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

Year 2010 *Paper* 257

Targeted Maximum Likelihood Estimation of the Parameter of a Marginal Structural Model

Michael Rosenblum* Mark J. van der Laan†

*Johns Hopkins University, mrosenbl@jhsph.edu

†University of California - Berkeley, laan@berkeley.edu

Targeted Maximum Likelihood Estimation of the Parameter of a Marginal Structural Model

Michael Rosenblum and Mark J. van der Laan

Abstract

Targeted maximum likelihood estimation is a versatile tool for estimating parameters in semiparametric and nonparametric models. We work through an example applying targeted maximum likelihood methodology to estimate the parameter of a marginal structural model. In the case we consider, we show how this can be easily done by clever use of standard statistical software. We point out differences between targeted maximum likelihood estimation and other approaches (including estimating function based methods). The application we consider is to estimate the effect of adherence to antiretroviral medications on virologic failure in HIV positive individuals.

1 Introduction

Targeted maximum likelihood estimation (van der Laan and Rubin, 2006; Moore and van der Laan, 2007; Polley and van der Laan, 2009; van der Laan et al., 2009; Rosenblum and van der Laan, 2010; van der Laan, 2010a,b) is a versatile tool for estimating parameters in semiparametric and nonparametric models. For example, in the area of causal inference, it can be used to estimate (i) the effect of static or dynamic treatments, (ii) direct and indirect effects of treatments/exposures, and (iii) the parameters or marginal structural models and structural nested models. Targeted maximum likelihood estimation can be applied in analyzing cross-sectional as well as longitudinal data, and data with censoring and missing values. It is useful for analyzing data from observational studies as well as from randomized trials.

We work through an example applying the methodology of targeted maximum likelihood to estimate the parameter of a marginal structural model. Robins (1997) introduced marginal structural models, an important tool in causal inference. Our example is a slightly simplified version of an analysis of the REACH cohort, an observational study of HIV positive, marginally housed and homeless individuals in San Francisco (Rosenblum et al., 2009); we estimate the impact of percent adherence to antiretroviral therapy on virologic failure, conditioning on duration of past HIV suppression.

We point out differences between targeted maximum likelihood estimation and other approaches, including estimating function based methods (Robins and Rotnitzky, 1992; Robins et al., 1994; Robins, 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan, 2002; van der Laan and Robins, 2002). We do not give an overview of past work related to targeted maximum likelihood in this paper, but instead refer the reader to (van der Laan et al., 2009). We do, however, note that in important special cases the estimators we present are examples of estimators in (Scharfstein et al., 1999, Rejoinder to Comments, pages 1141-2).

The papers (van der Laan, 2010a,b) give general methods for applying targeted maximum likelihood methodology to estimate the causal effect of single and multiple time point interventions. Here we describe a particular application of these methods to single time point interventions.

In the next section, we give an overview of the targeted maximum likelihood algorithm. Then, in Section 3, we apply targeted maximum likelihood methodology in a setting where baseline variables, a treatment/exposure, and an outcome are observed for n individuals; here, the goal is to estimate the mean outcome, at a set level of treatment/exposure, adjusting for the baseline variables. This is a special case of the more general problem we tackle

in Section 4, where we work through an example giving a targeted maximum likelihood estimate of the parameter of a marginal structural model.

2 Targeted Maximum Likelihood Estimation

Targeted maximum likelihood estimation is an algorithm[1] for constructing a substitution (or "plug-in") estimator for a given parameter ψ, in a (often nonparametric or semiparametric) model $\mathcal{M}$. Here, by parameter, we generally mean a smooth[2] function from the data generating distribution p to a real-valued vector $\psi(p)$. Examples of such smooth parameters include the mean treatment difference in a randomized trial, the odds ratio of survival at a given time point under a static or dynamic treatment regime, and the parameter of a marginal structural model. We assume that the model $\mathcal{M}$ can be represented as a set of densities with respect to some known measure μ.

Targeted maximum likelihood estimation involves the following three steps: (i) constructing an initial estimate $\hat{p}_0$ of the density of the data generating distribution, (ii) using the efficient influence function of the parameter to find a better fit $\hat{p}_1$ targeted at minimizing mean squared error for estimation of the parameter ψ, and (iii) computing the substitution estimator $\psi(\hat{p}_1)$ at this estimated density. In general, step (ii) is iterated until convergence (defined below), though in many examples a single iteration suffices. Below, we elaborate on the three steps of the targeted maximum likelihood algorithm. Then, in the following sections, we apply the targeted maximum likelihood algorithm to estimate a treatment specific mean, and then later to estimate the parameter of a marginal structural model.

Targeted Maximum Likelihood Algorithm

The first step in the targeted maximum likelihood algorithm involves constructing an initial estimator of the density of the data generating distribution. This can be done in a variety of ways, e.g. by kernel smoothing, by fitting parametric working models, by machine learning algorithms such as classification and regression trees (Breiman et al., 1984), or by other approaches. One option is to use cross-validation to select among from a variety of estimation methods, as in (van der Laan et al., 2007). We denote the initial estimator for

[1] Technically, targeted maximum likelihood estimation is a template for an algorithm, since it allows the user to make choices in its implementation, such as the choice of initial density estimators. Nevertheless, we refer to it simply as an algorithm here.

[2] By "smooth" we mean pathwise differentiable, as defined in e.g. (van der Vaart, 1998).

the density of the data generating distribution by $\hat{p}_0$.

The second step in the targeted maximum likelihood algorithm is to update the density estimate $\hat{p}_0$ to a new density estimate $\hat{p}_1$, where the goal, intuitively, is to minimize the mean squared error of the resulting substitution estimator for the parameter of interest ψ. This is done by constructing a parametric model $\{p(\epsilon) : \epsilon \in (-\delta, \delta)\}$ in the overall model $\mathcal{M}$ that (i) equals the initial density estimate $\hat{p}_0$ at $\epsilon = 0$ and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence function of the parameter ψ at $\hat{p}_0$. We then use maximum likelihood estimation in the parametric model $\{p(\epsilon) : \epsilon \in (-\delta, \delta)\}$ to get an estimate $\hat{\epsilon}$ for ϵ. Our updated density estimate is then set to be $\hat{p}_1 := p(\hat{\epsilon})$. In general, we iterate this step, replacing $\hat{p}_0$ by $\hat{p}_1$, until convergence (that is, until the estimated coefficient $\hat{\epsilon}$ is sufficiently small).

The motivation for updating the density estimate in this way is that we'd like to reduce bias in our estimate for the parameter ψ, while minimally increasing the variance in this estimate. The parameter is most sensitive to small changes in the data generating distribution in the direction corresponding to the efficient influence function, to first order. Thus, we hope that by restricting our update of the density to the (estimated) direction in which the parameter is most sensitive, we will achieve bias reduction at the smallest expense in increased variance. We give examples in the following sections of how to construct parametric models having the properties (i) and (ii) of the previous paragraph.

Recall that the parameter ψ, by definition, maps each density p in the model $\mathcal{M}$ to a real-valued vector denoted by $\psi(p)$. The last step in the targeted maximum likelihood algorithm is to compute the substitution (or "plug in ") estimator for the parameter ψ at our final density estimate. That is, we evaluate the parameter ψ at the final density estimate. For example, if the parameter is the mean of a distribution, we would output the mean of the distribution corresponding to the final density estimate output in step two of the targeted maximum likelihood algorithm. Similarly, if the parameter were the odds ratio of survival at time t, we would compute this odds ratio, as if the final density estimate were the true data generating distribution, and report this odds ratio as our estimate of the parameter.

