter $\Psi_I(p)$ is only a function of $Q(p)$, and the density factorizes as $p(O) = p_W(W)g(p)(A \mid W)Q(p)(Y \mid A, W)$. 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 $D_I(p)$ of β_I at p into three subcomponents as follows:

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

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

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 $W_1, \ldots, W_n$. The above decomposition of the efficient influence curve $D_I(p)$ shows that a submodel $p_n^0(\epsilon)$ through p_n^0 with score $D_I(p_n^0)$ at $\epsilon = 0$ can be selected to only vary $Q(p_n^0)$ with a score $D_I(p_n^0)$ at $\epsilon = 0$. Such a submodel will now be presented.

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

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

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 in $\mathcal{M}_I$ through Q_n^0 at $\epsilon = 0$. Let $\beta(\epsilon) \equiv \beta(Q_n^0) + \epsilon$ and $\theta_n^0(\epsilon) = \theta(Q^0) + \epsilon^\top r_I$. It remains to find a function $r_I(W)$ so that the score of $Q_n^0(\epsilon)$ at $\epsilon = 0$ equals the efficient influence curve $D_I(p_n^0)$.

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

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

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

$$\begin{aligned} &h_I(p_n^0)(A,W)(Y - E_{Q^0}(Y \mid A,W)) = \\ &\tfrac{1}{\sigma^2(A,W)} \left\{ \tfrac{d}{d\beta(Q_n^0)}\psi_{I,\beta(Q_n^0)}(A,W) - r(W) \right\} (Y - E_{Q_n^0}(Y \mid A,W)). \end{aligned}$$

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

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

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

$$S(0) = h_I(p_n^0)(A,W)(Y - \psi_{I,\beta(Q_n^0)}(A,W)) - \theta(Q_n^0)(W)),$$

where

$$\begin{aligned} h_I(p_n^0)(A,W) &= \frac{1}{\sigma^2(A,W)} \frac{d}{d\beta(Q_n^0)}\psi_{I,\beta(Q_n^0)}(A,W)) \\ &\quad - \frac{1}{\sigma^2(A,W)} \frac{E_{p_n^0}\left(\frac{d}{d\beta(Q_n^0)}\psi_{I,\beta(Q_n^0)}(A,W))/\sigma^2(A,W) \mid W \right)}{E_{p_n^0}(1/\sigma^2(A,W) \mid W)} \end{aligned} \quad (29)$$

This choice $h_I(p_n^0)$ corresponds with the efficient influence curve. 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 for β_I.

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

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

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

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} \frac{1}{\sigma^2(A_i, W_i)} \left(Y_i - \psi_{I,\beta_n^0+\epsilon}(A_i, W_i) - \theta_n^0(W_i) - \epsilon^\top r_I(p_n^0)(W_i)\right)^2 (30)$$

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

$$0 = P_n \frac{\left\{\frac{d}{\beta_n^0(\epsilon)}\psi_{I,\beta_n^0(\epsilon)} - r_I(p_n^0)\right\}(Y - \psi_{I,\beta_n^0(\epsilon)} - \theta_n^0 - \epsilon r_I(p_n^0))}{\sigma^2}.$$

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

$$0 = P_n \frac{\{m_I - r_I(p_n^0)\}(Y - (\beta_n^0 + \epsilon_n)m_I - \theta_n^0 - \epsilon_n^\top r_I(p_n^0))}{\sigma^2}. \quad (31)$$

Firstly, we note that ϵ_n exist in closed form:

$$\epsilon_n = A_n^{-1} P_n \frac{\{m_I - r_I(p_n^0)\}(Y - \beta_n^0 m_I - \theta_n^0}{\sigma^2},$$

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

$$A_n \equiv P_n \frac{1}{\sigma^2}\left\{m_I - r_I(p_n^0)\right\}(m_I + r_I(p_n^0))^\top.$$

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 when this first step targeted MLE $p_n^1 \equiv p_n^0(\epsilon_n)$ already solves the efficient score equation: $P_n D_I(p_n^1) = P_n D_I(p_n^0(\epsilon_n)) = 0$. We have that $P_n D_I(p_n^0(\epsilon_n))$ is given by

$$P_n \frac{\{m_I - r_I(p_n^0(\epsilon_n))\}(Y - (\beta_n^0 + \epsilon_n)m_I - \theta_n^0 - \epsilon_n^\top r_I(p^0(\epsilon_n)))}{\sigma^2}.$$

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

$$P_n \frac{\{m_I - r_I(p_n^0)\}(Y - (\beta_n^0 + \epsilon_n)m_I - \theta_n^0 - \epsilon_n r_I(p_n^0))}{\sigma^2},$$

but the latter equals zero by the fact that $P_n S(\epsilon_n) = 0$. This proves that, if $\psi_{I,\beta}$ 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 $P_n D_I(p) = 0$. If one would also update $\sigma^2(A, W)$ in the submodel $p_n^0(\epsilon)$, then the algorithm would have to be iterated to converge to a targeted MLE solving $P_n D_I(p) = 0$. For nonlinear models $\psi_{I,\beta}$ the targeted MLE algorithm will also need to be iterated till convergence.

Collection of targeted MLE(s).

The initial density estimator p_n^0 is indexed by I because $E_{p_n^0}(Y \mid A, W) - E_{p_n^0}(Y \mid A = 0, W) = \psi_{I,\beta(p_n^0)}(A, W))$. In addition, it might also be indexed by different choices of the fit of the treatment mechanism and $E(Y|A = 0, W)$: for example, both of these nuisance parameters might be fitted with a machine learning algorithm indexed by various fine tuning constraints measuring how agressive the algorithm searched the space of possible regressions. We will denote these latter choices with l. Therefore, the targeted MLE of β_I is indexed by I and l. For example, to obtain an initial estimator $E_{p_n^0}(Y \mid A, W)$ with parametric component $\psi_{I,\beta}$ one could use the backfitting algorithm, which would then be indexed by fine tuning parameters measuring how nonparametric the nonparametric component $E_{p_n^0}(Y \mid A = 0, W)$ is fitted.

