An Inverse Probability Weighting Approach for Instrumental Variable Estimation of Marginal Structural Mean Models

Haben Michael
Yifan Cui
Eric J. Tchetgen Tchetgen
Department of Statistics
The Wharton School
University of Pennsylvania

Abstract

Robins (1998) introduced marginal structural models (MSMs), a general class of counterfactual models for the joint effects of time-varying treatment regimes in complex longitudinal studies subject to time-varying confounding. He established identification of MSM parameters under a sequential randomization assumption (SRA), which rules out unmeasured confounding of treatment assignment over time. We consider sufficient conditions for identification of the parameters of a subclass, Marginal Structural Mean Models (MSMMs), when sequential randomization fails to hold due to unmeasured confounding, using instead a time-varying instrumental variable. Our identification conditions require that no unobserved confounder predicts compliance type for the time-varying treatment, the longitudinal generalization of the identifying condition of Wang and Tchetgen Tchetgen (2018). We describe a simple weighted estimator and examine its finite-sample properties in a simulation study.

KEY WORDS: marginal structural models, time-varying endogeneity, instrumental variables

1 Introduction

Robins ([10, 11, 12]) introduced marginal structural models (MSMs), a class of counterfactual models that encode the joint causal effects of time-varying treatment in the presence of time-varying confounding. For identification, Robins relied on a sequential randomization assumption (SRA), which rules out unmeasured confounding of the time-varying treatment. MSMs have since become the standard analytic approach to evaluate causal effects in time-varying epidemiological studies [6, 5] . However, SRA may be hard to justify in many such settings, and unmeasured confounding bias may invalidate causal claims inferred by the approach. In the case of a point treatment, a large literature in causal inference has developed over the years on the instrumental variable method aiming to address unmeasured confounding [2, 3, 1]. Instead of assuming that there is no unmeasured confounding, the IV approach relies on the key assumption that one has observed a pretreatment variable that can affect the outcome only through its effects on the treatment. IV methods for time-varying settings are far less developed. In this paper, we consider sufficient conditions for identification the parameters of a Marginal Structural Mean Model (MSMM) with the aid of a time-varying instrumental variable when sequential randomization fails to hold due to unmeasured confounding. In doing so, we firmly establish the IV approach in the context of MSMs for complex longitudinal settings, an extension previously believed out of reach. Our identification conditions require longitudinal generalizations of (i) IV relevance, (ii) exclusion restriction, and (iii) IV independence assumptions, together with a key assumption (iv) that no unobserved confounder predicts compliance type for the time-varying treatment, a longitudinal generalization of the identification condition of [14]. Under these assumptions, we establish identification of the MSMM parameter and propose a simple estimation procedure analogous to inverse-probability weighted (IPW) estimation, the most common approach for estimating MSMs under SRA.

Prior to the current work, Robins [9] developed a general framework for identification and estimation of causal effects of time-varying endogenous treatments using a time-varying instrumental variable under a structural nested mean model (SNMM). As described in [12], the parameters of an SNMM can under certain homogeneity conditions be interpreted as MSMM parameters, in which case [9] provides alternative estimators to ours. In contrast, the proposed methodology is

more general as it directly targets MSMM parameters irrespective of whether or not they can be interpreted as parameters of an equivalent SNMM.

The remainder of the paper is organized as follows. In Section 2 we describe identification and estimation of MSMM parameters under SRA. In Section 3 we present an alternative set of identification conditions to SRA, making use of a time-varying instrumental variable. In Section 4 we describe a simple weighted estimator for MSMM parameters using our instrumental variable approach. In Section 5 we describe methods for sampling data under a MSMM and apply this method to examine the finite-sample performance of our proposed estimator. We conclude in Section 6 with an assessment and description of future work.