The targeted maximum likelihood estimator has many desirable properties. First, in many settings, it is a doubly robust, locally efficient estimator for censored data and causal inference models, as described in (van der Laan and Rubin, 2006) and in the papers (van der Laan, 2010a,b) in this issue. The targeted maximum likelihood estimator approximately solves the efficient influence curve estimating function, for nuisance parameter values obtained by a substitution estimator at the final density estimate. Often the targeted maxi-

mum likelihood algorithm can be easily implemented using standard statistical software.

We now briefly point out some differences between the targeted maximum likelihood estimator and estimating-function based estimators, such as those in (Robins and Rotnitzky, 1992; Robins et al., 1994; Robins, 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan, 2002; van der Laan and Robins, 2002). In contrast to targeted maximum likelihood estimators, estimating-function based estimators (i) may lead to multiple solutions to the estimating function, without criteria for selecting among them, (ii) may not respect global model constraints, such as parameters being restricted to the range $[0, 1]$, (iii) are only defined for situations where the efficient influence function is an estimating function in the parameter of interest. Targeted maximum likelihood estimators do not suffer from these problems.

3 Targeted Maximum Likelihood Estimate of the Treatment Specific Mean

We apply the targeted maximum likelihood algorithm from the previous section to a particular example: estimating the treatment specific mean adjusting for baseline confounders. This is a special case of estimating the parameter of a marginal structural model, and may be helpful for understanding the development in Section 4. At the end of this section, we describe how in an important special case, our estimator coincides with an estimator given in (Scharfstein et al., 1999, Rejoinder to Comments).

We assume our set of data consists of n independent, identically distributed realizations $\{(W_i, A_i, Y_i)\}_{i=1}^n$ of a random vector (W, A, Y), where W represents baseline variables, A is a binary treatment, and Y is a binary outcome. For example, $Y = 1$ may represent virologic failure, $A = 1$ may indicate adherence to an antiretroviral regimen in the last month (i.e. taking all pills as prescribed), and W may be a list of baseline potential confounders of the effect of adherence on virologic failure, such as CD4 count, past adherence, depression, etc.

Our model for the joint density of (W, A, Y) is nonparametric. (By "joint density" we mean density in the general sense, so that for discrete variables it represents a frequency function.) We will use $p = p_Y(Y|A, W)p_A(A|W)p_W(W)$ to denote the joint density of (W, A, Y). The parameter we're interested in is the mean of Y given $A = 1$, adjusting for the baseline variables W; more

precisely, the parameter we want to estimate is

$$\psi(p) := E_p[p(Y = 1|A = 1, W)],$$

where the outer expectation is with respect to the marginal distribution of W according to the density p. We call this parameter the "treatment specific mean." Under certain assumptions (see Section 4 for these assumptions), ψ can be interpreted as the causal effect of setting $A = 1$ on the population mean of Y.

Before applying the targeted maximum likelihood algorithm, we note that the efficient influence function for ψ in the nonparametric model, at a density p is

$$\frac{A(Y - p(Y = 1|A = 1, W))}{p(A = 1|W)} + p(Y = 1|A = 1, W) - E_p[p(Y = 1|A = 1, W)]. \quad (1)$$

This, and the efficient influence functions for a variety of parameters and models can be found in e.g. (Bickel et al., 1993; van der Laan and Robins, 2002). We now show one way to implement the targeted maximum likelihood algorithm, as described in the previous section, to the problem of estimating the treatment specific mean.

Step 1 of Targeted Maximum Likelihood Algorithm Applied to Estimating the Treatment Specific Mean: Initial Density Estimate

The first step of the targeted maximum likelihood algorithm is to compute an initial density estimator for the joint density of (W, A, Y). For simplicity we use parametric working models to estimate the density of Y given A, W and the density of A given W. However, we note that in general one could use more nonparametric approaches. We estimate the marginal distribution of the baseline variables W with the empirical distribution of W, and denote it by $\hat{p}_{0W}$.

We use a logistic regression working model for the density of Y given A, W:

$$P(Y = 1|A, W) = \text{logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 W + \beta_3 AW). \quad (2)$$

This is just one possible choice of terms for the model–any set of terms can be included. We fit the model with maximum likelihood estimation, based on our n independent, identically distributed observations $\{(W_i, A_i, Y_i)\}_{i=1}^n$, to produce $\hat{\beta}$, and set $\hat{p}_{0Y}(Y|A, W)$ to be the density corresponding to the fit model:

$$\hat{p}_{0Y}(Y = 1|A, W) := \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW). \quad (3)$$

Using a similar procedure, we estimate the density of A given W using logistic regression with terms $1, W, W^2$, and obtain a fit

$$\hat{p}_{0A}(A = 1|W) := \text{logit}^{-1}(\hat{\gamma}_0 + \hat{\gamma}_1 W + \hat{\gamma}_2 W^2). \tag{4}$$

Combining the above, we let $\hat{p}_0 = \hat{p}_{0Y}(Y|A, W)\hat{p}_{0A}(A|W)\hat{p}_{0W}(W)$ be our initial estimator for the joint density of (W, A, Y). It follows from our decision to set $\hat{p}_{0W}(W)$ to be the empirical distribution of W, that the substitution estimator for ψ at the initial density estimate $\hat{p}_0$ is

$$\hat{\psi}_0 := \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{0Y}(Y = 1|A = 1, W_i). \tag{5}$$

This can be equivalently expressed as

$$\sum_{i=1}^{n} [\hat{p}_{0Y}(Y = 1|A = 1, W_i) - \hat{\psi}_0] = 0. \tag{6}$$

Step 2 of Targeted Maximum Likelihood Algorithm Applied to Estimating the Treatment Specific Mean: Constructing and Fitting a Certain Parametric Model

The second step in the targeted maximum likelihood algorithm involves constructing a parametric working model $\{p(\epsilon) : \epsilon \in (-\delta, \delta)\}$ that (i) equals the initial density estimate $\hat{p}_0$ at $\epsilon = 0$ and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence function of the parameter ψ at $\hat{p}_0$ as given in (1). We then fit this parametric model using maximum likelihood estimation to obtain an updated density estimate we denote by $\hat{p}_1$.

Our method for constructing a parametric working model satisfying (i) and (ii) is to define a logistic regression model with two terms in the linear part: the "clever covariates":

$$C_1(A, W) := A/\hat{p}_{0A}(A = 1|W),$$

$$C_2(W) := \hat{p}_{0Y}(Y = 1|A = 1, W) - \hat{\psi}_0.$$

This type of procedure works in many situations, and is not limited to binary outcomes; depending on the problem at hand, instead of using a logistic regression working model, one can use a linear model, Poisson regression model, or any of a set of commonly used generalized linear models with canonical link functions (Rosenblum and van der Laan, 2010; van der Laan, 2010a,b).

