Multiply robust, locally efficient estimation of MSM parameters in the presence of unmeasured confounding.

May 31, 2023

SUMMARY: Recently proposed methods leverage time-varying instrumental variables to draw causal conclusions from longitudinal data in the presence of potentially time-varying unknown confounders. While an effective proof of concept, the estimators requires specification of several nuisance models, and are typically inconsistent when any of these specifications fails to capture the data. Moreover, the estimators do not make efficient use of all available data. Building on the work of Tchetgen Tchetgen et al. 2018, we present a multiply robust, locally efficient estimator. We apply this estimator to marginal structural mean and Cox model parameters.
KEYWORDS: Causal inference, longitudinal data, instrumental variables.

Observational data are increasingly used to draw inferences and make decisions. While controlled experiments remain the gold standard for inferring causal relationships, cost, ethics, or expedience often compel analysts and decision-makers to rely on observational data. Furthermore, the increasing size and richness of observational data being collected invites analysis. Electronic health records, claims data, sensor data, and social media data are examples of the massive amounts data being collected, typically in a time series. The role of observational data as complementary to the traditional controlled experimental data is increasingly recognized by regulatory bodies, private R&D departments, and has been written into the federal US code (Li et al., 2021).

The central challenge when making decisions based on observational data is the possibility of confounding, i.e., factors affecting both the choice of treatment and outcome under study. A large toolbox has been developed over the last 40 years for drawing causal inferences from longitudinal observational data on the assumption that all confounders are available to the analyst. While not assuming treatment is randomized, these approaches still assume that treatment is randomized conditional on a set of available data. This *sequential ramdonization assumption (SRA)* may therefore be nearly as difficult to meet or nearly as implausible as randomization. Nevertheless, reviews of the literature show that published studies frequently rely on this assumption uncritically, leading to spurious inferences (Clare et al., 2019; Kreif et al., 2013). Longitudinal observational data is particularly susceptible to unknown confounding, due to the manifold complex pathways through which factors may induce dependence between treatment assignment and the outcome.

Instrumental variables are the standard technique for controlling unmeasured confounding (Baiocchi et al., 2014; Martens et al., 2006). An IV may be viewed as a surrogate for a treatment that possesses the crucial property of being random with respect to the outcome under study. An IV therefore brings in the benefits of randomization for drawing causal conclusions, in exchange for a possible loss of efficiency in relation to the surrogacy. While instrumental variables have a long history in the field of economics and the social sciences, their use has been largely limited to non-longitudinal data. Indeed, in an article from 2000, James Robins, one of the founders of the field of longitudinal causal inference, expressed doubt that time-varying IVs could ever be used to identify causal effects. Many of the core applications motivating causal inference, such as longitudinal drug trials and cohort studies, have thus been out of reach to IV methods.

Cui et al. (2023); Michael et al. (2023); Tchetgen Tchetgen et al. (2018) have recently proposed tools to leverage time-varying instrumental variables to draw causal conclusions from longitudinal data in the presence of potentially time-varying unknown confounders. Specifically, Michael et al. (2023) uses IVs to identify a *marginal structural mean model*, the most common model for obtaining the causal effect of a time-varying treatment on an outcome. Cui et al. (2023) uses IVs to identify the parameters of a *marginal structural Cox model*, the most common model for evaluating the causal effect of a time-varying treatment on a censored failure time outcome. These identification results lead to naive estimators for the MSM. While an effective proof of concept, the naive estimator suffers from inefficiency, non-robustness, and unnecessary limitations on the data. These problems limit the widespread deployment of IV MSM estimation.

1 Background/notation

We begin by providing background on MSMs. We are interested in the causal relationship between a course of treatments $A_1, \dots, A_T$, on an outcome Y. Included among the data is a process of covariates $L_1, \dots, L_T$. We denote by $\mathcal{A}$ the common sample space of the treatments, and we initially assume this space is discrete. We adopt the potential outcomes framework, which postulates potential outcomes $Y_{\overline{a}}(\omega)$, random variables indexed by treatment levels $\overline{a} = (a_1, \dots, a_T)$. The potential outcome $Y_{\overline{a}}(\omega)$ is interpreted as the response of a unit ω if, possibly contrary to fact, treatment $\overline{a}$ were applied to ω. Potential outcomes are only partially observed since not all units receive all treatments. What is observed is $Y_{\overline{A}(\omega)}(\omega)$, the potential outcome of unit ω under the treatment $\overline{A}(\omega)$ received by unit ω. We use $\perp\!\!\!\perp$ to denote statistical independence and $\mathbb{P}_n$ to denote expectation with respect to the empirical distribution of the data. We use overbars to indicate the history of a quantity, e.g., $\overline{a} = (a_1, \dots, a_T)$.