2 Problem Formulation

We consider longitudinal data, some of which is fully observed and some only partially observed. There are $J \leq \infty$ discrete times, which we identify as $0, 1, 2, \ldots$ to simplify notation. Given a time j and time-varying variable X, let an overbar denote the history of the process up to and including j, e.g., $\overline{X}(j) = (X(0), \ldots, X(j))$. The observed data consist of iid vectors $\overline{L}(J)$ interpreted as J vectors of covariates and $\overline{A}(J)$ interpreted as J binary treatments. A variable $Y \in L(J)$ is singled out as an outcome of interest, and $V \in L(0)$ as baseline covariates. $L(j)$ is assumed to precede $A(j)$. The statistical significance of these temporal relations are conditional independence relationships implied by a temporally ordered DAG such as given in Fig. 1, in which a node is independent of non-descendants conditional on its parent nodes; see [7] for details. The partially observed data consist of 2^{J+1} variables $Y_{\overline{a}}$ indexed by $\overline{a} \in \{0,1\}^{J+1}$. The observed and partially observed data are related by the "consistency" assumption [10],

$$Y = Y_{\overline{A}(J)} \equiv Y_{\overline{a}}\big|_{\overline{a}=\overline{A}(J)} \text{ a.s..} \tag{1}$$

Thus $\overline{a}$ has the interpretation of a particular nonrandom treatment regime, and $Y_{\overline{a}}$ as a potential outcome, i.e., the distribution of Y were everyone in the population to follow treatment regime $\overline{a}$.

A marginal structural mean model [10] is a model on the marginal mean of potential outcomes, possibly conditional on baseline covariates V. Consider the following MSMM that specifies the joint effect of treatment as linear in cumulative treatment history:

$$\mathbb{E}(Y_{\overline{a}}) = \beta_0 + \beta_1 \sum_{j=0}^{J-1} a_j, \tag{2}$$

with $\beta \in \mathbb{R}^2$ parameterizing the model and encoding the incremental effect of a unit of treatment. A link function can be introduced to accommodate binary and count outcome variables, i.e., (2) generalizes to

$$\mathbb{E}(Y_{\overline{a}}) = \mu_\beta(\overline{a}) \tag{3}$$

where $\mu_\beta(\cdot)$ is a known function that can incorprate an appropriate link function, e.g., logit or log link, and β is an unknown parameter of interest.

Since the model is defined using partially unobserved data, this parameter is not in general identified by the observed data. Robins [10] provides the standard sufficient condition for identification, the sequential randomization assumption

$$Y_{\overline{a}}(j) \amalg A(j) \mid \overline{L}(j), \overline{A}(j-1) \qquad 0 \leq j < J, \quad \overline{a} \in \{0,1\}^{J+1}, \qquad \text{(SRA)} \tag{4}$$

using $\amalg$ to denote statistical independence. SRA will hold if all common causes of Y and $A(j)$ are included in $(\overline{L}(j), \overline{A}(j-1))$. Assuming also that the "positivity" assumptions holds, $\mathbb{P}(\overline{A}(j-1) = \overline{a}_{j-1}, \overline{L}(j) = \overline{l}_j) > 0$ implies $\mathbb{P}(A(j) = a_j \mid \overline{A}(j-1) = \overline{a}_{j-1}, \overline{L}(j) = \overline{l}_j) > 0$ for any $\overline{a}_j, \overline{l}_j$, Robins identifies MSMM parameters using inverse-probability-of-treatment weights. Specifically, let f denote the conditional density of A_j given prior data $(\overline{A}_{j-1}, \overline{L}_{j-1})$ and let f^* denote any density of A_j given prior treatment data A_{j-1}. f^* need not be the true data-generating mechanism. Define time-varying weights as their ratio,

$$W_{SRA}(j) = \frac{f(A_j \mid \overline{A}_{j-1}, \overline{L}_{j-1})}{f^*(A_j \mid \overline{A}_{j-1})}, \quad j = 0, \ldots, J-1.$$

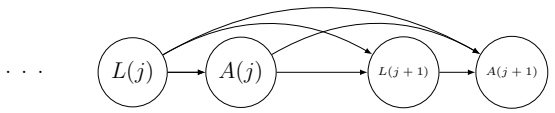


Figure 1: Causal DAG describing longitudinal confounding.