For each $\epsilon = (\epsilon_1, \epsilon_2)$, we define the density in the parametric model

$$p(\epsilon)(Y, A, W) := p_Y(\epsilon)(Y|A, W)p_A(\epsilon)(A|W)p_W(\epsilon)(W),$$

where we set

$$\begin{aligned} p_Y(\epsilon)(Y = 1|A, W) &:= \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW \\ &\qquad + \epsilon_1 C_1(A, W)), \qquad (7) \\ p_A(\epsilon)(A|W) &:= \hat{p}_{0A}(A|W), \\ p_W(\epsilon)(W) &:= s_{\epsilon_2} \exp(\epsilon_2 C_2(W))\hat{p}_{0W}(W), \end{aligned}$$

where the constant $s_{\epsilon_2} := 1/[\frac{1}{n}\sum_{i=1}^n \exp(\epsilon_2 C_2(W_i))]$ is chosen so that $p_W(\epsilon)(w)$ integrates to 1 for each ϵ. In this definition of our parametric model, we consider $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\psi}_0$ as fixed numbers (having been computed above). (Here we slightly abuse notation, sometimes writing $\epsilon = 0$ to mean $\epsilon = (0, 0)$.) It follows that conditions (i) and (ii) hold for this model, since substituting 0 for ϵ results in the initial density $\hat{p}_0$, and the components of the score at $\epsilon = 0$ are:

$$\begin{aligned} \frac{d}{d\epsilon_1}[\log p(\epsilon)(Y, A, W)]|_{\epsilon=0} &= \frac{d}{d\epsilon_1}[\log p_Y(\epsilon)(Y|A, W)]|_{\epsilon=0} \\ &= (Y - \hat{p}_{0Y}(Y = 1|A = 1, W))C_1(A, W), \quad (8) \end{aligned}$$

and

$$\frac{d}{d\epsilon_2}[\log p(\epsilon)(Y, A, W)]|_{\epsilon=0} = \frac{d}{d\epsilon_2}[\log p_W(\epsilon)(W)]|_{\epsilon=0} = C_2(W). \qquad (9)$$

Substituting the definitions of $C_1(A, W)$ and $C_2(W)$ in (8) and (9), we see that the linear span of these components of the score at $\epsilon = 0$ includes the efficient influence function at $\hat{p}_0$ as given in (1). Thus condition (ii) above is satisfied.

We fit the above parametric model with maximum likelihood estimation, to obtain estimates $(\hat{\epsilon}_1, \hat{\epsilon}_2)$. Because in our case the log likelihood $\log p(\epsilon)(W, A, Y)$ can be written as a sum of a function only of ϵ_1 and a function only of ϵ_2, we can compute the components $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ of the maximum likelihood estimate separately. To get the maximum likelihood estimate $\hat{\epsilon}_1$, we use standard software to fit the logistic regression (7); we enter the expression $\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW$ in (7) as an offset, since we consider $\hat{\beta}$ to be fixed, so that only ϵ_1 can be varied.

We now compute $\hat{\epsilon}_2$. We have from (9) that at $\epsilon_2 = 0$, the derivative of the log likelihood with respect to ϵ_2 is

$$\sum_{i=1}^n C_2(W_i) = \sum_{i=1}^n [\hat{p}_{0Y}(Y = 1|A = 1, W_i) - \hat{\psi}_0] = 0, \qquad (10)$$

where the last equality follows from the property (6) of $\hat{\psi}_0$. Since the log likelihood here is a strictly concave function of ϵ_2, we then have that the maximum of the log likelihood is achieved at $\epsilon_2 = 0$. Thus, the maximum likelihood estimate for ϵ_2 is $\hat{\epsilon}_2 = 0$, and our updated density is then $\hat{p}_1 := p((\hat{\epsilon}_1, 0))$.

In general, we would now replace $\hat{p}_0$ by our update $\hat{p}_1$ and repeat step two of the targeted maximum likelihood algorithm; in this example, however, this is unnecessary, since repeating the above procedure with $\hat{p}_1$ in place of $\hat{p}_0$ would lead to no change in the resulting density estimate. This follows since the clever covariate $C_1(A, W)$ depends only on our current estimate of $p(A|W)$, which is not changed in the above procedure; the clever covariate $C_2(W)$ is updated, but retains the property (10) so that the maximum likelihood estimate $\hat{\epsilon}_2$ would still be 0.

Step 3 of Targeted Maximum Likelihood Algorithm Applied to Estimating the Treatment Specific Mean: Computing the Substitution Estimator

The third and final step in the targeted maximum likelihood algorithm is to compute the substitution estimator for ψ. Recall that the parameter ψ was defined as $\psi(p) := E_p[p(Y = 1|A = 1, W)]$. We will thus compute $\psi(p)$ evaluated at $p = \hat{p}_1$. Under $\hat{p}_1$, the density of Y given A, W, is

$$\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW + \hat{\epsilon}_1 C_1(A, W)), \tag{11}$$

and the marginal density of W is the empirical distribution of W. We then have

$$\begin{aligned} \psi(\hat{p}_1) &= E_{\hat{p}_1}[\hat{p}_1(Y = 1|A = 1, W)] \\ &= E_{\hat{p}_1}\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 W + \hat{\beta}_3 W + \hat{\epsilon}_1 C_1(1, W)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 W_i + \hat{\beta}_3 W_i + \hat{\epsilon}_1 C_1(1, W_i)). \end{aligned} \tag{12}$$

Thus, our final estimator $\psi(\hat{p}_1)$ for ψ is given by (12).

We point out that the derivation of (12) above involves setting A to 1 in each term in the final logistic regression fit

$$\hat{p}_1(Y = 1|A, W) = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW + \hat{\epsilon}_1 C_1(A, W),$$

including setting A to 1 in the clever covariate $C_1(A, W)$.

To summarize the procedure derived above: we first fit the initial logistic regression models (3) and (4); we then update the fit for the logistic regression model (3) by adding the clever covariate as in (7) and refitting the logistic regression with offset; lastly, we compute the sum (12).

Relationship of Above Estimator to that of (Scharfstein et al., 1999) in Important Special Case

Scharfstein et al. (1999) on page 1141 present a class of doubly robust, locally efficient, regression-based estimators for the treatment specific mean, which is the same parameter estimated in this section. To our knowledge, their class of estimators is the first to include the inverse of the propensity score as a term in the regression model used. Scharfstein et al. (1999) state on page 1141, regarding this class of regression-based estimators:

> A straightforward generalization of this estimator solves the long-standing problem in the analysis of treatment effects of how to add the propensity score to a regression model to guarantee consistency, without needing to smooth.

In addition, Scharfstein et al. (1999) point out a useful property of generalized linear models with canonical link functions, when estimating the treatment specific mean; they show that the score under such a model, when the inverse of the propensity score is included as a term in the linear part, includes a component of the efficient influence function for this parameter. They use this property to show double robustness and local efficiency of the corresponding regression-based estimator. This property is leveraged in Step 2 of the targeted maximum likelihood algorithm above. We note that using generalized linear models with canonical link functions is a convenient way to implement targeted maximum likelihood with standard statistical software, but that in general targeted maximum likelihood estimation does not require the use of such working models. We point out that we did not make any claims in our paper that we were the first to present doubly robust, locally efficient, regression-based estimators for treatment effects; to the best of our knowledge, credit for this brilliant result goes to Scharfstein et al. (1999).

The above class of estimators of Scharfstein et al. (1999) is identical to an important special case of applying the targeted maximum likelihood algorithm described above. In particular, this class of estimators of Scharfstein et al. (1999) results if

1. The initial estimator used in Step 1 of the targeted maximum likelihood

algorithm is the fit[3] of a logistic regression model that already contains as one of its terms

$$C_1(A, W) := A/\hat{p}_{0A}(A = 1|W),$$

which we referred to as a "clever covariate"; and,

2. Step 2 of the targeted maximum likelihood algorithm uses the parametric fluctuation as in (7).

In this case, Step 2 of the targeted maximum likelihood algorithm would result in no update to the initial density, since the initial density estimate already contains the covariate $C_1(A, W)$.

We also point out that in (Robins, 2002), the same class of estimators as above is given (before they give an important generalization to time-dependent treatments), and the covariate $1/\hat{p}_{0A}(A = 1|W)$ is referred to there as a "robustifying covariate," on page 1665. The general targeted maximum likelihood algorithm described in Section 2 of our paper does not necessarily involve the use of such covariates, though in many cases (such as those presented in Sections 3 and 4 of our paper), the use of such covariates allows for simple implementation using standard statistical software.

