Supplementary Materials - Impact of discretization of the timeline for longitudinal causal inference methods

Steve Ferreira Guerra[1,2], Mireille E. Schnitzer[1,2], Amélie Forget[1,3], Lucie Blais[1,3]

[1]Faculté de Pharmacie, Université de Montréal, Montréal, Canada
[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Canada
[3]Research Center, Hôpital du Sacré-Coeur de Montréal, Montréal, Canada

1 Consistency violation example

Consider an example where the finest discretizations includes 4 time-points and that the (true) observed exposure at the finest discretization for a given individual consists of $\bar{A} = (1, 0, 0, 1)$. Consider also a single final observed outcome $Y(4)$. If consistency holds at the finest discretization: $Y(4) = Y^{\bar{a}}(4) = Y^{(1,0,0,1)}(4), \text{ if } \bar{A} = \bar{a}$.

Now imagine some discretization approach, say reducing the timeline to 2 time-points, which results in the observed exposure on the discretized data to be $\bar{A} = (1, 1)$. The observed outcome, $Y(4)$, which remains is not affected by discretization, since it is observed at the end of study, may not be equal to the counterfactual outcome that would have been observed had we intervened under the discretized version of the exposure at each of the two time-points, i.e. $Y(4) = Y^{(1,0,0,1)}(4) \neq Y^{(1,1)}(4)$. Under our definition of the hypothetical intervention, the regimen (1,1) corresponds to sustained exposure from time-point 1 to 4. However, this does not correspond to the observed exposure which is not sustained. As such, we would not necessarily expect the counterfactual outcome under the hypothetical intervention on the discretized timeline to correspond to the observed outcome. Hence consistency does not hold.

2 Pooled LTMLE algorithm

Recall from section 6.1 of the main manuscript, that $\bar{Q}_t^{\bar{a}_r}(t) = E(Y^{\bar{a}_r}(t) \mid \bar{A}(t-1) = \bar{a}_r(t-1), \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1))$. and, for $j = t, ..., 1$, $\bar{Q}_t^{\bar{a}_r}(j) = E(\bar{Q}_t^{\bar{a}_r}(j+1) \mid \bar{A}(j-1) = \bar{a}_r(j-1), \bar{\boldsymbol{L}}(j-1), \bar{Y}(j-1))$ where $\bar{Q}_t^{\bar{a}_r}(t+1) := Y^{\bar{a}_r}(t)$. Further recall that, for every $j = t, \ldots, 1$, $\bar{g}^{\bar{a}_r}(j) = \prod_{k=0}^{j} P(A(k) = a(k) \mid \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$. The pooled LTMLE algorithm for our specific example where $Y(t)$ is a binary indicator of a failure type outcome by time t and the target parameters are the MSM coefficients defined in equation (1) of the main manuscript can be implemented as follows:

First, calculate the estimator $\bar{g}_n^{\bar{a}_r}(j)$ of $\bar{g}^{\bar{a}_r}(j) = \prod_{k=0}^{j} P(A(k) = a(k) \mid \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$, by estimating every component $\bar{g}_{k,n}^{\bar{a}_r}(j) = P_n(A(k) = a(k) \mid \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$. Denote $\bar{g}_n^{\bar{a}_r}(j) = \prod_{k=0}^{j} \bar{g}_{k,n}^{\bar{a}_r}(j)$ as their product.

Define $\bar{Q}_{t,n}^{\bar{a}_r,*}(t+1) \equiv Y(t)$.

For $t = K_r + 1, \ldots, 1$ set $j = t$:

Step 1: Initial estimation of $\bar{Q}_t^{\bar{a}_r}(j)$

Calculate the initial estimate by evaluating the conditional expectation $\bar{Q}_t^{\bar{a}_r}(j) = E(\bar{Q}_{t,n}^{\bar{a}_r,*}(j+1) \mid \bar{A}(j-1), \bar{\boldsymbol{L}}(j-1)), \bar{Y}(j-1))$. Then predict $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ for each possible value of $\bar{a}_r$ and every subject $i = 1, \ldots, n$. It follows that $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ is of length $n \times$ *number of potential regimes*.