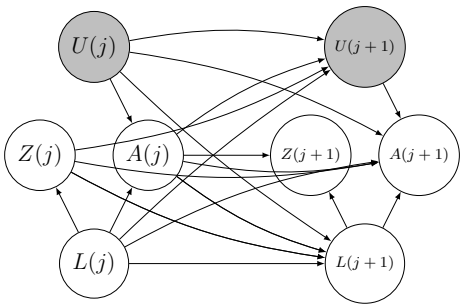


Figure 2: Causal DAG describing confounding with unobserved confounders and IV.

Then under SRA and positivity Robins showed that β in (2) is the solution to the following weighted estimating equation:

$$0 = \mathbb{E}\left(\begin{pmatrix} 1 \\ \sum_j A_j \end{pmatrix} \frac{Y - \beta_0 - \beta_1 \sum_j A_j}{\overline{W}_{SRA}(J-1)}\right).$$

More generally, given an MSMM (3), let h be a vector-valued function of $\overline{A}$ of dimension equal to that of β. Then under SRA and positivity, β is a solution to

$$0 = \mathbb{E}\left(h(\overline{A})\frac{Y - \mu_\beta(\overline{A})}{\overline{W}_{SRA}(J-1)}\right).$$

In Lemma 1 below we establish an analogous result using IV assumptions in lieu of SRA.

3 Identification of causal model parameters

Suppose that there is some additional variable $U \notin L$ associated with $\overline{A}$ and $\overline{L}$ (including any $Y \in L(J)$) beyond what is observed in $\overline{L}$, so that the sequential randomization assumption (4) is not warranted. On the other hand, suppose there is observed some time-varying variable Z associated with $\overline{A}$ that, given observed data, is unassociated with the potential outcomes. We propose to use such an "instrumental variable" to supply sufficient conditions to identify an MSM parameter in the absence of SRA.

Specifically, in addition to the data described in Section 2, suppose a binary-valued process Z is also observed and let U be an unobserved process. See Fig. 2. We make the following assumptions

for all $0 \leq j \leq J-1$ and $\overline{a}, \overline{z} \in \{0,1\}^J$:

Assumption 1 (Latent SRA) $\overline{Y}_{\overline{a}} \amalg A(j) \mid \overline{A}(j-1) = \overline{a}(j-1), \overline{L}(j), \overline{U}(j), \overline{Z}(j)$

Assumption 2 (IV Relevance) $Z(j) \not\amalg A(j) \mid \overline{A}(j-1), \overline{L}(j), \overline{Z}\,(j-1)$

Assumption 3 (Exclusion Restriction) $Y_{\overline{a}\overline{z}} = Y_{\overline{a}}$

Assumption 4 (IV Independence) $\left(\overline{U}, \overline{Y}_{\overline{a}}\right) \amalg Z(j) | \overline{A}(j-1) = \overline{a}(j-1), \overline{L}(j), \overline{Z}(j-1)$

Assumption 5 (IV Positivity) $0 < \mathbb{P}(Z(j) = 1 | \overline{A}(j-1), \overline{L}(j), \overline{Z}(j-1)) < 1$

Assumption 6 (Independent Compliance Type)

$$\mathbb{E}\left[A(j) | \overline{U}(j), \overline{L}(j), \overline{A}\,(j-1)\,, \overline{Z}(j-1), Z(j) = 1\right] - \mathbb{E}\left[A(j) | \overline{U}(j), \overline{L}(j), \overline{A}\,(j-1)\,, \overline{Z}(j-1), Z(j) = 0\right]$$
$$= \delta_j\left(\overline{L}(j), \overline{A}\,(j-1)\,, \overline{Z}\,(j-1)\right)$$

Assumption 1 is a variant of SRA, requiring SRA hold were U included among the observables. Assumptions 2–5 are longitudinal generalizations of standard IV assumptions. Note that Assumption 1 is more easily satisfied with a larger set U, whereas Assumption 4 is more easily satisfied with smaller U. Assumption 6, interpreted as asserting that compliance type [4] is conditionally independent of the unknown confounder, is not standard in the SRA or IV literature. It is a longitudinal generalization of an assumption used in the point exposure setting of our problem [14].