4 Targeted Maximum Likelihood Estimate of the Parameter of a Marginal Structural Model

We show how to apply targeted maximum likelihood methodology to estimate the parameter of a marginal structural model. The previous section gave a special case, corresponding to a saturated marginal structural model. The setting below is a simplified version of that in the analysis described in (Rosenblum et al., 2009). In that analysis, data on each subject was longitudinal, and repeated measures regression was used. We first consider a simpler data generating process, and then discuss a case involving repeated measures in Section 4.2. At the very end of this section we explain how in an important special case our estimator coincides with an estimator of (Scharfstein et al., 1999, Rejoinder to Comments).

[3]When we say the "fit of a regression model," we mean the conditional distribution obtained by estimating the model coefficients with maximum likelihood estimation.

4.1 Estimating Parameter of Marginal Structural Model, where Each Subject Contributes a Single Vector of Observations

Below, we assume our set of data consists of n independent, identically distributed realizations $\{(V_i, W_i, A_i, Y_i)\}_{i=1}^n$ of a random vector (V, W, A, Y), where V, W are baseline variables, A is a treatment with four levels, and Y is a binary outcome. This might occur, for example, in a cross-sectional survey where a vector of data is collected at a single time point for each subject. These variables have the following interpretations: $Y = 1$ represents virologic failure; $A \in \mathcal{A} := \{1, 2, 3, 4\}$ indicates adherence to an antiretroviral regimen in the last month, at levels 0-49%, 50-74%, 75-89%, and 90-100%, respectively; W are baseline potential confounders of the effect of adherence on virologic failure, such as $CD4$ count, past adherence, depression, etc.; V denotes number of consecutive months of past viral suppression since initial suppression was achieved. We restrict attention to the first twelve months since initial suppression, and let $\mathcal{V} := \{1, \ldots, 12\}$ denote the possible values taken by V. For now we assume each subject i contributes just one vector (V_i, W_i, A_i, Y_i) to our data set, and that each subject's data is independent of all other data; in Section 4.2 we extend to the case where subjects contribute multiple such vectors of data.

We are interested in the impact of adherence level A on virologic failure $Y = 1$, within subpopulations defined by number of consecutive months of past suppression V. Thus, we would ideally like to estimate the response curve:

$$r(a, v) := E_p[Y_a \mid V = v], \text{ for } a \in \mathcal{A}, v \in \mathcal{V},$$

where Y_a is the counterfactual response that would have been observed had treatment assignment A been set to a. See e.g. (van der Laan, 2006) for discussion of counterfactual outcomes and their relationship to the parameter considered here. Under the following assumptions, which we make throughout this section, the response curve $r(a, v)$ is identifiable from the distribution of the observed data (V, W, A, Y):

- Time Ordering Assumption: V, W precede A, which precedes Y.
- Consistency Assumption: For all $a \in \mathcal{A}, Y_a = Y$ on the event $A = a$.
- No Unmeasured Confounding Assumption: $\{Y_a\}_{a \in \mathcal{A}} \perp\!\!\!\perp A \mid V, W$.

These assumptions are described in (Robins, 1997; van der Laan, 2006). Under these assumptions, the response curve $r(a, v)$ is identified by

$$r(a, v) = E_p[p(Y = 1 | A = a, V, W) \mid V = v], \text{ for } a \in \mathcal{A}, v \in \mathcal{V}.$$

When A, V have many levels, as is the case here, it may be too ambitious to estimate $r(a, v)$ for all $a \in \mathcal{A}, v \in \mathcal{V}$ directly. Instead, we introduce a working model for $r(a, v)$ defined by

$$m(a, v, \psi) = \text{logit}^{-1}(\psi_0 + \psi_1 a_1 + \psi_2 a_2 + \psi_3 a_3 + \psi_4 v), \tag{13}$$

where a_1, a_2, a_3 are "dummy" indicator variables for the first three levels of adherence, respectively. (That is, for any $a \in \mathcal{A}$, a_1 is the indicator that $a = 1$, a_2 is the indicator that $a = 2$, and a_3 is the indicator that $a = 3$.) We define our parameter of interest ψ to be:

$$\arg\max_{\psi'} \sum_{a \in \mathcal{A}} E_p h(a, V) \log \left[m(a, V, \psi')^{Y_a} (1 - m(a, V, \psi'))^{1 - Y_a} \right], \tag{14}$$

for a given, bounded, measurable, weight function $h(a, V) > 0$. The definition (14) can be considered a maximization of a (weighted) expected log likelihood, under the logistic working model $m(a, V, \psi)$, with weights $h(a, V)$.

There will either be a unique solution or no solution to (14), since the corresponding Hessian matrix is negative definite at all ψ'. In what follows, we assume that there is a solution to (14).[4] Under this assumption, the unique solution ψ^* to (14) is also the unique solution to the estimating equation:

$$\sum_{a \in \mathcal{A}} E_p h(a, V)(Y_a - m(a, V, \psi'))(1, a_1, a_2, a_3, V)' = \mathbf{0}, \tag{15}$$

as well as the unique solution to the estimating equation:

$$\sum_{a \in \mathcal{A}} E_p h(a, V)(p(Y = 1 | A = a, V, W) - m(a, V, \psi'))(1, a_1, a_2, a_3, V)' = \mathbf{0}. \tag{16}$$

The working model $m(a, v, \psi)$ is called a marginal structural model (Robins, 1997). Rather than assume that the working model $m(a, v, \psi)$ is a correctly specified model for $r(a, v)$, we have instead defined ψ nonparametrically by (14). When the working model (13) is correctly specified, ψ is the parameter of the corresponding marginal structural model for the response curve $r(a, v)$; when it is not correctly specified, which in general will be the case, our parameter is still well defined.

Before estimating the parameter ψ with targeted maximum likelihood estimation, we need its efficient influence function, in the nonparametric model

[4] For large enough sample size, violation of this assumption will be detectable with probability tending to 1.

(which is the model we'll be assuming), which is, up to a normalizing matrix:

$$D(p)(Y,A,V,W) := \left[\frac{h(A,V)(Y-p(Y=1|A,V,W))}{p(A|V,W)}(1,A_1,A_2,A_3,V)' + \sum_{a\in\mathcal{A}} h(a,V)\left(p(Y=1|A=a,V,W)-m(a,V,\psi)\right)(1,a_1,a_2,a_3,V)'\right] \quad (17)$$

where A_1, A_2, A_3 are indicator variables of $A = 1, A = 2$, and $A = 3$, respectively. Define the normalizing matrix $M := -E_p \frac{d}{d\psi} D(p)(Y,A,V,W)$. Then the efficient influence function for ψ in the nonparametric model is $M^{-1}D(p)(Y,A,V,W)$. This can be derived, using the general procedure given in (van der Laan and Robins, 2002), and outlined in (van der Laan, 2006). For the working model defined in (13), we have $\frac{d}{d\psi}m(a,v,\psi) = m(a,v,\psi)(1-m(a,v,\psi))(1,a_1,a_2,a_3,v)'$, so that we could replace $(1,a_1,a_2,a_3,v)'$ in (17) by $\frac{d}{d\psi}m(a,v,\psi)/[m(a,v,\psi)(1-m(a,v,\psi))]$.