A marginal structural mean model ("MSMM") is a model on the marginal means of the potential outcomes $Y_{\overline{a}}, \overline{a} \in \mathcal{A}^T$ (Robins, 1997). For example, the effect of a treatment regime may be modeled as linear in the cumulative treatment taken,

$$\mathbb{E}(Y_{\overline{a}}) = \beta_0 + \beta_1 \sum_{t=0}^{T} a_t. \tag{1}$$

In this example, $\beta \in \mathbb{R}^2$ parameterizes the MSMM and encodes the incremental effect of a unit of treatment. A link function can be introduced to accommodate binary or count outcome variables, e.g., $\mathbb{E}(Y_{\overline{a}}) = (1 + \exp(\beta_0 + \beta_1 \sum_{t=0}^{T} a_t))^{-1}$ for binary Y. In general we write

$$\mathbb{E}(Y_{\overline{a}}) = m_\beta(\overline{a}) \tag{2}$$

to describe an MSMM, where $m_\beta : \mathcal{A}^T \to \mathbb{R}$ belongs to a family of functions parameterized by finite-dimensional β. The model parameter β is the target of inference.

A marginal structural Cox model ("Cox MSM") is a certain model on the hazard function of the potential outcomes $Y_{\overline{a}}, \overline{a} \in \mathcal{A}^T$,

$$\lambda_{Y_{\overline{a}}} = \lambda_0(t) \exp(m(\overline{a}_t, t, \beta, V)). \tag{3}$$

Here, λ_0 is an unspecified baseline hazard function, $V \in L(0)$ are baseline covariates, and the function m satisfies $m(0, t, \beta, V) = 0$ for all t, β, V. The target of inference as before is β, a finite-dimensional parameter indexing m and, through m, the hazard function.

Since a MSM is a model on the incompletely observed potential outcomes $Y_{\overline{a}}$, there is no guarantee that the observed data alone can pick out the parameter β of the data generating process. That is, without further assumptions, the observed data may not uniquely identify the MSM parameter. We describe here two identification results, one classical and the other the subject of Cui et al. (2023); Michael et al. (2023). Both are formally the same: They work by relating the law of the potential outcomes $Y_{\overline{a}}$ to a moment condition on the observed data $(Y, \overline{A}, \overline{Z}, \overline{L})$,

$$\mathbb{E}\left(h(\overline{A}) \frac{Y - m_\beta(\overline{A})}{\overline{W}} \right) = \sum_{\overline{a}} \mathbb{E}\left(h(\overline{a})(Y_{\overline{a}} - m_\beta(\overline{a})) \right) = 0 \tag{4}$$

for an MSMM, and

$$\mathbb{E}\left(\int \frac{dN(t)}{\overline{W}} \left(h(\overline{A}) - \frac{\mathbb{E}\left(h(\overline{A}) \exp(m(\overline{A}(t), t, \beta, V))\{Y \geq t\}/\overline{W} \right)}{\mathbb{E}\left(\exp(m(\overline{A}(t), t, \beta, V))\{Y \geq t\}/\overline{W} \right)} \right) \right) = 0 \tag{5}$$

for a Cox MSM. The difference between the two lies in the choice of weights $\overline{W}$ and the assumptions justifying their use. In both cases h is an arbitrary function on $\mathcal{A}^T$ of the same dimension as β, to be discussed further below. Whichever set of assumptions and weights one relies on, the MSM parameter may be estimated as the solution to the estimating equation obtained as the empirical form of (4):

$$\mathbb{P}_n \left(h(\overline{A}) \frac{Y - \mu_\beta(\overline{A})}{\overline{W}} \right) = 0 \tag{6}$$

for an MSMM, and analogously for a Cox MSM. With the weights in hand, this estimation is a weighted regression that may be carried out in many popular software packages.