Let f_Z denote the density of $Z(j)$ conditional on the prior observed history and let f^* be any given density of the treatment process $A(\cdot)$. We define the following modified time varying weights:

$$\overline{W}\,(j) = \prod_{j=1}^{J-1} W_{j,1} W_{j,2}$$
$$\overline{W} = \overline{W}\,(J-1)$$

where

$$W_{j,1} = \frac{f_Z\left(Z(j)|\overline{L}(j), \overline{A}(j-1), \overline{Z}(j-1)\right)\delta_j\left(\overline{L}(j), \overline{A}\left(j-1\right), \overline{Z}\left(j-1\right)\right)}{(-1)^{1-Z(j)}} \tag{5}$$

and

$$W_{j,2} = \frac{1}{(-1)^{1-A(j)} f^*\left(A(j)|V, \overline{A}(j-1)\right)}$$

Lemma 1 *Suppose that together with consistency (1), Assumptions 1–6 hold. Given an MSMM* $\mathbb{E}(Y_{\overline{a}}) = \mu_\beta(\overline{a})$ *let* h *be a vector-valued function of* $\overline{a}$ *of the same dimension as* β*, and let* f^* *denote an arbitrary joint density for the treatment process* $A(\cdot)$*. Then,*

$$\mathbb{E}\left(h(\overline{A})(Y - \mu_\beta(A))/\overline{W}\right) = \sum_{\overline{a}} h(\overline{a})\left(\mathbb{E}(Y_{\overline{a}}) - \mu_\beta(\overline{a})\right)\prod_{j=0}^{J-1} f^*\left(a_j|\overline{a}_{j-1}\right),$$

where the summation is taken over all tuples $\overline{a} \in \{0,1\}^{J-1}$*.*

Proof.

$$\begin{aligned}\mathbb{E}(Y/\overline{W}) &= \sum_{\overline{a}} \mathbb{E}(Y_{\overline{a}}\{\overline{A} = \overline{a}\}/\overline{W}) \\ &= \sum_{\overline{a}} f^*(\overline{a}) \prod_{j=0}^{J-1}(-1)^{1-a_j}\mathbb{E}(Y_{\overline{a}}\{\overline{A} = \overline{a}\}/\overline{W}_1)\end{aligned}$$

By an application of Assumption (1) the inner expectation is

$$\begin{aligned}\mathbb{E}(Y_{\overline{a}}\{\overline{A} = \overline{a}\}/\overline{W}_1) &= \mathbb{E}(\mathbb{E}(Y_{\overline{a}} \mid \overline{AZLU}_{J-1})\{\overline{A} = \overline{a}\}/\overline{W}_1) \\ &= \mathbb{E}(\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{J-2}, \overline{LU}_{J-1})\{\overline{A} = \overline{a}\}/\overline{W}_1)\end{aligned}$$

and we proceed by induction. Suppose for $1 \leq j \leq J-1$

$$\mathbb{E}(Y_{\overline{a}}\{\overline{A} = \overline{a}\}/\overline{W}_1) = \prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}(\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_j = \overline{a}_j\}/\overline{W}_1(j)).$$

