The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 38

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

Iván Díaz Muñoz, *University of California, Berkeley*
Mark J. van der Laan, *University of California, Berkeley*

Recommended Citation:

Díaz Muñoz, Iván and van der Laan, Mark J. (2011) "Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 38.

DOI: 10.2202/1557-4679.1356

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

Iván Díaz Muñoz and Mark J. van der Laan

Abstract

In this paper, we present a histogram-like estimator of a conditional density that uses cross-validation 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 covariate vector is large. We demonstrate its applicability to estimation of Marginal Structural Model (MSM) parameters in which an initial estimator of the exposure 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.

KEYWORDS: causal inference, conditional density, cross validation, efficient estimation, marginal structural models, oracle selector

1 Introduction

Conditional density estimation is one of the most important problems in statistics, as a parameter of interest itself, or as an input for estimation of other parameters. The latter is the case of many of the problems in causal inference in which the causal relationship between an exposure and an outcome is confounded by a set of covariates (Pearl, 2000). The marginal density of the exposure given the covariates (also called the exposure mechanism) is the continuous equivalent of the propensity score for binary outcomes, which has historically played a central role in the estimation of causal effects in observational studies (Rosenbaum and Rubin, 1983, R. Mansson, 2007). Therefore, it is not surprising that estimation of causal parameters for continuous exposures requires the specification of an initial estimator of the exposure mechanism. Two examples of causal parameters for continuous treatments are given by Marginal Structural Models (Robins, Hernan, and Brumback, 2000, Neugebauer and van der Laan, 2007) and causal parameters corresponding to stochastic interventions (Díaz-Munoz and van der Laan, 2011). In this paper we develop a machine learning estimator of a conditional density that can be used to estimate the exposure mechanism and demonstrate its use in the estimation of MSM parameters.

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

Parametric models such as generalized linear models intend to estimate the conditional density of a variable given a set of predictors by assuming a functional form that is known up to a finite-dimensional vector of real parameters. If the assumptions made about the functional form of the conditional density reflect characteristics of the true data generating mechanism, maximum likelihood 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 we knew that the true density belonged 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 the conditional density and compute 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. This is an important issue in causal inference, since the number of confounders is often large.

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

In Section 2 we propose a conditional density estimator that starts with a list of histogram-like density estimators indexed by the number of bins, their position, and the choice of an estimator for the histogram probabilities, and then uses the super learner to find the optimal estimator in the library given by the convex hull of this list of candidate estimators. We use the super learner itself to estimate the histogram probabilities of each of the estimators in the initial list.

A review of Marginal Structural Models as described by Neugebauer and van der Laan (2007) is provided in Section 3, as well as three MSM estimators that require an initial estimator of the exposure mechanism. The performance of our method in the context of MSM estimation is assessed through a simulation study in which the three estimators described in Section 3 are computed under three different estimators of the exposure mechanism: a correctly specified parametric model, an incorrectly specified parametric model and our super learner based estimator. The results of this simulation are presented in Section 4, and Section 5 includes an application example of in which the causal effect of physical activity on all cause mortality is defined through an MSM. Finally, Section 6 provides some concluding remarks and directions of future research.

2 Density Estimator

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

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

2.1 Super Learner

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

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

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

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

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

risk is given by the following expression, in which the parameter is estimated on a training set and the risk is estimated in the corresponding validation set:

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

Assume that we have a list of candidate estimators $\hat{\Psi}_j : j \in J$. The discrete super learner is defined as the estimator in this list for which the cross validated risk in (1) is the smallest. Consider now a library of candidate estimators given by all possible convex linear combinations of the candidates $\hat{\Psi}_j$. It can be shown (van der Laan et al., 2007) that the candidate in this library with the smallest cross validated risk is given by

