University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2011 *Paper* 282

Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models

Ivan Diaz Munoz* Mark J. van der Laan†

*University of California, Berkeley, School of Public Health - Division of Biostatistics, idiaz@jhu.edu

†University of California, Berkeley; School of Public Health - Division of Biostatistics, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper282

Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models

Ivan Diaz Munoz and Mark J. van der Laan

Abstract

In this paper we present a histogram-like estimator of a conditional density that uses super learner crossvalidation to estimate the histogram probabilities, as well as the optimal number and position of the bins. This estimator is an alternative to kernel density estimators when the dimension of the problem is large. We demonstrate its applicability to estimation of Marginal Structural Model (MSM) parameters in which an initial estimator of the treatment %mechanism is needed. MSM estimation based on the proposed density estimator results in less biased estimates, when compared to estimates based on a misspecified parametric model.

1 Introduction

Conditional density estimation is one of the most important problems in statistics. Parametric models such as generalized linear models intend to estimate the conditional density of an outcome given a set of predictors by assuming a functional form that is known up to a finite-dimensional vector of real parameters. If the assumptions made about the functional form of the conditional density reflect characteristics of the true data generating mechanism, maximum likelihood estimation techniques yield consistent and efficient estimators of the parameters of the model and consequently of the conditional density (van der Vaart, 1998).

However, it is very common to find applications in fields such as epidemiology and social studies in which little information about the true data generating mechanism is known, and the researcher does not have enough scientific knowledge to assume a functional form for the conditional density. For such cases, non parametric estimators such as kernel density estimators, which do not assume a pre-specified functional form have been proposed. Kernel estimation was introduced by Rosenblatt (1969), and has been extensively studied in the statistics literature since then. As a remarkable property, under certain conditions on the true density, the univariate kernel density estimator has been proven to have mean integrated square (MISE) error of order $n^{-4/5}$, which is only $n^{-1/5}$ times larger than the MISE of a parametric density estimator if the true density was known to belong to some parametric model (van der Vaart, 1998). A comprehensive description of univariate and multivariate kernel density estimators and their statistical properties can be found in Wand and Jones (1995) and Scott (1992). The multivariate kernel density estimator can be used to find estimates of the joint densities involved in the definition of any conditional density and produce a plug-in estimator. Nevertheless, unless the number of covariates is very small (Wand and Jones (1995) suggests less than 6) or the sample size is extremely large, these estimators suffer from the curse of dimensionality, and the resulting estimates are highly biased.

Cross validation selection from a library of candidates of estimators has been proven to have optimal properties in terms of the risk of the resulting estimator (van der Vaart et al., 2006). In particular, the super learner (van der Laan et al., 2007) is a machine learning technique that uses cross-validated risks to find an optimal convex combination of candidate estimators in a user-supplied library. One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution).

In Section 2, we propose a histogram-like estimator of the density of a continuous outcome conditioned on a set of covariates. Our proposed estimator uses cross validation to find an optimal convex combination of candidates in a library consisting of histogram-like density estimators indexed by the number and the position of the bins, and the choice of an estimator for the histogram probabilities. In this paper, the histogram probabilities are estimated by using the super learner as described by van der Laan et al. (2007), whose properties have been widely studied (van der Laan et al. (2004) and van der Laan and Dudoit (2003)).

One of the various applications of conditional density estimation to inference problems has to do with the estimation of exposure mechanisms in causal models and their use to estimate causal parameters. In Section 3 we use simulation to compare the performance of three different estimators of causal parameters in a marginal structural model when three different estimators of the exposure mechanism are used: a correctly specified parametric model, an incorrectly specified parametric model and our histogram-like estimator. As expected, the estimator of the causal

parameter that uses a correctly specified parametric model for the treatment density outperforms the other two, but the estimator based on the proposed histogram-like density estimator is better than the estimator that uses the misspecified parametric model.