The two sets of identification assumptions and associated weights are

1. *SRA MSM estimation.* If one is willing to assume that all confounders have been accounted for, (Robins, 1998, 1997) give the classical identification result for the MSM parameter. Specifically, under the SRA and the positivity assumption,

$$Y_{\overline{a}} \perp\!\!\!\perp A_t \mid \overline{L}_t, \overline{A}_{t-1}, \quad 1 \leq t \leq T \qquad \text{(SRA)} \tag{7}$$

$$0 < f_{A_t|\overline{A}_{t-1},\overline{L}_t}(a_t \mid \overline{A}_{t-1}, \overline{L}_t) \text{ a.s.}, \quad a_t \in \mathcal{A}, 1 \leq t \leq T \qquad \text{(positivity)} \tag{8}$$

   the weights $\overline{W}$ in (4) may be chosen to be

$$\overline{W}^{(SRA)} = \overline{W}_T^{(SRA)} = \prod_{t=1}^{T} W_t^{(SRA)} \tag{9}$$

$$W_t^{(SRA)} = f_{A_t|\overline{A}_{t-1},\overline{L}_t}(A_t \mid \overline{A}_{t-1}, \overline{L}_t) \qquad 1 \leq t \leq T. \tag{10}$$

   SRA will hold if the cumulative observed data at each time point capture all systematic associations between the treatment and outcome of interest. Positivity will hold when, among all subpopulations defined by covariates $\overline{L}_t$ and a treatment regime $\overline{A}_{t-1}, t \leq T$, there are further subpopulations at each possible treatment level $a_t \in \mathcal{A}$.

It is frequently impossible or imprudent to assume that there are no unmeasured confounders. Instrumental variables provide a way to manage unknown confounding, rather than stipulating that none is present. Informally, an IV is a random variable associated with the treatment of interest that has no association with the outcome of interest except as mediated by the treatment. For example, a prescription randomly assigned to study subjects is orthogonal to any pretreatment variables, including unobserved confounders, but would usually influence whether the subject actually follows the prescription. Therefore the prescription may serve as an IV for the treatment actually taken or not taken. More formally, a random variable is an IV when it satisfies (i) (IV relevance) the IV Z must be associated with the treatment A; (ii) (exclusion restriction) the IV Z may not be a direct cause of the treatment A; and (iii) (IV independence) there are no unknown confounders of the relation between the IV Z and outcome Y. Whereas SRA requires that the treatment of interest not share any unknown confounders with the outcome, the IV approach allows the analyst to meet this requirement by any other available quantity that may act as a surrogate for the treatment.

2. *IV MSM estimation.* Cui et al. (2023); Michael et al. (2023); Tchetgen Tchetgen et al. (2018) establish that time-varying binary-valued IVs may be used to identify MSM parameters. Besides straightforward longitudinal generalizations of standard IV assumptions, the key assumption needed is Independent Compliance Type:

$$\delta_t \perp\!\!\!\perp \overline{U}_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t, \quad 1 \leq t \leq T \tag{11}$$

$$\text{where } \delta_t = f\left(a_t|\overline{A}_{t-1}, \overline{Z}_{t-1}, Z_t = 1, \overline{L}_t, \overline{U}_t\right) - f\left(a_t|\overline{A}_{t-1}, \overline{Z}_{t-1}, Z_t = 0, \overline{L}_t, \overline{U}_t\right) \tag{12}$$

   The ICT assumption states that while $\overline{U}_t$ may confound the causal effects of $\overline{A}_t$, no component of $\overline{U}_t$ interacts with Z_t in its additive effects on A_t. The name of the assumption comes from the aforementioned design in which the IV Z is a doctor's prescription and the treatment A is taking the prescribed medicine. A patient is grouped into 1 of 4 *compliance classes* depending on whether the patient was or was not prescribed the treatment, and whether the patient did or did not take the medicine. The ICT assumption is that unknown confounders do not interact with compliance class at any stratum of the population under consideration. The assumption will be met if enough data on subjects is collected to capture all systematic differences in compliance type. Under the ICT assumption and IV assumptions, the weights in (4) may be chosen as $\overline{W}^{(IV)} = \prod_{t=1}^{T} W_t^{(IV)}$ with

$$W_t^{(IV)} = (-1)^{1-Z_t} f_{Z_t}\left(Z_t \mid \overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1}\right) \delta_t\left(\overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1}\right). \tag{13}$$

2 Summary of results

We use instrumental variable to manage unmeasured confounding. Michael et al. (2023) and Cui et al. (2023) present simple IV estimators for MSMM and Cox MSM parameters that are not robust to misspecifications to nuisance models. That is, this estimator depends on several unknown quantities that must be estimated from the data under parametric assumptions, and if any of these parametric assumptions fails to be met, the resulting estimator is likely to be inconsistent. Besides sensitivity to model misspecification, the naive estimator does not make efficient use of the data. The notion of efficiency used here is drawn from semiparametric theory, which provides benchmarks for efficiency not met by the naive estimator.