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

where

$$\beta = (\beta_1, \dots, \beta_J) = \arg \min_{\beta} \frac{1}{S} \sum_{s=1}^S \frac{1}{n_s} \sum_{i \in V(s)} L\left\{O_i, \sum_{j \in J} \beta_j \hat{\Psi}_j(P_{T(s)})\right\}, \quad (2)$$

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

2.2 Candidates

Consider a sequence of values $\alpha_0, \dots, \alpha_k$ 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:

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

where we note that the choice of the values α_t ($t = 0, \dots, k$) 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 (3) as a mapping that takes the empirical distribution P_n and maps it into an estimate of the conditional density of A given W ,

this notation will be helpful later in the section when we define the cross-validated estimator. Note also that

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

The likelihood for model (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

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

include the following logistic model with only main terms:

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

where we assume the dimension of W is p , and $I_{[\alpha_{j-1}, \infty)}(A)$ denotes an indicator of $A \in [\alpha_{j-1}, \infty)$. Another candidate might be given by a logistic model including double interaction terms. In general, any estimator that has the potential of providing an accurate representation of the underlying true data generating mechanism can be postulated as a candidate for estimation of (4), including a super learner algorithm that takes all available candidate estimators and finds an optimal convex combination of them. Each candidate estimator in (3) 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 $x + cy = bh$, $b = 1, \dots, k + 1$; where c and h are parameters defining the slope and the distance between lines, respectively. The parameter h identifies the number of bins k . The authors note that $c = 0$ gives the usual histogram, whereas $c \rightarrow \infty$ 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

$$\beta = (\beta_1, \dots, \beta_J) = \arg \max_{\beta} \frac{1}{S} \sum_{s=1}^S \frac{1}{n_s} \sum_{i \in V(s)} \log \sum_{j \in J} \beta_j g_{n,j}(P_{T(s)})(A_i|W_i), \quad (6)$$

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. Our library of estimators includes all the estimators given by convex combinations of $g_{n,j}(A|W)$ for $j \in J$, and the oracle selector is given by the candidate estimator in the library that minimizes the Kullback-Leibler divergence with respect to the true data-generating distribution

The minimization in (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

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

The consistency of the MSM estimators presented in this section will be used in Section 4 to assess the performance of the density estimator proposed in Section 2 when it is used to estimate the exposure mechanism g_0 .

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

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

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 $g_0(A|W) \equiv P_0(A|W)$, $\bar{Q}_0(A, W) \equiv E_0(Y|A, W)$, $Q_{W,0}(W) \equiv P_0(W)$, and $Pf = \int f dP$ for a function f of O . Causal inference parameters are usually defined in terms of the distribution of the counterfactual outcome Y_a that one would obtain in a controlled experiment in which the equation corresponding to A in (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:

$$\gamma_0 = \arg \min_{\gamma \in B \subset \mathbb{R}^d} \int_{\mathcal{A}} L\{m(a), m_\gamma(a)\} h(a) d\mu(a), \quad (9)$$

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 the true causal curve $m(a)$ with $m_\gamma(a)$, and μ is an appropriate measure (i.e., the Lebesgue or the counting measure). If $L\{m(a), m_\gamma(a)\}$ is a convex function of γ , the parameter can also be defined as the value $\gamma_0 = (\gamma_{01}, \dots, \gamma_{0d})'$ that solves the system of equations

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

The most intuitive loss function to use is

$$\begin{aligned} L\{m(a), m_\gamma(a)\} &= \{m(a) - m_\gamma(a)\}^2 \\ \frac{\partial}{\partial \gamma_j} L\{m(a), m_\gamma(a)\} &= -2\{m(a) - m_\gamma(a)\} \frac{\partial m_\gamma(a)}{\partial \gamma_j}, \end{aligned} \quad (10)$$

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

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

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

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

$$\int_{\mathcal{A}} \frac{\partial}{\partial \gamma_j} L\{E(\bar{Q}_0(a, W)), m_\gamma(a)\} h(a) d\mu(a) = 0; \quad j = 1, \dots, d. \quad (11)$$