2 Density Estimator

Let A be a random variable representing an outcome of interest, and let W be a random vector containing a set of predictors of A. We are interested in finding an estimator of $g_0(A|W)$, the true conditional density function of A given W. As explained in the introduction, we will use the super learner to choose a convex combination of estimators among a library of candidates consisting of histogram density estimators defined by hazard functions. In the following subsections we will define the super learner, the candidate estimators in the library, and present the cross validated estimator of the conditional density.

2.1 Super Learner

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

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

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

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

Let $v \in \{1, \ldots, V\}$ index a sample split into a validation sample $V(v) \subset \{1, \ldots, n\}$ and a training sample $T(v) = (V(v))^c$. Here we note that the union of the validation samples equals the total sample: $\cup_{v=1}^{V} V(v) = \{1, \ldots, n\}$, and the validations samples are disjoint: $V(v_1) \cap V(v_2) = \emptyset$ for $v_1 \neq v_2$. Let $P_{T(v)}$ be the empirical distribution of the training sample v. The cross validated estimator of the risk is given by the following expression, in which the parameter is estimated on a training set and the risk is estimated in the corresponding validation set:

$$E_{B_n} R\{\hat{\Psi}(P_{T(v)}), P_{V(n)}\} = E_{B_n} \int L\{o, \hat{\Psi}(P_{T(v)})\} dP_{V(v)}(o). \quad (1)$$

Now, if we have a library of candidate estimators $\hat{\Psi}_j : j \in J$, the so-called discrete super learner will select the estimator in this library for which the cross validated risk in (1) is the smallest. If we want to expand our library of candidate estimators by considering all possible convex linear

combinations of the candidates $\hat{\Psi}_j$, it can be shown that the candidate in the new library with the smallest cross validated risk will be given by

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

where

$$\beta = (\beta_1, \ldots, \beta_J) = \arg\min_{\beta} \sum_{v=1}^{V} \sum_{i \in v} L\left\{ O_i,\ \sum_{j \in J} \beta_j \hat{\Psi}_j(P_{T(v)}) \right\}, \tag{2}$$

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

2.2 Candidates

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

$$g_{n,\alpha}(P_n)(a|W) = \frac{Pr_n(A \in [\alpha_{t-1}, \alpha_t)|W)}{\alpha_t - \alpha_{t-1}},\ \text{for } \alpha_{t-1} \leq a < \alpha_t, \tag{3}$$

where we note that the choice of the values α_t $(t = 0, \ldots, k)$ implies defining the number and position of the bins. Here Pr_n denotes an estimator of the true probability $Pr(A \in [\alpha_{t-1}, \alpha_t)|W)$ obtained through a hazard specification and use of a model for binary variables in a pooled repeated measures dataset, as explained below. Note that we consider the estimator in (2.2) as a mapping that takes the empirical distribution P_n and maps it into an estimate of the conditional density of A given W. This notation will be helpful later in the section when we define the cross-validated estimator. Note also that

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

The likelihood for model (3) is now proportional to

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

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

$$Pr(A \in [\alpha_{t-1}, \alpha_t)|A \geq \alpha_{t-1}, W) \tag{4}$$

include the following logistic model with only main terms:

$$\text{logit}\left\{Pr(A \in [\alpha_{t-1}, \alpha_t) | A \geq \alpha_{t-1}, W)\right\} = \sum_{j=1}^{t} \gamma_j I_{[\alpha_{j-1}, \alpha_j)}(A) + \sum_{l=1}^{p} \theta_l W_l, \tag{5}$$

where we assume the dimension of W is p, and $I_{[\alpha_{j-1}, \alpha_j)}(A)$ denotes an indicator of $A \in [\alpha_{j-1}, \alpha_j)$. Another candidate might be given by a logistic model including double interaction terms. In general, any estimator that has the potential of providing an accurate representation of the underlying true data generating mechanism can be postulated as a candidate for estimation of (4), including a super learner algorithm that takes all available candidate estimators and finds an optimal convex combination of them. Each candidate estimator in (2.2) is now indexed by choice of the values α_t and choice of an algorithm for estimating (4).