A multiply robust, locally efficient estimator mitigates these problems. Multiple robustness is the property that the estimator remains asymptotically normal even when certain of the nuisance models are misspecified. If it so happens that all nuisance models are correctly specified, the estimator achieves the semiparametric efficiency bound: Its asymptotic variance is minimal among all regular asymptotically linear estimators subject to the IV assumptions and MSM. The wide class of RAL estimators is of interest for benchmarking purposes since these are the estimators known to be stable under small changes to the data generating process (Bickel et al., 1993).

We build on Tchetgen Tchetgen et al. (2018) to obtain a multiply robust, locally efficient IV MSM estimator. That technical report presents a multiply robust, locally efficient estimator in the more general setting of marginal structural models. When the identification conditions hold, the estimator may be obtained as the solution in β of

$$\begin{aligned} o_p(n^{-1/2}) = {} & \mathbb{P}_n\left(\frac{D_{sm}(h,\beta)}{\overline{W}^{(IV)}}\right) \\ & -\mathbb{P}_n\left(\sum_{t=1}^{T} \frac{1}{\overline{W}_{t-1}^{(IV)}}\left(\frac{(-1)^{1-Z_t}}{f(Z_t \mid \overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1})}\left(\psi_t(\beta) - \frac{\tilde{\psi}_t(\beta)(A_t - \mathbb{E}(A_t \mid \overline{A}_{t-1}, \overline{L}_t, \overline{Z}_t))}{\delta_t}\right) - \tilde{\psi}_t(\beta)\right)\right) \end{aligned}$$

where

$$\psi_T(\beta) = \mathbb{E}\left(\frac{D_{sm}(h,\beta)}{\delta_T}\middle|\overline{A}_{T-1}, \overline{L}_T, \overline{Z}_t\right),$$

for $t = 1, \ldots, T-1$,

$$\psi_t(\beta) = \mathbb{E}\left(\tilde{\psi}_{t+1}/\delta_t \mid \overline{A}_{t-1}, \overline{L}_t, \overline{Z}_t\right), \tag{14}$$

and for $t = 1, \ldots, T$,

$$\tilde{\psi}_t(\beta) = \mathbb{E}(\psi_t(\beta) \mid \overline{A}_{t-1}, \overline{L}_t, \overline{Z}_{t-1}). \tag{15}$$

The estimating function $D_{sm}(h, \beta)$ is, in the case of an MSMM,

$$D_{sm}(h, \beta) = h(\overline{A})(Y - \mu_\beta(\overline{A}))$$

and, in the case of a Cox MSM,

$$D_{sm}(h, \beta) = \int dN(t) \left(h(\overline{A}) - \frac{\mathbb{P}_n\left(h(\overline{A})\exp(m(\overline{A}(t), t, \beta, V))\{Y \geq t\}\right)}{\mathbb{P}_n\left(\exp(m(\overline{A}(t), t, \beta, V))\{Y \geq t\}\right)} \right).$$

The complicated form taken by the estimating equation is one of the hurdles to be overcome in this paper. In particular, the recursive form of the nuisance parameters $\psi_t(\beta)$ and $\tilde{\psi}_t(\beta)$ poses modeling and computational challenges. Despite the promises of multiple robustness and local efficiency, to date no estimator has been developed taking advantage of this result.

For practical use, this estimator requires estimation of $\psi_t(\beta), \delta$, $f_{Z_t|\overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1}}$, and $\epsilon_t = \mathbb{E}(A_t \mid \overline{A}_{t-1}, \overline{L}_t, \overline{Z}_t)$, $t = 1, \ldots, T$. Theorem 1 argues that, under mild conditions, the solution to the estimating equation (6) remains asymptotically normal whenever models for any of the following 3 sets of quantities are correctly specified: (i) $\mathcal{M}_1 : f_{Z_t|\overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1}}$ and δ_t, (ii) $\mathcal{M}_2 : f_{Z_t|\overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1}}$ and $\tilde{\psi}_t(\beta)$, (iii) $\mathcal{M}_3 : \tilde{\psi}_t(\beta), \psi_t(\beta)$, and $\mathbb{E}(A_t \mid \overline{A}_{t-1}, \overline{L}_t, \overline{Z}_t)$, $t = 1, \ldots, T$. That is, the estimator is asymptotically normal whenever the data generating process for the nuisance terms lies in the union $\mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$. Building on results in Tchetgen Tchetgen et al. (2018), Theorem 2 establishes that, when all 3 models are correctly specified, i.e., the data generating process for the nuisance terms lies in $\mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3$, the solution to the estimating equation (6) with a particular choice of h achieves the semiparametric efficiency bound.