The RHS is

$$
\begin{aligned}
&\prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}(\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_j = \overline{a}_j\}/\overline{W}_1(j)) \\
&= \prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}\left(\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_{j-1} = \overline{a}_{j-1}\}\mathbb{P}(A_j = a_j \mid \overline{LUZ}_j, A_{j-2})/\overline{W}_1(j)\right) \\
&= \prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}\left(\frac{\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_{j-1} = \overline{a}_{j-1}\}(-1)^{Z_j}\mathbb{P}(A_j = a_j \mid \overline{LUZ}_j, A_{j-2})}{\overline{W}_1(j-1)\Delta_j f_Z(Z_j \mid \overline{A}_{j-1}, \overline{Z}_{j-1}, \overline{L}_j)}\right) \\
&= \prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}\left(\frac{\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_{j-1} = \overline{a}_{j-1}\}}{\overline{W}_1(j-1)\Delta_j}\times\right. \\
&\qquad \left.\sum_{z_j\in\{0,1\}}\frac{(-1)^{z_j}\mathbb{P}(A_j = a_j \mid \overline{LU}_j, z_j, A_{j-2})\mathbb{P}(z_j \mid \overline{A}_{j-1}, \overline{Z}_{j-1}\overline{L}_j, \overline{U}_j)}{f_Z(z_j \mid \overline{A}_{j-1}, \overline{Z}_{j-1}\overline{L}_j)}\right) \\
&= \prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}\left(\frac{\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_{j-1} = \overline{a}_{j-1}\}}{\overline{W}_1(j-1)\Delta_j}\sum_{z_j\in\{0,1\}}(-1)^{z_j}\mathbb{P}(A_j = a_j \mid \overline{LU}_j, z_j, A_{j-2})\right) \\
&= \prod_{k=j+1}^{J-1}(-1)^{1-a_k}\mathbb{E}\left(\frac{\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_{j-1} = \overline{a}_{j-1}\}}{\overline{W}_1(j-1)\Delta_j}(-1)^{1-a_j}\Delta_j\right) \\
&= \prod_{k=j}^{J-1}(-1)^{1-a_k}\mathbb{E}\left(\mathbb{E}(Y_{\overline{a}} \mid \overline{AZ}_{j-1}, \overline{LU}_j)\{\overline{A}_{j-1} = \overline{a}_{j-1}\}/\overline{W}_1(j-1)\right).
\end{aligned}
$$

The third-to-last equality follow from an application of Assumption (4) and the second-to-last by (6). Thus

$$\mathbb{E}(Y_{\overline{a}}\{\overline{A} = \overline{a}\}/\overline{W}_1) = \prod_{k=1}^{J-1}(-1)^{1-a_k}\mathbb{E}(Y_{\overline{a}}),$$

as required. ■ As the range of $\overline{W}$ includes negative values, $\overline{W}$ are not weights in the usual sense, and we only refer to them as such in analogy with the weights $\overline{W}_{SRA}$ defined in Section 2.

4 Estimation and Inference

Lemma 1 suggests an estimator for an MSMM parameter (3). Provided the assumptions of the lemma hold, $\mathbb{E}\left\{(Y - \mu_\beta(\overline{A}))/\overline{W})\right\} = 0$, and

$$o_P(n^{-1/2}) = n^{-1} \sum_k h(\overline{A}_k)(Y_k - \mu_\beta(\overline{A}_k))/\overline{W}_k$$

may serve as an estimating equation for β, using k to index an iid sample of size n. Specifically, provided also that the usual regularity conditions for M-estimation hold, the solution $\hat{\beta}$ is asymptotically linear with

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\mathbb{E}\left\{h(\overline{A})\frac{\partial}{\partial\beta}(\mu_\beta(\overline{A}))/\overline{W}\right\}\right)^{-1} n^{-1/2} \sum_k h(\overline{A}_k)(Y_k - \mu_\beta(\overline{A}_k))/\overline{W}_k + o_P(1). \quad (6)$$

As the weights will usually not be known, suppose we have $n^{1/2}$-consistent estimators for the terms $\delta(j)$ and the conditional density f_Z for Z given the observed earlier variables, and let $\widehat{W}(j)$ denote the weights formed by plugging these estimators into (5), $0 \leq j \leq J-1$. Then

$$o_P(n^{-1/2}) = n^{-1} \sum_k h(\overline{A}_k)(Y_k - \mu_\beta(\overline{A}_k))/\widehat{\overline{W}}_k$$

is an estimating equation for β. Inference may be carried out with the nonparametric bootstrap or the "sandwich" asymptotic variance estimator using (6)