This provides us with a collection of candidate targeted MLE density estimators p_{nII}^* indexed by I and starting density choices l, and corresponding targeted MLE $\hat{\Psi}_{Il}(P_n) = \Psi_I(p_{nIl}^*)$ of ψ_{I0}. It remains to select (I, l).

Data adaptive targeted MLE for given constraints.

Sieve on the parameter space: Consider now a sieve $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ indexed by s. To be specific, given a vector function m measuring various measures of complexity of a candidate subset of basis functions, we define

$$\mathbf{\Psi}_s = \cup_{\{I:m(I)\leq s\}} \psi_{\mathbf{I}}.$$

For example, $m_1(I) =\mid I \mid$ might represent the number of non-zero coefficients, and $m_2(I)$ might represent a maximal complexity (e.g., the order) of the basis functions ϕ_j with $j \in I$.

(s,l)**-specific targeted MLE(s):** For each choice of l and s, we define the following subset estimator of basis functions

$$I_n(s,l) = I(s,l)(P_n) \equiv arg \max_{\{I:m(I)\leq s\}} P_n \log p^*_{nIl}.$$

Since the targeted MLE's only differ in their fit of the regression $E(Y|A,W)$ this maximization problem corresponds with minimizing the residual sum of squared errors, possibly weighted by $1/\sigma^2(A,W)$ over all (I,l)-specific regression fits $E_{p^*_{nIl}}(Y \mid A,W)$. To compute this estimator $I_n(s,l)$ requires choosing the I-specific targeted MLE among all I with $m(I) \leq s$ with the maximal value of the log-likelihood, or, in this case, minimal value of RSS. In practice, to approximate this maximization problem aggressive algorithms searching among subsets I might be required such as the DSA algorithm for variable selection in regression (Sinisi and van der Laan (2004)).

We now define the (s,l)-specific targeted MLE density estimator by

$$p^*_{nsl} \equiv p^*_{nI_n(s,l)l}.$$

Let's denote the algorithm from the data P_n to this (s,l)-specific targeted MLE with $\hat{\Phi}_{sl}(P_n)$: $p^*_{nsl} = \hat{\Phi}_{sl}(P_n)$.

Likelihood based cross-validation to select s,l**:** We now select l (indicating the initial density used in the targeted MLE's) and s (indicating constraints on the parameter space of the parameter of interest) with likelihood based cross-validation. Thus we select s,l data adaptively with

$$(s_n,l_n) \equiv \arg\min_{s,l} E_{B_n} P^1_{n,B_n} \log \hat{\Phi}_{sl}(P^0_{n,B_n}).$$

Because the densities p^*_{nsl} are all normally distributed and only differ in their regression fit, it follows that (s_n,l_n) simply indicates the regression estimator of $E(Y \mid A,W)$ among all (s,l)-specific regression fits $E_{p^*_{nsl}}(Y \mid A,W)$ with minimal cross-validated RSS. We now obtain a cross-validated targeted MLE density estimator

$$p^*_n \equiv p^*_{ns_nl_n},$$

and corresponding cross-validated targeted MLE of ψ_0 given by

$$\psi_n \equiv \Psi(p^*_n).$$

13 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 which are not necessarily dominated by a single dominating measure.

Given a density estimator we defined a targeted density estimator through an iterative maximum likelihood algorithm along hardest sub models with a score equal to the efficient influence curve of the parameter of interest. This tool allows us to map any candidate density into its targeted version. 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. We also used the increase of the log-likelihood during the targeted MLE algorithm as a powerful criteria for evaluating nuisance parameter fits, thereby providing a road map for a new and general class of targeted MLE algorithms which also fully target the fitting of the nuisance parameters: see Section 8. 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.

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 a pathwise differentiable parameter of a full data distribution F_X in CAR censored data models, as in (van der Laan and Robins (2002)), if one applies the targeted MLE to an initial $p_n^0 = g_n^0 Q_n^0$ of $p_0 = g_0 Q_0$, 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 (2002).

References

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

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

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

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

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.

L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

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.

I. M. Johnstone. Oracle inequalities and nonparametric function estimation. *Journal der Deutchen Mathematiker Vereinigung, Proc. of the International Congress of Mathematicians, Berlin 1998*, III:267–278, 1998.

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

M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1996.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. Technical report, Department of Mathematics, Paris-Sud, 1998.

R. Neugebauer and M. J. van der Laan. Why prefer double robust estimates? Illustration with causal point treatment studies. Technical Report 115, Division of Biostatistics, University of California, Berkeley, 2002.

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 and S. Mark. 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.

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.

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

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

M. Talagrand. A new look at independence. *Ann. Probab.*, 24:1–34, 1996a.

M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996b.

M. J. van der Laan, S. Dudoit, and S. Keleş. Asymptotic optimality of likelihood based cross-validation. Technical Report 125, Division of Biostatistics, University of California, Berkeley, 2003.

M.J. van der Laan. Causal effect models for intention to treat and realistic individualized treatment rules. Technical report 203, Division of Biostatistics, University of California, Berkeley, 2006a.

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

M.J. van der Laan. Identity for npmle in missing data and biased sampling models. *Bernoulli*, 1(4):335–341, 1995a.

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

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 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, 2002.

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

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.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 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.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 2006.

Y. Wang, O. Bembom, and M.J. van der Laan. Data adaptive estimation of the treatment specific mean. *Journal of Statistical Planning and Inference*, 2006.

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