As a proof of concept, we consider a simple model where the covariates and treatment are binary. One might encounter data appropriate to this model in a SMART study where a binary or dichotomized baseline measure (e.g., early versus advanced disease, high versus low CD4 count) informs treatment assignment (the IV) and is then tracked over a period of time (the covariate process) at the end of which a final measurement is taken as the outcome. When all the data is binary-valued the $\psi_t(\beta)$ terms may be estimated nonparametrically; parametric estimation of ψ_t, which presents certain challenges, will be discussed shortly. We consider $T = 2$ time points.

Multiple Robustness. Table 1 gives the result of an examination of the multiple robustness of the proposed estimator with a sample of size $n = 1000$. We test multiple robustness by moderate misspecifications to the models for the 4 nuisance terms. An example of such a misspecification is to model $f(Z_t \mid \overline{L}_t, \overline{A}_{t-1}, \overline{Z}_{t-1})$ as bernoulli with success probability $\text{expit}(\gamma_0 + \gamma_1 L_t)$ while sampling the data using a success probability $\text{expit}(\gamma_0 + \gamma_1' L_t)$ for $\gamma_1 \neq \gamma_1'$. Thus the data generating process does not lie in $\mathcal{M}_1 \cup \mathcal{M}_2$. In each case, these

misspecification				naive estimator			efficient estimator		
f_Z	δ	ψ_t	ϵ	bias	SD	MSE	bias	SD	MSE
x				1.35	22.23	495.71	0.02	0.54	0.30
	x			-0.01	15.07	226.98	-0.08	2.33	5.44
		x		0.35	17.40	302.70	-0.00	0.77	0.59
x	x			-0.03	5.82	33.89	-0.08	1.15	1.33
			x	0.54	11.89	141.47	-0.03	0.98	0.96

Table 1: Pilot study: The performance of the naive and the proposed IV MSMM estimator under misspecification to the models for nuisance parameters.

moderate misspecifications are enough to drive up the bias or variance of the naive estimator to the point that it becomes impractical. The performance of the multiply robust estimator, however, remains reasonable.

Local efficiency. When all nuisance models are correctly specified, the efficient estimator achieves the semiparametric efficiency bound. Figure 1 describes the efficiency of the estimator in this situation. In this pilot study, the figure shows that the MSE of the efficient estimator is about 2/3 that of the naive estimator, across a range of sample sizes.

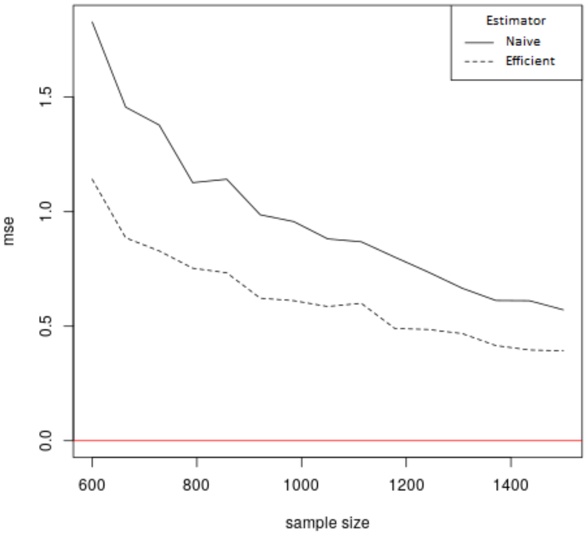

Figure 1: Pilot study: MSE of the naive and proposed IV MSMM estimator.

While this pilot study is encouraging, more realistic applications require consideration of, at least, non-binary covariates; vectors of covariates; dependence between the IV at each time step and past covariates; and direct dependence of the treatment on earlier time points (i.e., a non-markov treatment process). With more complex data such as these, the nuisance parameters such as ψ_t typically cannot be modeled nonparametrically due to the curse of dimensionality and will require parametric specification. The main challenge here is that the parametric models must be chosen so as to respect the recursive relationships (14,15).

Many natural models, such as a logistic model for the treatment, are incompatible with the integrations.

The problem of parametrizing mutually consistent models for time-varying parameters is not specific to efficient IV estimation of MSM parameters. Recent work of Babino et al. (2019), building on the foundational work of Robins (2000), addresses the issue in the setting of MSMs under SRA. Next we propose a parametrization of the likelihood for the data that implies a consistent sequence of models.