5 Simulation

In order to simulate data under an MSMM $\mu_\beta(\overline{a})$ we use a variant of the G-formula [8] to relate the MSMM to the non-counterfactual data,

$$\mu_\beta(\overline{a}) = \mathbb{E}(Y_{\overline{a}}) = \int \mathbb{E}(Y \mid \overline{A}, \overline{L}, \overline{U}) \prod_j f(l_j, u_j \mid \overline{a_{j-1}}\overline{l_{j-1}}\overline{u_{j-1}}) d\nu(l_j, u_j),$$

using f to refer to the conditional densities with respect to some dominating measure ν of (L_j, U_j) on past treatment and covariates . This formula follows from the "Latent SRA" assumption (1) by a straightforward induction; alternatively, interpret $\overline{U}$ along with $\overline{L}$ as observed variables and apply the standard G-formula of [8], the Latent SRA assumption supplying the needed sequential randomization assumption under this interpretation.

For example, we may sample outcomes as

$$\mathbb{E}(Y \mid \overline{A}, \overline{L}, \overline{U}) = \sum_j \left(f_j(L_j, U_j) - \mathbb{E}(f_j(L_j, U_j) \mid \overline{L}_{j-1}, \overline{U}_{j-1}, \overline{A}_{j-1}) \right) + \mu_\beta(\overline{A}), \tag{7}$$

once we have decided on a sampling scheme for $\overline{L}, \overline{U}, \overline{A}$, and specified functions f_j.

We use this method to examine the finite-sample behavior of the simple weighted estimator described in Section 4 under a particular data-generation process in which SRA does not hold. For all $0 \leq j \leq J-1$, $U(j)$ is sampled as standard normal and $Z(j)$ is bernoulli with success probability 1/2, all mutually independent. In order to enforce the independent compliance type requirement (6), the treatments $A(j)$ and covariates $L(j)$ are sampled as:

$$\begin{aligned}
\mathbb{E}(L(j)) &= \lambda_0 + \lambda_1 A(j-1) \\
\Phi^{-1}(\delta(j+1)) &\equiv \Phi^{-1}[\mathbb{P}(A(j+1) = 1 \mid \overline{L}(j+1), \overline{U}(j+1), \overline{A}(j), \overline{Z}(j), Z(j+1) = 1) \\
&\quad - \mathbb{P}(A(j+1) = 1 \mid \overline{L}(j+1), \overline{U}(j+1), \overline{A}(j), \overline{Z}(j), Z(j+1) = 0)] \\
&= \gamma_0 + \gamma_1 L(j+1) \\
\mathbb{P}(A(j+1) = 1 \mid \overline{LUZ}(j+1), \overline{A}(j)) &= \Phi(\alpha_0 + \alpha_1 L(j+1) + \alpha_2 U(j+1)) \times (1 - \delta(j+1)) \\
&\quad + Z(j+1) \times \delta(j+1).
\end{aligned} \tag{8}$$

Φ denotes the standard normal CDF and $\lambda_0, \lambda_1, \gamma_0, \gamma_1, \alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$, control, among other things, the strength of the IV and degree of confounding. We take $\lambda_0 = \lambda_1 = .5, \gamma_0 = \gamma_1 = .3, \alpha_0 = -.2$, and $\alpha_1 = \alpha_2 = .2$ in the simulation described below.

The MSMM is

$$\mathbb{E}(Y_{\overline{a}}) = \beta_0 + \beta_1 \sum_j a_j.$$

Outcome variables Y are generated as

$$Y = \sum_{j=0}^{J} (\tau_j L_j + \rho_j U_j) + \mathbb{E}(Y_{\overline{a}}) + \epsilon = \sum_{j=0}^{J} \rho_j U_j + \beta_0 + \beta_1 \sum_j a_j + \epsilon,$$

with $\rho_j, \tau_j \in \mathbb{R}$ and ϵ standard normal. We take $\rho_j = \tau_j = 1$ in our simulation.