Step 1 of Targeted Maximum Likelihood Algorithm for Marginal Structural Model: Initial Density Estimate

Step 1 of the targeted maximum likelihood algorithm is to select an initial density estimator. Analogous to our choices in Section 3, we use parametric regression models. However, we note that in general one could use more nonparametric approaches. We use a logistic regression working model for the density of Y given A, V, W:

$$P(Y=1|A,V,W) = \text{logit}^{-1}(\beta_0+\beta_1 A+\beta_2 V+\beta_3 W+\beta_4 AV+\beta_5 AW). \quad (18)$$

This is just one possible choice of terms for the model–any set of terms can be included. We fit the model with maximum likelihood estimation, based on our n independent, identically distributed observations $\{(V_i,W_i,A_i,Y_i)\}_{i=1}^n$, to produce $\hat{\beta}$, and set $\hat{p}_{0Y}(Y|A,V,W)$ to be the density corresponding to the fit model:

$$\hat{p}_{0Y}(Y=1|A,V,W) := \text{logit}^{-1}(\hat{\beta}_0+\hat{\beta}_1 A+\hat{\beta}_2 V+\hat{\beta}_3 W+\hat{\beta}_4 AV+\hat{\beta}_5 AW). \quad (19)$$

We estimate the density of A given V, W using multinomial logistic regression with terms $1, V, W$, and obtain a fit using maximum likelihood estimation, which we denote by $\hat{p}_{0A}(A|V,W)$. We estimate the density of V, W using the empirical distribution, which we denote by $\hat{p}_{0VW}(V,W)$. Our initial estimator

for the density of (V, W, A, Y) is $\hat{p}_0 = \hat{p}_{0Y}(Y|A, V, W)\hat{p}_{0A}(A|V, W)\hat{p}_{0VW}(V, W)$. The substitution estimator $\hat{\psi}_0$ for ψ at this joint density $\hat{p}_0$ then satisfies (by (16))

$$\sum_{a\in\mathcal{A}}\sum_{i=1}^{n} h(a, V_i)(\hat{p}_{0Y}(Y = 1|A = a, V_i, W_i) - m(a, V_i, \hat{\psi}_0))(1, a_1, a_2, a_3, V_i)' = \mathbf{0}. \tag{20}$$

This is a generalization of Equation (6) from Section 3.

Step 2 of Targeted Maximum Likelihood Algorithm for Marginal Structural Model: Constructing and Fitting a Certain Parametric Model

The second step in the targeted maximum likelihood algorithm involves constructing a parametric model $\{p(\epsilon) : \epsilon \in (-\delta, \delta)\}$ that (i) equals the initial density estimate $\hat{p}_0$ at $\epsilon = 0$ and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence function $M^{-1}D(p)(Y, A, V, W)$ of the parameter ψ at $\hat{p}_0$. To construct such a parametric model, we first define clever covariates, which are generalizations of the clever covariates defined in Section 3. Define

$$C_1(A, V, W) := \frac{h(A, V)}{\hat{p}_{0A}(A|V, W)}(1, A_1, A_2, A_3, V)',$$

and $C_2(V, W) :=$

$$\sum_{a\in\mathcal{A}} h(a, V)\left(\hat{p}_{0Y}(Y = 1|A = a, V, W) - m(a, V, \hat{\psi}_0)\right)(1, a_1, a_2, a_3, V)'.$$

We can define a parametric model that only varies the components $p(Y|A, V, W)$ and $p(V, W)$ of $p(Y, A, V, W)$, leaving $p(A|V, W)$ unchanged. For each $\epsilon = (\epsilon_1, \epsilon_2)$, we define the corresponding density in the parametric model

$$p(\epsilon)(Y, A, V, W) := p_Y(\epsilon)(Y|A, V, W)p_A(\epsilon)(A|V, W)p_{V,W}(\epsilon)(V, W),$$

where we set

$$\begin{aligned} p_Y(\epsilon)(Y = 1|A, V, W) &:= \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 V + \hat{\beta}_3 W + \hat{\beta}_4 AV + \hat{\beta}_5 AW \\ &\qquad + \epsilon_1 C_1(A, V, W)), \qquad (21) \\ p_A(\epsilon)(A|V, W) &:= \hat{p}_{0A}(A|V, W), \qquad (22) \\ p_{V,W}(\epsilon)(V, W) &:= s_{\epsilon_2}\exp(\epsilon_2 C_2(V, W))\hat{p}_{0VW}(V, W), \qquad (23) \end{aligned}$$

where the constant $s_{\epsilon_2} := 1/[\frac{1}{n}\sum_{i=1}^{n}\exp(\epsilon_2 C_2(V_i, W_i))]$ is chosen so that for all ϵ, $p_{V,W}(\epsilon)(v, w)$ integrates to 1. It is straightforward to check, just as was done in Section 3, that this parametric model equals $\hat{p}_0$ at $\epsilon = (0, 0)$, and that the score at $\epsilon = (0, 0)$ of this parametric model contains the efficient influence function at $\hat{p}_0$.

The above parametric model is fit using maximum likelihood estimation, resulting in estimates $(\hat{\epsilon}_1, \hat{\epsilon}_2)$. The value of $\hat{\epsilon}_1$ is found by fitting the logistic regression (21), where the expression involving $\hat{\beta}$ is treated as an offset. Since by (20) we have $\sum_{i=1}^{n} C_2(V_i, W_i) = 0$, this implies, by an analogous argument as given in Section 3, that $\hat{\epsilon}_2 = 0$. We then set our updated density to be $\hat{p}_1 := p((\hat{\epsilon}_1, \hat{\epsilon}_2))$.

A single iteration of the above step suffices, since just as was argued in Section 3, the clever covariate $C_1(A, V, W)$ defined above is not changed during the update in that step, and the clever covariate $C_2(V, W)$ will still satisfy $\sum_{i=1}^{n} C_2(V_i, W_i) = 0$.

Step 3 of Targeted Maximum Likelihood Algorithm for Marginal Structural Model: Substitution Estimator

The third step of the targeted maximum likelihood algorithm is to compute the substitution estimator for the parameter at the final density $\hat{p}_1$. That is, we want to compute, based on the property (16) of our parameter, the solution in ψ' to

$$\sum_{a\in\mathcal{A}}\sum_{i=1}^{n} h(a, V_i)(\hat{p}_{1Y}(Y = 1|A = a, V_i, W_i) - m(a, V_i, \psi'))(1, a_1, a_2, a_3, V_i)' = \mathbf{0}. \tag{24}$$

The solution is the substitution estimator $\hat{\psi}$.

Equation (24) can be solved using iteratively reweighted least squares as follows: We first construct a new data set, with outcomes
$\hat{p}_{1Y}(Y = 1|A = a, V_i, W_i)$ for each $a \in \mathcal{A}, i : 1 \leq i \leq n$; these outcomes are regressed on the working model $m(a, V_i, \psi)$ using weighted least squares with weights $h(a, V_i)/[m(a, V_i, \psi)(1 - m(a, V_i, \psi))]$. In the statistical programming language R, this can be done using the generalized linear model **glm** function, with family binomial and logistic link, using $h(a, V_i)$ as weights. If the iteratively reweighted least squares algorithm converges to a value ψ, it necessarily is a solution to (24). Furthermore, if the iteratively reweighted least squares algorithm converges to a value ψ, it is the *unique* solution to (24). This is proved in the Appendix.

In summary, we gave one possible implementation of the targeted maxi-

mum likelihood algorithm to estimate the parameter of a marginal structural model. This involved first computing initial estimators, such as (19), of the components of the density of the data generating distribution. Next, clever covariates were computed and the initial density estimates were updated; this involved simply using maximum likelihood estimation to fit the logistic regression model (21). Lastly, the substitution estimator for ψ at the final density estimate was computed using iteratively reweighted least squares as just described.