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

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

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

We now define our library of candidate estimators for the conditional density as a collection of estimators in (2) by defining values of the vector α through different choices of c and k, and defining an estimator for the probabilities in (4). The use of this approach will result in estimators that are able to identify regions of high density as well as provide a good description of the tails and outliers of the density. For the sake of simplicity, we will only consider one candidate for estimation of (4): the super learner itself with candidates that may include, for example, the logistic model in (5). Since the choice of each α only depends on c and k, the candidate estimators $g_{n,\alpha}$ in (3) will now be denoted by $g_{n,j}$, where $j \in J$ is an index identifying a combination of c and k.

2.3 Cross Validation

Consider the cross validation scheme presented in section 2.1. We define our estimator of the conditional density of A given W as

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

where

$$\beta = (\beta_1, \ldots, \beta_J) = \arg\min_{\beta} \sum_{v=1}^{V} \sum_{i \in v} \log \sum_{j \in J} \beta_j \; g_{n,j}(P_{T(v)})(A_i|W_i), \tag{6}$$

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

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

The minimization in (6) is carried out by using the augmented Lagrange multiplier method as implemented in the R function `solnp()` (Ghalanos and Theussl, 2010). Technical details about the implementation of this method can be found in Ye (1987).

3 Marginal Structural Model Estimation

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

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(A, W, U_Y), \end{aligned} \tag{7}$$

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

$$P_0(O) = P_0(Y|A, W) P_0(A|W) P_0(W), \tag{8}$$

where we denote $g_0(A|W) \equiv P_0(A|W)$, $\bar{Q}_0(A, W) \equiv E_0(Y|A, W)$, and $Q_{W,0}(W) \equiv P_0(W)$. Causal inference parameters are usually defined in terms of the distribution of the counterfactual outcome Y_a that one would obtain in a controlled experiment in which the equation corresponding to A in (7) is removed from the SCM and the treatment A is set to be equal to some pre-specified value a deterministically.

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

$$\beta_0 = \arg \min_{\beta \in B \subset \mathbb{R}^d} \int_{\mathcal{A}} L\{m(a), m_\beta(a)\} h(a)\ d\mu(a), \tag{9}$$

where $\mathcal{A}$ is the support of A, $h(a)$ is a stabilizing weight function, L is a loss function that describes the loss obtained by approximating $m(a)$ with $m_\beta(a)$, and μ is an appropriate measure. If $L\{m(a), m_\beta(a)\}$ is a convex function of β, the parameter can also be defined as the value $\beta_0 = (\beta_{01}, \ldots, \beta_{0d})'$ that solves the system of equations

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

The most intuitive loss function to use is

$$\begin{aligned} L\{m(a), m_\beta(a)\} &= \{m(a) - m_\beta(a)\}^2 \\ \frac{\partial}{\partial \beta_j} L\{m(a), m_\beta(a)\} &= -2\{m(a) - m_\beta(a)\} \frac{\partial m_\beta(a)}{\partial \beta_j}, \end{aligned} \tag{10}$$

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

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

In this paper we focus in the estimation of parameters defined in terms of (10), but similar calculations can be made for other parameters defined by different loss functions.

Since $m(a)$ is identified as a function of the distribution of the observed data by $E(\bar{Q}_0(a, W))$, the parameter of interest is identified as the value β_0 that solves

$$\int_{\mathcal{A}} \frac{\partial}{\partial \beta_j} L(E(\bar{Q}_0(a, W)), m_\beta(a)) h(a)\ d\mu(a) = 0;\ j = 1, \ldots, d. \tag{11}$$

4 Estimators

In this section we describe three possible estimators for the parameter β_0 of a MSM defined in the previous section. The first estimator is an IPTW estimator that requires a consistent estimator of the exposure mechanism in order to be consistent. The second estimator is an augmented IPTW (A-IPTW) that solves the efficient influence curve equation and requires initial estimators of $\bar{Q}_0$ and g_0; it is consistent if either of them is consistent, and it is efficient if both are consistent. The third one is a targeted maximum likelihood estimator (TMLE) that has the same properties as the A-IPTW, plus some additional advantages, like being a substitution estimator and not having multiple solutions.