The parameter of interest, β, is estimated using OLS regression on the treatments $\overline{a}$ weighted as described in Section 4. That is, for a sample size of n, letting D_n be the design matrix containing the observed regressors, Y_n the vector of observed outcomes, and W_n the diagonal matrix formed from the inverted weights, we estimate β as

$$\hat{\beta} = (D_n^T W_n D_n^T)^{-1} D_n^T (W_n Y_n).$$

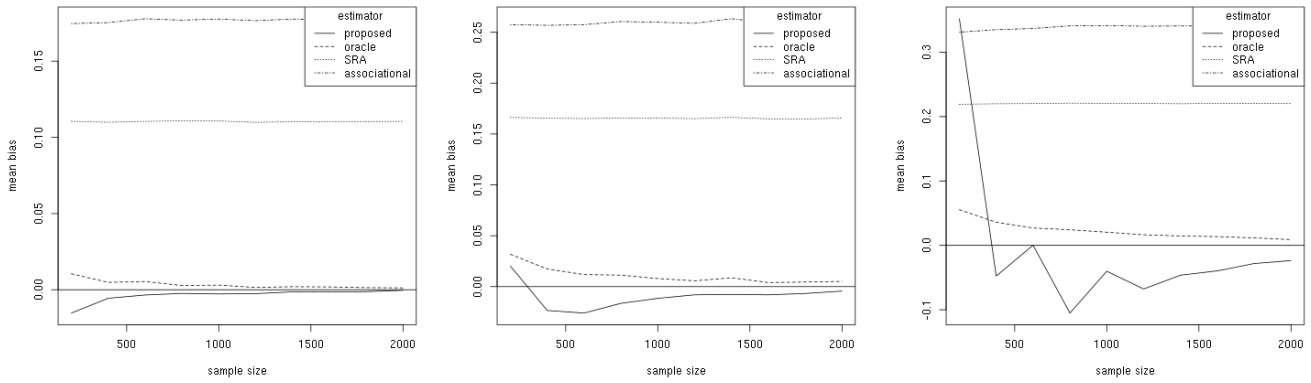
Also computed for comparison were an "oracle" estimator, an SRA estimator, and the associational or "crude" estimator. The oracle estimator uses inverse probability weighting with the true propensity score $P(A(j) = 1 \mid \overline{A}(j-1), \overline{L}(j), \overline{Z}(j), \overline{U}(j))$, i.e., treating $\overline{U}$ as known and taking into account all confounders. The SRA estimator uses inverse probability weighting with the propensity score taking into account only observed confounders, $P(A(j) = 1 \mid \overline{A}(j-1), \overline{L}(j), \overline{Z}(j))$. The associational estimator uses no weights, ignoring all confounding.

For the first estimator, $\delta_j, 0 \leq j \leq J-1$, requires consistent estimation. The MLE is used, obtained by maximizing the full observed data likelihood for $A(j)$,

$$\mathbb{P}(A(j+1) = 1 \mid \overline{LUZ}(j+1), \overline{A}(j)) = \Phi\left((\alpha_0 + \alpha_1 L(j+1))/\sqrt{1+\alpha_2^2}\right) \times (1 - \delta(j+1)) + Z(j+1)\delta(j+1),$$

with

$$\Phi^{-1}(\delta(j+1)) = \gamma_0 + \gamma_1 L(j+1),$$



(a) J=2 (b) J=3 (c) J=4

Figure 3: Mean bias versus sample size of the proposed weighted estimator, for J=2, 3, and 4, time points, compared with oracle (weights including observed and unobserved confounders), SRA (weights including observed confounders), and associational (no weighting) estimators.

obtained by integrating $U(j+1)$ out of (8), the maximization being taken over the coefficients α, viewed as nuisance parameters, and γ, the parameters of interest for estimating δ.