4.2 Estimating Parameter of Marginal Structural Model, where Each Subject Contributes Multiple Observations

Above, we assumed data consisted of i.i.d. vectors $\{(V_i, W_i, A_i, Y_i)\}_{i=1}^n$. This corresponds to a situation where each subject i contributes one vector of data (V_i, W_i, A_i, Y_i). We now consider the case in which each subject contributes multiple observations, corresponding to measurements on that subject at monthly intervals. This is similar to the setting in (Rosenblum et al., 2009).

Consider the case in which each subject i contributes 12 time points of data:

$$(V_i(1), W_i(1), A_i(1), Y_i(1), \ldots, V_i(12), W_i(12), A_i(12), Y_i(12)).$$

We denote this vector of observed data by O_i. Here we ignore missing data, though this can be handled as described in (van der Laan, 2010a,b). For each month $t \in \{1, \ldots, 12\}$ since initial viral suppression was achieved, the variables $V(t), W(t), A(t), Y(t)$ have the following interpretations: $Y(t) = 1$ represents virologic failure at end of month t; $A(t) \in \mathcal{A} := \{1, 2, 3, 4\}$ indicates adherence to an antiretroviral regimen during month t, at levels 0-49%, 50-74%, 75-89%, and 90-100%, respectively; $W(t)$ are summaries up through month $t-1$ of potential confounders of the effect of adherence on virologic failure, such as $CD4$ count, past adherence, depression, etc.; $V(t)$ is an indicator of whether continuous viral suppression was achieved during all of months $1, \ldots, t-1$; that is, $V(t) = 1$ if $Y(t') = 0$ for all $t' < t$.

We would like to estimate the causal response curve:

$$r'(a, t) := E(Y_a(t)|V(t) = 1),$$

where $Y_a(t)$ denotes the counterfactual outcome (virologic failure or not) had adherence $A(t)$ been set to a. This is an example of a history-adjusted marginal

structural model (Petersen et al., 2007). Note that we only consider the effect of setting adherence during a single month on the outcome at the end of that month; we do not consider the effect, for example, of setting adherence at multiple time points simultaneously.

Under the following assumptions, which are similar to those given in Section 4.1, the response curve $r'(a,t)$ is identified by

$$r'(a,t) = E_p[p(Y(t)=1|A(t)=a,V(t),W(t)) \mid V(t)=1],$$

for $a \in \mathcal{A}, t \in \{1,\ldots,12\}$. The assumptions are: for all $t \in \{1,\ldots,12\}$,

- Time Ordering Assumption: $V(t), W(t)$ precede $A(t)$, which precedes $Y(t)$.
- Consistency Assumption: For all $a \in \mathcal{A}, Y_a(t) = Y(t)$ on the event $A(t)=a$.
- No Unmeasured Confounding Assumption:

$$\{Y_a(t)\}_{a\in\mathcal{A}} \perp\!\!\!\perp A(t) \mid V(t)=1, W(t).$$

We consider a working model $m'(a,t,\psi)$ for $r'(a,t)$, defined by

$$m'(a,t,\psi) = \text{logit}^{-1}(\psi_0 + \psi_1 a_1 + \psi_2 a_2 + \psi_3 a_3 + \psi_4 t). \tag{25}$$

We define our parameter of interest to be $\psi :=$

$$\arg\max_{\psi'} \sum_{t\in\{1,\ldots,12\}} \sum_{a\in\mathcal{A}} E_p h(a,t)V(t) \log\left[m'(a,t,\psi')^{Y_a(t)}(1-m'(a,t,\psi'))^{1-Y_a(t)}\right], \tag{26}$$

for a given, bounded, measurable, weight function $h(a,t) > 0$. In what follows, we assume that there is a unique solution to (26).

The parameter ψ is a solution of

$$\sum_{t\in\{1,\ldots,12\}} \sum_{a\in\mathcal{A}} E_p h(a,t)V(t)(Y_a(t) - m'(a,t,\psi'))(1,a_1,a_2,a_3,t)' = \mathbf{0}. \tag{27}$$

and under the above assumptions is also a solution of

$$\sum_{t\in\{1,\ldots,12\}} \sum_{a\in\mathcal{A}} E_p h(a,t)V(t) \times[p(Y(t)=1|A(t)=a,V(t)=1,W(t)) - m'(a,t,\psi')](1,a_1,a_2,a_3,t)' = \mathbf{0}. \tag{28}$$

The efficient influence function for this parameter in the nonparametric model (which is the model we'll be assuming), is, up to a normalizing matrix:

$$D'(p)(O) := \sum_{t\in\{1,\ldots,12\}} \left[\frac{h(A(t),t)V(t)(Y(t) - p(Y(t)=1|A(t),V(t),W(t)))}{p(A(t)|V(t),W(t))}(1,A_1(t),A_2(t),A_3(t),t)' + \sum_{a\in\mathcal{A}} h(a,t)V(t)\left(p(Y(t)=1|A(t)=a,V(t),W(t)) - m'(a,t,\psi)\right)(1,a_1,a_2,a_3,t)' \right].$$

where $A_1(t), A_2(t), A_3(t)$ are indicator variables of $A(t) = 1, A(t) = 2$, and $A(t) = 3$, respectively. Define the normalizing matrix $B := -E_p \frac{d}{d\psi} D'(p)(O)$. Then the efficient influence function for ψ in the nonparametric model is $B^{-1}D'(p)(O)$. This can be derived, using the general procedure given in (van der Laan and Robins, 2002), and outlined in (van der Laan, 2006).

Implementation of the targeted maximum likelihood algorithm for the parameter ψ defined in (26) is a generalization of that for the parameter (14) defined earlier. Since by (28) the parameter ψ is a function only of the marginal distributions $p^t(V(t), W(t), A(t), Y(t))$, for each $t \in \{1, \ldots, 12\}$ (and not, for example, the joint distribution of these variables at different values of t), we will estimate just these parts of the overall density of the random vector $O = (V(1), W(1), A(1), Y(1), \ldots, V(12), W(12), A(12), Y(12))$. In fact, we can make a further refinement by only estimating $p^t(Y(t), A(t)|W(t), V(t) = 1)$ and $p^t(W(t), V(t))$, for each $t \in \{1, \ldots, 12\}$, since the parameter ψ only depends upon these parts of the density of the full observation O.

We now define a repeated measures data set that we will use below. It will have $12n$ rows and 5 columns, where n is the number of subjects. For each $t \in \{1, \ldots, 12\}$ we create a separate row of data for each subject i and each month $t \in \{1, \ldots, 12\}$, consisting of the vector $(t, V_i(t), W_i(t), A_i(t), Y_i(t))$, to be used in the steps that follow. We refer to the columns of this data set as (t, V, W, A, Y) below.

The first step of the targeted maximum likelihood algorithm is to get an initial density fit for $p^t(Y(t), A(t)|W(t), V(t) = 1)$ and $p^t(W(t), V(t))$, for each $t \in \{1, \ldots, 12\}$. Though we could obtain the initial density estimate for each $p^t(Y(t), A(t)|W(t), V(t) = 1)$ based only observations at month t, we instead will borrow information across t, assuming a degree of smoothness over t, as we now describe. We fit a logistic regression model for the Y column of our repeated measures data set on the columns t, W, A, using only rows for which $V = 1$. For example we could use the regression model:

$$P(Y = 1|t, W, A, V = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 t + \beta_2 W + \beta_3 A + \beta_4 At + \beta_5 AW). \quad (29)$$

Denote the fit model by $\hat{p}_{0Y}(Y|t, W, A, V = 1)$. We set the initial density estimate for $p^t(Y(t)|W(t), A(t), V(t) = 1)$ to be $\hat{p}_{0Y}(Y|t, W, A, V = 1)$, for each t. We similarly fit a multinomial logistic regression model for the A column of our repeated measures data set on the columns t, W, again using only rows for which $V = 1$. We denote the fit model by $\hat{p}_{0A}(A|t, W, V = 1)$, and set the initial density estimator for $p^t(A(t)|W(t), V(t) = 1)$ to be $\hat{p}_{0A}(A|t, W, V = 1)$, for each t. We use the empirical distribution as initial density estimate for $p^t(W(t), V(t))$. Let $\hat{\psi}_0$ denote the substitution estimator of ψ at these initial density estimates.