3.1 Estimators

In this section we describe three possible estimators for the parameter γ_0 of a MSM defined in the previous section. The first estimator is an Inverse Probability of Treatment Weighted (IPTW) estimator which requires a consistent estimator of the exposure mechanism in order to be consistent. The second estimator is an augmented

IPTW that solves the efficient influence curve equation and requires initial estimators of $\bar{Q}_0$ and g_0 , it is consistent if either of them is consistent, and it is efficient if both are consistent. The third estimator is a targeted maximum likelihood estimator (TMLE) that has the same properties as the A-IPTW, plus additional advantages that include being a substitution estimator and not having multiple solutions.

3.1.1 IPTW

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

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

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

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

The IPTW is an asymptotically linear estimator with influence function D_{IPTWj} , therefore the variable $\sqrt{n}(\gamma_{n,1,j} - \gamma_{0j})$ converges in distribution to a random variable distributed as $N\{0, P_0 D_{IPTWj}^2(\cdot | g_0, \gamma_0)\}$, whose variance can be estimated by $P_n D_{IPTWj}^2(\cdot | g_n^0, \gamma_{n,1})$, where P_n denotes the empirical measure. Van der Laan and Robins (2003) prove that this variance estimator is conservative. We will use notation $D_{IPTW}(O)$ or $D_{IPTW}(O|g, \gamma)$ depending on whether it is necessary to emphasize the dependence on g and γ .

3.1.2 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

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

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

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

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

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

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

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

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

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

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

where

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

3.1.3 Targeted Maximum Likelihood Estimator

In order to define a targeted maximum likelihood estimator for γ_0 , we need first to define three elements: (1) A loss function $L(Q)$ for the relevant part of the likelihood required to evaluate γ_0 , which in this case is $Q = (\bar{Q}, Q_W)$. This function must satisfy $Q_0 = \arg \min_Q E_{P_0} L(Q)(O)$, where Q_0 denotes the true value of Q ; (2)

An initial estimator Q_n^0 of Q_0 ; (3) A parametric fluctuation $Q(\varepsilon)$ through Q_n^0 such that the linear span of $\frac{d}{d\varepsilon}L\{Q(\varepsilon)\}|_{\varepsilon=0}$ contains all the components of the efficient influence curve $D(O)$ defined in (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:

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

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

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

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

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

Initial Estimators

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

Targeted Maximum Likelihood Estimator

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

$$\int_{\mathcal{A}} \frac{\partial}{\partial \gamma_j} L\{E_{Q_{W,n}} \bar{Q}_n^1(\varepsilon_n)(a, W), m_\gamma(a)\} h(a) d\mu(a) = 0; \quad j = 1, \dots, d. \quad (14)$$

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

4 Simulation

Consider the following data generating process

$$W_1 \sim \text{Unif}\{0, 1\}.$$

$$W_2 \sim \text{Ber}\{0.7\}.$$

$$A \sim \text{Gamma}\{.3 + 3\log(W_1 + 1) + 2.2\exp(W_1)W_2^{-1}, 1\}.$$

$$Y \sim \text{Ber}\{\text{expit}(-1 + .05A - .02AW_2 + .2A \tan(W_1^2) - .02W_1W_2 + .1AW_1W_2)\}.$$

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

$$m_\gamma(a) = \frac{1}{1 + \exp(-\gamma^0 - \gamma^1 a)}, \quad (15)$$

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

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

A , the approximation of m_{γ_0} to m is almost perfect in areas of high density, at the cost of a poor approximation in the areas in which A has low density.

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; and two different initial estimators were considered for the expectation of Y given A and W : (1) correctly specified parametric model, and (2) logistic regression with only linear terms. The choice of the misspecification of the models performed in (b) and (2) comes from usual practice in parametric modeling in epidemiology, in which for the sake of ease of interpretation and calculation, linear models without interactions are usually assumed.