4.1 IPTW

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

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

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

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

We will use notation $D_{IPTW}(O)$ or $D_{IPTW}(O|g,\beta)$ depending on whether it is necessary to emphasize the dependence on g and β.

4.2 Augmented IPTW

The efficient influence curve $D(O)$ of (9) in the non-parametric model can be found through the IPTW estimating function $D_{IPTW}(O)$ as

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

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

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

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

$$D_j(O|\bar{Q},g,\beta) = (Y - \bar{Q}(A,W))\frac{h(A)}{g(A|W)}\frac{\partial m_\beta(A)}{\partial \beta_j} + \int_{\mathcal{A}} (\bar{Q}(a,W) - m_\beta(a))\frac{\partial m_\beta(a)}{\partial \beta_j}h(a)\ d\mu(a), \quad (12)$$

and the A-IPTW estimator is defined as the value $\beta_{n,2}$ that solves the system of estimating equations

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

Note that the efficient influence curve can be decomposed into three components corresponding to the orthogonal decomposition of the tangent space implied by the factorization (8) as:

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

where

$$
\begin{aligned}
D_{j1}(O) &= D_j(O) - E(D_j(O)|A,W) = (Y - \bar{Q}(A,W))\frac{h(A)}{g(A|W)}\frac{\partial m_\beta(A)}{\partial \beta_j}, \\
D_{j2}(O) &= E(D_j(O)|A,W) - E(D_j(O)|W) = 0, \\
D_{j3}(O) &= E(D_j(O)|W) - E(D_j(O)) = \int_{\mathcal{A}} (\bar{Q}(a,W) - m_\beta(a))\frac{\partial m_\beta(a)}{\partial \beta_j} h(a)\ d\mu(a).
\end{aligned} \tag{13}
$$

This decomposition will be useful in the next section to define the TMLE.

4.3 Targeted Maximum Likelihood Estimator

In order to define a targeted maximum likelihood estimator (van der Laan and Rubin, 2006) for β_0, we need first to define three elements: (1) A loss function $L(Q)$ for the relevant part of the likelihood required to evaluate β_0, which in this case is $Q = (\bar{Q}, Q_W)$, that must satisfy $Q_0 = \arg\min_Q E_{P_0} L(Q)(O)$, where Q_0 denotes the true value of Q; (2) An initial estimator Q_n^0 of Q_0; (3) A parametric fluctuation $Q(\epsilon)$ through Q_n^0 such that the linear span of $\frac{d}{d\epsilon} L(Q(\epsilon))|_{\epsilon=0}$ contains all the components of the efficient influence curve $D(P)$ defined in (12). These three elements are defined below:

Loss Function

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

Parametric Fluctuation

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

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

$$
\text{logit}\, \bar{Q}_n^1(\epsilon)(A,W) = \text{logit}\, \bar{Q}_n^0(A,W) + \sum_{j=1}^d \epsilon_j H_j^0(A,W),
$$

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

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

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

Initial Estimators

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

Targeted Maximum Likelihood Estimator

The TMLE of β_0 is now defined as the value $\beta_{n,3}$ that solves the equations

$$\int_{\mathcal{A}} \frac{\partial}{\partial \beta_j} L\{E_{Q_{W,n}} \bar{Q}_n^1(\epsilon_n)(a, W), m_\beta(a)\} h(a) \; d\mu(a) = 0; \; j = 1, \ldots, d. \tag{14}$$

5 Simulation

Consider the following data generating process

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

We are interested in estimating the parameter defined in (9) with

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

and $h(a)$ equal to the marginal density of A. Note that the efficient influence curve calculations made in the previous sections remain valid in this case, and that estimators of g_0 and $Q_{W,0}$ define an estimator of h. The true value of the parameter for this data generating distribution is $\beta_0^0 = -1.0067$ and $\beta_0^1 = 0.1520$.