For few time points, $2 \leq J \leq 4$, the bias of the proposed estimator falls off at a comparable rate to that of the oracle estimator. As expected, the SRA and associational estimators are biased. See Figure 3 for plots of the mean bias versus sample size. The estimator is relatively noisy, however, with standard deviations on the order of 1/10 when the bias is on the order of 1/1000. See Table 1 for measures of scale. A semiparametric efficient estimator, to be reported elsewhere, mitigates the noisiness [13].

For inference, we use the nonparametric bootstrap. Although a sandwich variance estimator is easily obtained for this simple linear model, the resulting CIs appear overly conservative. We examine the empirical coverage of a nominal 95% bootstrap CI, varying the sample size n and total number of time points J. The coverage is close to the nominal level for smaller J and larger n, and overconservative for larger J and smaller n. Table 1 presents the detailed results. Also presented are the empirical mean bias and square root of the mean squared error.

J \ n	200	400	600	800	1000
2	0.99,-0.01,0.82	0.98,-0.00,0.53	0.97,-0.01,0.43	0.96,-0.00,0.36	0.96,-0.00,0.32
3	1.00,0.05,21.94	1.00,-0.03,2.57	1.00,-0.02,1.34	0.99,-0.01,1.07	0.99,-0.02,0.93
4	0.99,-0.11,114.28	1.00,-0.04,75.19	1.00,-0.06,36.58	1.00,-0.13,16.54	1.00,-0.04,8.14

J \ n	1200	1400	1600	1800	2000
2	0.96,-0.00,0.29	0.96,-0.00,0.27	0.96,-0.00,0.25	0.95,0.00,0.24	0.95,-0.00,0.23
3	0.98,-0.01,0.83	0.98,-0.01,0.76	0.97,-0.01,0.70	0.97,-0.01,0.66	0.97,-0.01,0.62
4	1.00,-0.06,3.83	1.00,-0.04,2.62	1.00,-0.04,2.12	1.00,-0.03,1.85	1.00,-0.03,1.69

Table 1: Empirical coverage of a nominal 95% bootstrap CI, mean bias, and root mean squared error of the proposed estimator, varying the sample size n and total number of time points J.

6 Discussion

We have shown how IVs may be used to identify causal parameters in marginal structural models. Our assumptions are mainly variations of standard IV or MSM assumptions, the exception being the independent compliance type assumption (6). Whether this assumption may be further relaxed or replaced by an assumption of a different nature requries further research.

We have also described estimation in the case of linear mean-model MSMs. Besides the the sample size and number of time points, the strength or our estimator in any particular setting will depend on the strength of the IV–treatment association and the strength of confounding. We expect improved performance will be obtained from a robust, semiparametric efficient estimator.

References

[1] Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002.

[2] Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

[3] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[4] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[5] Miguel A Hernán, Babette A Brumback, and James M Robins. Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in medicine*, 21(12):1689–1709, 2002.

[6] Hyejin Ko, Joseph W Hogan, and Kenneth H Mayer. Estimating causal treatment effects from longitudinal hiv natural history studies using marginal structural models. *Biometrics*, 59(1):152–162, 2003.

[7] Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

[8] James Robins. Marginal Structural Models. In *1997 Proceedings of the American Statistical Association*, pages 1–10 of 1998 Section on Bayesian Statistical Science, 1997.

[9] James M Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.

[10] James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

[11] James M Robins. Association, causation, and marginal structural models. *Synthese*, 121(1-2):151–179, 1999.

[12] James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. 2000.

[13] E. J Tchetgen Tchetgen, H. Michael, and Y. Cui. Marginal Structural Models for Time-varying Endogenous Treatments: A Time-Varying Instrumental Variable Approach. *ArXiv e-prints*, September 2018.

[14] Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550, 2018.