The prohibitive computational cost of the cross-validation procedure resulting in the proposed conditional density estimator restrained us from using Monte Carlo simulation to 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 exposure mechanism estimate as well as the three estimates. Figure 2 shows the estimates and true value of the conditional densities for two given profiles, obtained by using a list of candidate estimators in (3) defined by all combination of values $k = 5, 7, 9, 11, 13$ and $c = 800, 500, 300, 100, 50, 10, 0.5, 0.01$. From this graph we can see that this estimator is very close to the true data exposure mechanism, which is a surprising fact given that it does not use any knowledge about the true density or the true parametric model.

Table 1 shows the three MSM estimates for model (15) along with their standard errors. Given the large sample size, a direct comparison of the estimates with the true value of the parameters provides an approximation to their bias. It is known that (up to positivity assumptions) the TMLE and the A-IPTW are double robust in the sense that they are unbiased if at least one of the initial estimators is consistent. The IPTW requires consistency of the estimator for the treatment mechanism in order to be unbiased.

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

Figure 2: (a) Estimated and true density for profile $W_1 = 0.09$, $W_2 = 1$. (b) Estimated and true density for profile $W_1 = 0.99$, $W_2 = 1$.

5 Application

With the objective of demonstrating the use of the exposure mechanism estimator provided in Section 2, we revisit the problem analyzed by Bembom and van der Laan (2007) and Diaz-Munoz and van der Laan (2011) of assessing the extent to which physical activity in the elderly is associated with reductions in cardiovascular morbidity and mortality, and improvement in, or prevention of metabolic abnormalities. Tager, Hollenberg, and Satariano (1998) followed a group of people over 55 years of age living around Sonoma, CA, over a time period of about ten years as part of a longitudinal study of physical activity and fitness (Study of Physical Performance and Age Related Changes in Sonomans - SPPARCS). The goal in analyzing the data that were collected as part of this study is to examine the effect of baseline vigorous LTPA (Leisure Time Physical Activity) on subsequent five-year all-cause mortality.

We use the same measure of LTPA used by Bembom and van der Laan (2007), which is a continuous score based on the number of hours that the participants were engaged in vigorous physical activities such as jogging, swimming,

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

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

bicycling on hills, or racquetball in the last seven days, and the standard intensity values in metabolic equivalents (MET: Metabolic Equivalent of Task) of such activities, where one MET is approximately equal to the oxygen consumption required for sitting quietly. The primary confounding factors that we adjust for are described in Table 2. Age and gender are natural confounders, and the rest of the variables intend to account for the subject's underlying level of general health. Of the 2092 subjects enrolled in the SPPARCS study, 40 were missing information in at least one of this variables; our analysis is based on the remaining 2052 subjects.

In the sequel of this section, the vector containing the confounders will be denoted by W , the continuous MET score by A , and the indicator of five-year all-cause mortality by Y . We are interested in summarizing the causal relationship between LTPA and all cause mortality based on the MSM provided in (15) through estimation of the parameters involved.

Figure 1 shows two contrasting estimated densities $g_n(A|W)$ for different profiles W , in which a subject with better general health status is more likely to have higher levels of physical activity. As pointed out before, this methodology allows the detection of high density areas in the exposure mechanism, like the one

Table 2: Confounders.

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

detected at zero in Figure 3 (a). This spike appears because this is a “zero-inflated” exposure, in which a large proportion of the population do not practice any amount of physical activity.