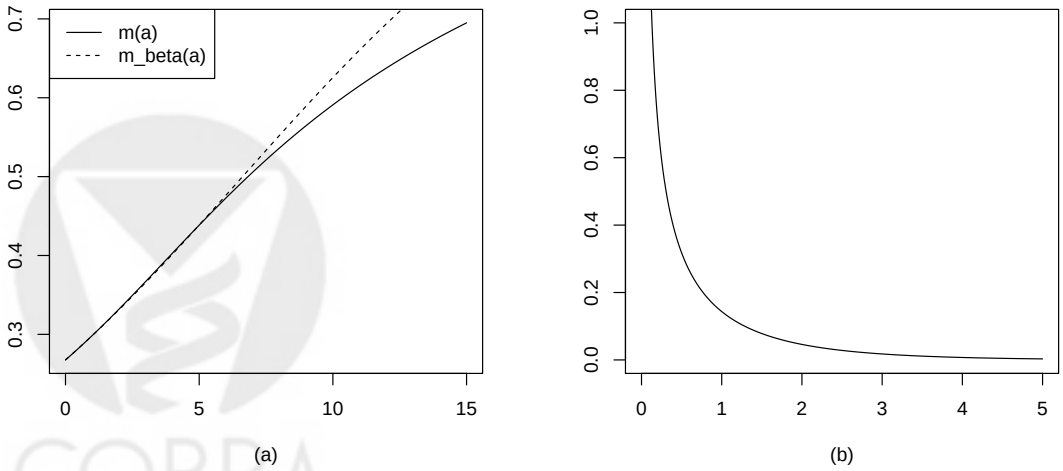

Figure 1: (a) True counterfactual expectation and MSM true curve. (b) marginal density of A.

Figure 1 presents the true counterfactual expectation $m(a)$ as well as the true MSM curve $m_\beta(a)$. Since the definition of the MSM parameter involves weighting by the marginal density of

A, the approximation of m_β to m is almost perfect in areas of high density, at the cost of a poor approximation in the areas in which A has low density.

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

The conditional density estimator proposed in section 2 involves two cross validation procedures: an internal one performed in order to estimates the probabilitites in (4), and an external one performed in order to estimate the risks of each estimator defined in (3). The high computational cost of these double cross validation made it prohibitive to use Monte Carlo simulation to assess the properties of the MSM estimator. Instead, we drew a sample of size 10.000 from the true data generating mechanism, and computed the three estimates defined in the previous section. Table 1 shows the results. Given the large sample size, a direct comparison of the estimates with the true

		(a)		(b)		(c)	
		β_0	β_1	β_0	β_1	β_0	β_1
(1)	IPTW	-1.0342	0.1171	-1.5406	0.3634	-1.0076	0.1055
	A-IPTW	-1.0331	0.1224	-1.9081	0.7968	-1.0127	0.1210
	TMLE	-1.0787	0.1402	-1.0006	0.1178	-1.0073	0.1376
(2)	IPTW	-1.0342	0.1171	-1.5406	0.3634	-1.0076	0.1055
	A-IPTW	-1.0301	0.1105	-1.9401	0.7935	-1.0064	0.0979
	TMLE	-1.0783	0.1360	-0.9923	0.0835	-1.0141	0.1342

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

value of the parameters provides an approximation to their bias. It is known that (up to positivity assumptions) the TMLE and the A-IPTW are double robust in the sense that they are unbiased if at least one of the initial estimators is consistent. The IPTW requires consistency of the estimator for the treatment mechanism in order to be unbiased.

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

References

L. Denby and C. Mallows. Variations on the histogram. *Journal of Computational and Graphical Statistics*, Vol. 18, Iss. 1:21–31, 2009.

A. Ghalanos and S. Theussl. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package*, 2010.

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

M. Rosenblatt. Conditional probability density and regression estimates. *Multivariate Analysis II, Ed. P.R. Krishnaiah*, 22:25–31, 1969.

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

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 D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.

M.P. Wand and M.C. Jones. *Kernel smoothing*. Monographs on statistics and applied probability. Chapman & Hall, 1995. ISBN 9780412552700.

Yinyu Ye. *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Department of EES, Stanford University, 1987.