References

Babino, L., A. Rotnitzky, and J. Robins (2019). Multiple robust estimation of marginal structural mean models for unconstrained outcomes. *Biometrics 75*(1), 90–99.

Baiocchi, M., J. Cheng, and D. S. Small (2014). Instrumental variable methods for causal inference. *Statistics in medicine 33*(13), 2297–2340.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993). *Efficient and adaptive estimation for semiparametric models*, Volume 4. Johns Hopkins University Press Baltimore.

Clare, P. J., T. A. Dobbins, and R. P. Mattick (2019). Causal models adjusting for time-varying confounding—a systematic review of the literature. *International journal of epidemiology 48*(1), 254–265.

Cui, Y., H. Michael, F. Tanser, and E. Tchetgen Tchetgen (2023). Instrumental variable estimation of the marginal structural cox model for time-varying treatments. *Biometrika 110*(1), 101–118.

Kreif, N., R. Grieve, and M. Z. Sadique (2013). Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health economics 22*(4), 486–500.

Li, M., S. Chen, Y. Lai, Z. Liang, J. Wang, J. Shi, H. Lin, D. Yao, H. Hu, and C. O. L. Ung (2021). Integrating real-world evidence in the regulatory decision-making process: a systematic analysis of experiences in the us, eu, and china using a logic model. *Frontiers in medicine 8*, 778.

Martens, E. P., W. R. Pestman, A. de Boer, S. V. Belitser, and O. H. Klungel (2006). Instrumental variables: application and limitations. *Epidemiology*, 260–267.

Michael, H., Y. Cui, S. A. Lorch, and E. J. Tchetgen Tchetgen (2023). Instrumental variable estimation of marginal structural mean models for time-varying treatment. *Journal of the American Statistical Association*, 1–12.

Robins, J. (1998). Marginal Structural Models. In *1997 Proceedings of the American Statistical Association*, pp. 1–10 of 1998 Section on Bayesian Statistical Science.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pp. 69–117. Springer.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, Volume 1999, pp. 6–10. Indianapolis, IN.

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

A Proof of multiple robustness

Establish multiple robustness by showing that the estimating equation remains unbiased,

$$E\left(\underset{(i)}{\frac{D_{sm}}{\overline{W}^*_T}} - \underset{(ii)}{\sum_{t=1}^{T} \frac{1}{\overline{W}^*_{t-1}} \left(\frac{(-1)^{1-Z_t}\psi^*}{f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} - \widetilde{\psi}^*_t \right)} - \underset{(iii)}{\sum_{t=1}^{T} \frac{\epsilon^*_t \widetilde{\psi}^*_t}{\overline{W}^*_t}} \right) = 0, \tag{16}$$

though certain subsets of the starred quantities may not be the same as the corresponding unstarred quantities. Specifically, the above holds whenever any one of the following hold:

1. $\overline{W}^*_t = \overline{W}_t$, i.e., $f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t) = f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$ and $\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t) = \Delta_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$, $t = 1, \ldots, T$.

2. $f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t) = f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$ and $\widetilde{\psi}^*_t = \widetilde{\psi}_t$, $t = 1, \ldots, T$.

3. $\psi^* = \psi$, $\widetilde{\psi}^*_t = \widetilde{\psi}_t$, and $\epsilon^*_t = \epsilon_t$, i.e., $E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t) = E(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t)$, $t = 1, \ldots, T$.

This result implies that the MSM parameter β may typically be consistently estimated using consistent estimators for the quantities in any one of the three above models.((we have convergence in prob, need convergence in mean)) In the above, E^*, f^* does not denote a type of expectation or density, and is just a notation for some substitute possibly random quantity for the corresponding unstarred quantity. The following mild assumptions are made on the starred quantities. For all $t, 1 \leq t \leq T$,

1. The starred quantities must be functions of the same sets of random variables as the corresponding unstarred quantities. That is, $\psi^*, \overline{W}^*_t, E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t) \in \sigma(\overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t)$, and $f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t), \widetilde{\psi}^*_t, \Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t) \in \sigma(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$.

2. $\widetilde{\psi}^*_t$ must be compatible with ψ^* in the sense that $\widetilde{\psi}^*_t = \psi^*|_{Z_t=1} - \psi^*|_{Z_t=0}$.

3. ϵ^*_t and $\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$ are compatible in the sense that $E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t)|_{Z_t=1} - E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t)|_{Z_t=0} = \Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$, equivalently, $E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t) = E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t)|_{Z_t=0} + Z_t\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$.