Step 2 of the targeted maximum likelihood algorithm involves updating these initial regression fits. We do this exactly as described above in Section 4.1, but now using the repeated measures data set and the following clever covariates:

$$C_1'(t, A, V, W) := V \frac{h(A, t)}{\hat{p}_{0A}(A|t, W, V = 1)}(1, A_1, A_2, A_3, t)',$$

and $C_2'(t, V, W) :=$

$$V \sum_{a \in \mathcal{A}} h(a, t) \left(\hat{p}_{0Y}(Y = 1|t, W, A = a, V = 1) - m'(a, t, \hat{\psi}_0) \right) (1, a_1, a_2, a_3, t)'.$$

The values of these covariates are added on to the repeated measures data set defined above. We construct a parametric model involving these clever covariates, analogous to that defined above in (21), (22), (23), and fit it using the repeated measures data set with maximum likelihood estimation. As in the case above, a single iteration of the above step suffices. Denote the resulting density estimate by $\hat{p}_1$.

The third step of the targeted maximum likelihood algorithm is to compute the substitution estimator for the parameter at the final density $\hat{p}_1$. That is, we want to compute, based on the property of the parameter in (28), the solution in ψ' to

$$\sum_{t\in\{1,\ldots,12\}}\sum_{a\in\mathcal{A}}\sum_{i=1}^{n}h(a,t)V_i(t)$$
$$\times[\hat{p}_1(Y=1|t,A=a,V_i(t)=1,W_i(t))-m'(a,t,\psi')](1,a_1,a_2,a_3,t)'=\mathbf{0}. \quad (30)$$

The solution is the substitution estimator $\hat{\psi}$. Equation (30) can be solved using iteratively reweighted least squares, as described at the end of Section 4.1, but now using the repeated measures data set described earlier, where each subject contributes 12 lines of data.

Construction of confidence intervals can be done with the nonparametric bootstrap, where the unit of sampling is the subject (not the subject-month).

4.3 Relationship of Above Estimator to that of (Scharfstein et al., 1999) in Important Special Case

James M. Robins informed us of another class of highly relevant estimators, presented in the Rejoinder to Comments in (Scharfstein et al., 1999), which coincide with our estimators in Section 4.1 in an important special case. Scharfstein et al. (1999), on page 1142, present a class of regression-based, doubly robust, locally efficient estimators for the parameter of a marginal structural model, which coincides with the parameter ψ defined in (14) of Section 4.1 of our paper. Again, to our knowledge, their class of estimators is the first to include the covariate $C_1(A,V,W)$, which incorporates a weight function and the inverse of the propensity score into a term in the regression working-model used; as Scharfstein et al. (1999) describe, this results in doubly robust, locally efficient estimators for the parameter of a marginal structural model.

The above class of estimators of Scharfstein et al. (1999) is identical to an important special case of applying the targeted maximum likelihood algorithm as described in Section 4.1 of our paper. In particular, this class of estimators of Scharfstein et al. (1999) results if

1. The initial estimator used in Step 1 of the targeted maximum likelihood algorithm is the fit of a logistic regression model that already contains as one of its terms $C_1(A,V,W)$; and,

2. Step 2 of the targeted maximum likelihood algorithm uses the parametric fluctuation as in (21), (22), (23) of our paper.

In this case, Step 2 of the targeted maximum likelihood algorithm would result in no update to the initial density, since the initial density estimate already

contains the covariate $C_1(A, V, W)$. We note that the targeted maximum likelihood algorithm is not restricted to using parametric regression estimators as initial estimators, and, in other work, Mark van der Laan has focused on more data-adaptive estimators (e.g. in Section 8 of van der Laan and Rubin (2006)).

Targeted maximum likelihood estimation is a general algorithm for constructing estimators in semiparametric and nonparametric models. As stated in Section 2 of this paper, it involves the following three steps:

> (i) constructing an initial estimate $\hat{p}_0$ of (a relevant part of) the density of the data generating distribution, (ii) using the efficient influence function of the parameter to find a better fit $\hat{p}_1$ targeted at minimizing mean squared error for estimation of the parameter ψ, and (iii) computing the substitution estimator $\psi(\hat{p}_1)$ at this estimated (relevant part of) density. In general, step (ii) is iterated until convergence (defined below), though in many examples a single iteration suffices.

In Section 2 of our article, we elaborated on the second step above, which involves

> constructing a parametric model $\{p(\epsilon) : \epsilon \in (-\delta, \delta)\}$ in the overall model $\mathcal{M}$ that (i) equals the initial density estimate $\hat{p}_0$ at $\epsilon = 0$ and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence function of the parameter ψ at $\hat{p}_0$. We then use maximum likelihood estimation in the parametric model $\{p(\epsilon) : \epsilon \in (-\delta, \delta)\}$ to get an estimate $\hat{\epsilon}$ for ϵ. Our updated density estimate is then set to be $\hat{p}_1 := p(\hat{\epsilon})$.

Under weak regularity conditions, targeted maximum likelihood estimation can be applied to arbitrary data structures, semiparametric models, and pathwise differentiable parameters (van der Laan and Rubin, 2006; van der Laan et al., 2009). Generalizations to parameters that are not pathwise differentiable are given in Sections 10-12 of (van der Laan and Rubin, 2006). Advantages of targeted maximum likelihood estimators are described in Section 2 of our paper, and in (van der Laan et al., 2009).

We also point out that the class of estimators of the effects of multiple time point interventions defined in the papers (Robins, 2000; Bang and Robins, 2005) is not a class of targeted maximum likelihood estimators (since, for example, they are not in general substitution estimators, as all targeted max-

imum likelihood estimators are).[5] Neither are the time-dependent estimators in Section 3.2 of Robins (2002), for the same reason. However, Robins (2000, 2002) and Bang and Robins (2005) made fundamental contributions, two of which we describe below.

1. Robins (2000, 2002) and Bang and Robins (2005) present important extensions of the regression-based estimators described above from (Scharfstein et al., 1999), to time-dependent censoring and/or treatments. In particular, the regression-based estimators in (Robins, 2000, 2002; Bang and Robins, 2005) are doubly robust, locally efficient, and include time-dependent generalizations of the inverse of the propensity score as terms in the regression models used.

2. In Section 2 of (Robins, 2000), the estimation algorithm involves an initial fit for a regression model for an estimating equation, followed by adding the inverse of the propensity score as a covariate, and then refitting the model while holding constant the previously fit coefficients. This procedure is used in Step 2 of both Sections 3 and 4 of our paper, but with different covariates than in (Robins, 2000); as we explained above, the class of estimators defined in (Robins, 2000) is not a class of targeted maximum likelihood estimators.

We also briefly describe newly published work relevant to our article. Gruber and van der Laan (2010) give a class of targeted maximum likelihood estimators that extend the class given in our article in Section 3 to handle continuous, bounded outcomes, in a way that avoids the poor finite sample performance described in (Robins et al., 2007) of some possible targeted maximum likelihood estimators.