As initial estimator of $\bar{Q}_0$ we also used the super learner (van der Laan et al., 2007). Table 3 shows the candidates used, their cross-validated risks, and their coefficients in the final super learner predictor. In order to get a consistent estimator of $\bar{Q}_0$ the library of candidate estimators should be as large as possible. Since this is an illustrating example, we allow ourselves to use this small library. Table 4 shows the three estimated values for each of the two parameters defined by the MSM in (15). Computation of these estimates required (as explained in Section 3) an initial estimator of the exposure mechanism. The simulation in Section 4 showed that misspecification of a parametric model for the exposure mechanism can lead to a substantial amount of bias in the MSM estimates, and that the use of the density estimator presented in Section 2 is preferred when the true exposure

Figure 3: Estimated conditional density of A given the profiles: (a) age = 77, gender = female, health = fair, nrb = 0.9, card = no, smoke = ex-smoker, decline = yes, chron = yes; and (b) age = 71, gender = male, health = good, nrb = 0.88, card = no, smoke = never smoked, decline = no, chron = no

Table 3: Super learner output for estimation of $\bar{Q}_0$.

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

mechanism is unknown. Although the three estimators for γ_1 differ in magnitude, all of them agree that there is a small protective effect of physical activity on all cause mortality, although that effect is not statistically significant, which does not mean that it is not relevant.

Table 4: Estimates of the MSM parameters (3) for the phisycal activity data.

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

6 Conclusion

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

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

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

References

- Bembom, O. and M. van der Laan (2007): “A practical illustration of the importance of realistic individualized treatment rules in causal inference,” *Electronic Journal of Statistics*.
- Bryan, J., Z. Yu, and M. van der Laan (2003): “Analysis of longitudinal marginal structural models,” *Biostatistics*, 5, 361–380.
- Denby, L. and C. Mallows (2009): “Variations on the histogram,” *Journal of Computational and Graphical Statistics*, Vol. 18, Iss. 1, 21–31.
- Diaz-Munoz, I. and M. J. van der Laan (2011): “Population intervention causal effects based on stochastic interventions,” Technical report, U.C. Berkeley.
- Ghalanos, A. and S. Theussl (2010): *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package*.
- Neugebauer, R. and M. J. van der Laan (2007): “Nonparametric causal effects based on marginal structural models,” *Journal of Statistical Planning and Inference*, 137, 419 – 434, URL <http://www.sciencedirect.com/science/article/pii/S0378375806000334>.
- Pearl, J. (2000): *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.
- R. Mansson, W. S. S. H., M.M. Joffe (2007): “On the estimation and use of propensity scores in case-control and case-cohort studies.” *Am J Epidemiol*, 166, 332–339.
- Robins, J. M., M. A. Hernan, and B. Brumback (2000): “Marginal structural models and causal inference in epidemiology,” *Epidemiology*, 11, 550–560.
- Rosenbaum, P. and D. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70, 41–55.
- Rosenblatt, M. (1969): “Conditional probability density and regression estimates,” *Multivariate Analysis II*, Ed. P.R. Krishnaiah, 22, 25–31.
- Scott, D. (1992): *Multivariate density estimation: theory, practice, and visualization*, Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley.
- Tager, I., M. Hollenberg, and W. Satariano (1998): “Self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population,” .
- van der Laan, M. and S. Dudoit (2003): “Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples,” Technical report, Division of Biostatistics, University of California, Berkeley.

- van der Laan, M., S. Dudoit, and S. Keles (2004): “Asymptotic optimality of likelihood-based cross-validation,” *Statistical Applications in Genetics and Molecular Biology*, 3.
- van der Laan, M., E. Polley, and A. Hubbard (2007): “Super learner,” *Statistical Applications in Genetics and Molecular Biology*, 6.
- van der Laan, M. and J. Robins (2003): *Unified methods for censored longitudinal data and causality*, Springer, New York.
- van der Laan, M. and D. Rubin (2006): “Targeted maximum likelihood learning,” *The International Journal of Biostatistics*, 2.
- van der Vaart, A., S. Dudoit, and M. van der Laan (2006): “Oracle inequalities for multi-fold cross-validation,” *Statistics and Decisions*, 24, 351–371.
- van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.
- Wand, M. and M. Jones (1995): *Kernel smoothing*, Monographs on statistics and applied probability, Chapman & Hall.
- Ye, Y. (1987): *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*, Ph.D. thesis, Department of EES, Stanford University.