4. $\overline{W}^*_t$ is comptaible with $f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$ and $\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$ in the sense that $\overline{W}^*_t = (-1)^{1-Z_t} f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)$.

1. For the first model, consider the 3 terms in ((ref))

   i. $E\left(\frac{D_{sm}}{\overline{W}^*_T}\right) = E\left(\frac{D_{sm}}{\overline{W}_T}\right) = 0$ by Theorem ((ref)) in ((JASA paper)).

ii.

$$E\sum_{t=1}^{T}\frac{1}{\overline{W}^{*}_{t-1}}\left(\frac{(-1)^{1-Z_t}\psi^{*}}{f^{*}(Z_t\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}-\widetilde{\psi}^{*}_{t}\right)\tag{17}$$

$$=\sum_{t=1}^{T}E\left(\frac{1}{\overline{W}_{t-1}}E\left(\frac{(-1)^{1-Z_t}\psi^{*}}{f^{*}(Z_t\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}-\widetilde{\psi}^{*}_{t}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right)\right)\tag{18}$$

and by the assumed comptaibility of ψ^{*} and $\widetilde{\psi}^{*}_{t}$.

$$E\left(\frac{(-1)^{1-Z_t}\psi^{*}}{f^{*}(Z_t\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}-\widetilde{\psi}^{*}_{t}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right)=\widetilde{\psi}^{*}_{t}-\widetilde{\psi}^{*}_{t}=0.\tag{19}$$

iii.

$$E\sum_{t=1}^{T}\frac{\epsilon^{*}_{t}\widetilde{\psi}^{*}_{t}}{\overline{W}^{*}_{t}}=E\sum_{t=1}^{T}\widetilde{\psi}^{*}_{t}E\left(\frac{\epsilon^{*}_{t}}{\overline{W}_{t}}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right).\tag{20}$$

Writing $\epsilon^{*}_{t}=A_t-E^{*}(A_t\mid\overline{A}_{t-1},\overline{Z}_t,\overline{L}_t)=A_t-(E^{*}(A_t\mid\overline{A}_{t-1}\overline{Z}_{t-1},Z_t=0,\overline{L}_t)+Z_t\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)$,which holds by the assumed compatibility of ϵ^{*}_{t} and $\Delta^{*}_{t}(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)$, the inner expectation is

$$E\left(\frac{\epsilon^{*}_{t}}{\overline{W}_{t}}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right)\tag{21}$$

$$=E\left(\frac{(-1)^{1-Z_t}(A_t-Z_t\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t))}{fzt\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right)\tag{22}$$

$$-E\left(\frac{E^{*}(A_t\mid\overline{A}_{t-1}\overline{Z}_{t-1},Z_t=0,\overline{L}_t)}{f(Z_t\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right).\tag{23}$$

The first term is

$$\frac{1}{\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}\sum_{z_t\in\{0,1\}}(-1)^{1-z_t}(E(A_t\mid\overline{A}_{t-1}\overline{Z}_{t-1},Z_t=z_t,\overline{L}_t)-z_t\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t))\tag{24}$$

$$=\frac{1}{\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}(\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)-\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t))=0.\tag{25}$$

The second term is

$$\frac{E^{*}(A_t\mid\overline{A}_{t-1}\overline{Z}_{t-1},Z_t=0,\overline{L}_t)}{\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}E\left(\frac{(-1)^{1-Z_t}}{f(Z_t\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}\mid\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t\right)\tag{26}$$

$$=\frac{E^{*}(A_t\mid\overline{A}_{t-1}\overline{Z}_{t-1},Z_t=0,\overline{L}_t)}{\Delta_t(\overline{A}_{t-1},\overline{Z}_{t-1},\overline{L}_t)}(1-1)=0.\tag{27}$$

2. For the second model, term (ii) vanishes as under the first model.

   For terms (i) and (iii), using the definitions

$$\widetilde{\psi}_t = E\left(\frac{(-1)^{1-Z_t}\psi}{f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t\right) \tag{28}$$

$$\psi = \frac{1}{\Delta_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} E\left(\widetilde{\psi}_{t+1} \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t\right) \tag{29}$$

   form the recurrence for $\widetilde{\psi}_t$

$$\widetilde{\psi}_t \Delta_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t) = E\left(\frac{(-1)^{1-Z_t} E\left(\widetilde{\psi}_{t+1} \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t\right)}{f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t\right) \tag{30}$$

$$= E\left(\frac{(-1)^{1-Z_t}\widetilde{\psi}_{t+1}}{f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t\right). \tag{31}$$