5 Discussion

Targeted maximum likelihood is a versatile estimation tool, extending some of the advantages of maximum likelihood estimation for parametric models to semiparametric and nonparametric models. In many problems it leads to doubly robust, locally efficient estimators. This is the case for the estimators we give above, under regularity conditions. Double robustness, in the scenarios we considered, means that whenever at least one of the initial estimators $\hat{p}_{0Y}$

[5] However, it is possible that there may be subclasses of these estimators of effects of multiple time point interventions that are examples of targeted maximum likelihood estimators. This is an open question and topic of further research.

and $\hat{p}_{0A}$ is correctly specified, the targeted maximum likelihood estimator is consistent. In contrast, standard propensity score methods and inverse probability weighting methods are generally not doubly robust, since they require the model for the density of A given baseline variables be correctly specified in order to be consistent.

The double robustness property of the targeted maximum likelihood estimator results, in part, from its being an approximate solution to the efficient influence function estimating equation (when such an estimating function in the parameter of interest exists). That is, the targeted maximum likelihood estimator has the property that if we write the efficient influence function as a function f of the parameter of interest $\psi(p)$, nuisance parameters $\eta(p)$, and an observation $O = (W, A, Y)$, then for $\hat{p}$ the final density output by the targeted maximum likelihood algorithm, we have that $\sum_{i=1}^{n} f(\psi(\hat{p}), \eta(\hat{p}), (W_i, A_i, Y_i)) \approx 0$. Important extensions of targeted maximum likelihood estimation to more general loss functions are developed in Appendix B of (van der Laan et al., 2009), and extensions to "collaborative" estimation of nuisance parameters are developed in (van der Laan and Gruber, 2009).

6 Appendix

Here we prove the claim made at the end of Section 4, that if the iteratively reweighted least squares algorithm described at the end of Section 4.1 converges to a value ψ, it is the unique solution to (24). Here, as above, we assume that V is univariate.

It is straightforward to show that if the iteratively reweighted least squares algorithm described at the end of Section 4 converges to a value ψ, then it solves (24). The task here is to show that if a solution to (24) exists, then it is unique.

Let $\hat{p}_1$ denote the final density output by the targeted maximum likelihood algorithm. The following expression has at most one point where its derivative is the vector $\mathbf{0}$, since its Hessian matrix is negative definite at all ψ':

$$\sum_{a \in \mathcal{A}} E_{\hat{p}_1} h(a, V) \log \left[m(a, V, \psi')^{\hat{p}_1(Y=1|A=a,V,W)} (1 - m(a, V, \psi'))^{1-\hat{p}_1(Y=1|A=a,V,W)} \right]. \tag{31}$$

The derivative with respect to ψ' of the previous display equals

$$\sum_{a \in \mathcal{A}} E_{\hat{p}_1} h(a, V) (\hat{p}_1(Y = 1|A = a, V, W) - m(a, V, \psi'))(1, a_1, a_2, a_3, V)'. \tag{32}$$

Thus, there can be at most one value of ψ' for which the expression (32) equals $\mathbf{0}$. The expression (32) equals $1/n$ times the expression on the left hand side of (24), since the marginal distribution of V, W under $\hat{p}_1$ is the empirical distribution of V, W. Thus, equation (24) can have at most one solution, completing the proof.

References

Bang, H. and J. M. Robins (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics 61*, 692–972.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: The Johns Hopkins University Press. Springer-Verlag.

Breiman, L., J. H. Friedman, R. Olshen, and C. J. Stone (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group.

Gruber, S. and M. J. van der Laan (2010, May). A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 265. http://www.bepress.com/ucbbiostat/paper265*.

Moore, K. L. and M. J. van der Laan (2007, April). Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 215. http://www.bepress.com/ucbbiostat/paper215*.

Neugebauer, R. and M. J. van der Laan (2002). Why Prefer Double Robust Estimates? Illustration with Causal Point Treatment Studies. Working Paper 115. http://www.bepress.com/ucbbiostat/paper115. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Petersen, M., S. Deeks, J. Martin, and M. van der Laan (2007). History-Adjusted Marginal Structural Models to Estimate Time-Varying Effect Modification. *American Journal of Epidemiology 166*(9), 985–993.

Polley, E. and M. van der Laan (2009). "Selecting Optimal Treatments Based on Predictive Factors". In K. E. Peace (Ed.), *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*, pp. 441–454. Boca Raton: Chapman and Hall/CRC.

Robins, J. and A. Rotnitzky (1992). Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In N. Jewell, K. Dietz, and V. Farewell (Eds.), *AIDS Epidemiology - Methodological Issues*. Boston, MA: Bikhäuser.

Robins, J., A. Rotnitzky, and L. Zhao (1994, September). Estimation of Regression Coefficients When Some Regressors are Not Always Observed. *Journal of the American Statistical Association 89*(427), 846–866.

Robins, J. M. (1997). Marginal Structural Models. *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, 1–10.

Robins, J. M. (2000). Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6–10.

Robins, J. M. (2002). Commentary on Using Inverse Weighting and Predictive Inference to Estimate the Effects of Time-Varying Treatments on the Discrete-time Hazard. *Statistics in Medicine 21*, 1663–1680.

Robins, J. M. and A. Rotnitzky (2001). Comment on the Bickel and Kwon article, "Inference for Semiparametric Models: Some Questions and an Answer". *Statistica Sinica 11*(4), 920–936.

Robins, J. M., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of Double-Robust Estimators when "Inverse Probability" Weights are Highly Variable. *Statistical Science 22*(4), 544–559.

Rosenblum, M., S. G. Deeks, M. van der Laan, and D. R. Bangsberg (2009, September). The Risk of Virologic Failure Decreases with Duration of HIV Suppression, at Greater than 50% Adherence to Antiretroviral Therapy. *PLoS ONE 4*(9), e7196.

Rosenblum, M. and M. van der Laan (2010). Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *The International Journal of Biostatistics. Article 13. DOI: 10.2202/1557-4679.1138 Available at: http://www.bepress.com/ijb/vol6/iss1/13 6*(1).

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for Non-Ignorable Drop-out Using Semiparametric Nonresponse Models, (with Discussion and Rejoinder). *Journal of the American Statistical Association 94*, 1096–1120 (1121–1146).

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

van der Laan, M. J. (2006). Statistical Inference for Variable Importance. *The International Journal of Biostatistics. DOI: 10.2202/1557-4679.1008. Available at: http://www.bepress.com/ijb/vol2/iss1/2. Issue 1 2*(2).

van der Laan, M. J. (2010a). Targeted Maximum Likelihood Based Causal Inference: Part I. *The International Journal of Biostatistics. Article 2. DOI: 10.2202/1557-4679.1211 Available at: http://www.bepress.com/ijb/vol6/iss2/2 6*(2).

van der Laan, M. J. (2010b). Targeted Maximum Likelihood Based Causal Inference: Part II. *The International Journal of Biostatistics. Article 3. DOI: 10.2202/1557-4679.1241 Available at: http://www.bepress.com/ijb/vol6/iss2/3 6*(2).

van der Laan, M. J. and S. Gruber (2009, April). Collaborative Double Robust Targeted Penalized Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 246. http://www.bepress.com/ucbbiostat/paper246*.

van der Laan, M. J. and J. M. Robins (2002). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

van der Laan, M. J., S. Rose, and S. Gruber (2009). Readings in Targeted Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 254. http://www.bepress.com/ucbbiostat/paper254*.

van der Laan, M. J. and D. Rubin (2006, October). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics 2*(1).

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.