The sum in (iii) telescopes using this recurrence,

$$\begin{aligned}
E\sum_{t=1}^{T}\frac{\epsilon^*_t\widetilde{\psi}^*_t}{\overline{W}^*_t} &= E\sum_{t=1}^{T}(-1)^{1-Z_t}\frac{(A_t - E^*(A_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t))\widetilde{\psi}^*_t}{\overline{W}^*_{t-1}\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \qquad (32)\\
&= E\sum_{t=1}^{T}(-1)^{1-Z_t}\frac{(A_t - E^*\left(A_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, Z_t = 0, \overline{L}_t\right) - Z_t\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t))\widetilde{\psi}^*_t}{\overline{W}^*_{t-1}\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \qquad (33)\\
&= \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)}E\left(\frac{(-1)^{1-Z_t}A_t}{f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t\right)\right) \qquad (34)\\
&- \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t E^*\left(A_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, Z_t = 0, \overline{L}_t\right)}{\overline{W}^*_{t-1}\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)}E\left(\frac{(-1)^{1-Z_t}}{f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t\right)\right) \qquad (35)\\
&- \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}E\left(\frac{(-1)^{1-Z_t}Z_t}{f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t\right)\right) \qquad (36)\\
&= \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t\Delta_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)}{\overline{W}^*_{t-1}\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)}\right) - \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \qquad (37)\\
&= \sum_{t=1}^{T}E\left(\frac{(-1)^{1-Z_t}\widetilde{\psi}_{t+1}}{\overline{W}^*_{t-1}\Delta^*_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)f(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)}\right) - \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \qquad (38)\\
&= \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_{t+1}}{\overline{W}^*_t}\right) - \sum_{t=1}^{T}E\left(\frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \qquad (39)\\
&= E\left(\frac{\widetilde{\psi}_{T+1}}{\overline{W}^*_T}\right) - E\left(\frac{\widetilde{\psi}_0}{\overline{W}^*_0}\right) \qquad (40)\\
&= E\left(\frac{\widetilde{\psi}_{T+1}}{\overline{W}^*_T}\right) \qquad (41)
\end{aligned}$$

Therefore, the difference of terms (i) and (iii) is 0.

3. For the third model, $\epsilon^*_t = \epsilon_t$ implies

$$E\sum_{t=1}^{T}\frac{\epsilon^*_t\widetilde{\psi}^*_t}{\overline{W}^*_t} = E\sum_{t=1}^{T}\frac{\widetilde{\psi}^*_t}{\overline{W}^*_t}E\left(\epsilon^*_t \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t\right) = 0. \tag{42}$$

The difference of terms (i) and (ii) is

$$E\left(\frac{D_{sm}}{\overline{W}^*_T} - \sum_{t=1}^{T} \frac{1}{\overline{W}^*_{t-1}} \left(\frac{(-1)^{1-Z_t}\psi^*}{f^*(Z_t \mid \overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)} - \widetilde{\psi}^*_t\right)\right) \tag{43}$$

$$= E\left(\frac{D_{sm}}{\overline{W}^*_T} - \sum_{t=1}^{T} \frac{\psi^* \Delta_t(\overline{A}_{t-1}, \overline{Z}_{t-1}, \overline{L}_t)}{\overline{W}^*_t} + \sum_{t=1}^{T} \frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \tag{44}$$

$$= E\left(\frac{D_{sm}}{\overline{W}^*_T} - \sum_{t=1}^{T} \frac{E(\widetilde{\psi}_{t+1} \mid \overline{A}_{t-1}, \overline{Z}_t, \overline{L}_t)}{\overline{W}^*_t} + \sum_{t=1}^{T} \frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \tag{45}$$

$$= E\left(\frac{D_{sm}}{\overline{W}^*_T} - \sum_{t=1}^{T} \frac{\widetilde{\psi}_{t+1}}{\overline{W}^*_t} + \sum_{t=1}^{T} \frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \tag{46}$$

$$= E\left(\frac{D_{sm}}{\overline{W}^*_T} = \frac{D_{sm}}{\overline{W}^*_T} - \sum_{t=1}^{T-1} \frac{\widetilde{\psi}_{t+1}}{\overline{W}^*_t} + \sum_{t=1}^{T} \frac{\widetilde{\psi}_t}{\overline{W}^*_{t-1}}\right) \tag{47}$$

$$= E\left(\frac{\widetilde{\psi}_t}{\overline{W}^*_0}\right) = 0. \tag{48}$$

B Proof of semiparametric efficiency

((write out differentiation wrt parametric sub-models))