Step 2: Updating step

For every individual and for every exposure regime $\bar{a}_r \in \bar{\mathcal{A}}_r$ calculate:

$$\boldsymbol{h}_j(\bar{a}_r, t) = I(\bar{A}(j-1) = \bar{a}_r(j-1)) \left(\frac{\partial}{\partial \boldsymbol{\beta}} \eta(\boldsymbol{\beta}, \bar{A}, t) \right) \Bigg|_{\bar{A}(t-1) = \bar{a}_r(t-1)} . \qquad (1)$$

The term $\frac{d}{d\boldsymbol{\beta}} \eta(\boldsymbol{\beta}, \bar{A}, t)$ is the derivative of the MSM formula with respect to $\boldsymbol{\beta}$ which is being evaluated at $\bar{A}(t-1) = \bar{a}_r(t-1)$. The term $I(\bar{A}(j-1) = \bar{a}_r(j-1))$ is an indicator function that is equal to 1 if the observed exposure up until time $j-1$ is equal to the exposure regime $\bar{a}_r(j-1)$. For an individual and a specific exposure regime $\bar{a}_r$, the dimension of $\boldsymbol{h}_j(\bar{a}_r, t)$ is equal to the dimension of $\boldsymbol{\beta}$. Therefore, $\boldsymbol{h}_j(\bar{a}_r, t)$ is of dimension ($n \times$ *number of potential regimes*) $\times$ *number of MSM coefficients*.

The following submodel defines the update path for $\bar{Q}^{\bar{a}_r,*}_{t,n}(j)$:

$$logit(\bar{Q}^{\bar{a}_r,*}_{t,n}(j)) = logit(\bar{Q}^{\bar{a}_r}_{t,n}(j)) + \boldsymbol{\epsilon}\boldsymbol{h}_j(\bar{a}_r, t). \tag{2}$$

$\bar{Q}^{\bar{a}_r,*}_{t,n}(j)$ is obtained by fitting an intercept-free weighted pooled logistic regression, over all $\bar{a}_r$, of $\bar{Q}^{\bar{a}_r,*}_{t}(j+1)$ on the covariates $\boldsymbol{h}_j(\bar{a}_r, t)$ using the previously calculated initial estimates $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ as offset and $1/\bar{g}^{\bar{a}_r}_n(j-1)$ as weights.

Subsequently, plug $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_n$ into equation (2) and evaluate it for each possible value of $\bar{a}_r$ to give $\bar{Q}^{\bar{a}_r,*}_{t,n}(j)$ for every subject under every possible $\bar{a}_r$.

For $j = t-1, \ldots, 1$:

Step 3: Estimation of $\bar{Q}^{\bar{a}_r}_{t}(j)$

For every $\bar{a}_r \in \bar{\mathcal{A}}_r$ separately, calculate the conditional expectation $\bar{Q}^{\bar{a}_r}_{t}(j) = E(\bar{Q}^{\bar{a}_r,*}_{t}(j+1) \mid \bar{A}(j-1), \bar{\boldsymbol{L}}(j-1), \bar{Y}(j-1))$ and evaluate the obtained estimate $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ at each respective $\bar{a}_r$. Consequently, we obtain, for every subject, a copy of $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ for each different regime $\bar{a}_r$. It follows that $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ is of length $n \times$ *number of regimes*.

Step 4: Update $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ to $\bar{Q}^{\bar{a}_r,*}_{t}(j)$

Construct the covariate $\boldsymbol{h}_j(\bar{a}_r, t)$ as before using equation (1). Then estimate $\boldsymbol{\epsilon}$ by weighted pooled logistic regression over all regimes $\bar{a}_r$ using submodel (2) and update $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ to $\bar{Q}^{\bar{a}_r,*}_{t,n}(j)$ as above.

Final Step:

The above steps provide $\bar{Q}^{\bar{a}_r,*}_{t,n}(1)$ for every t and every $\bar{a}_r$. The estimate of the pooled LTMLE estimator $\boldsymbol{\psi}_{r,n}$ is obtained by solving:

$$\begin{aligned}\boldsymbol{\psi}_{r,n} \quad = \quad & \text{argmax}_{\boldsymbol{\beta}}\, E_n \sum_{t=1}^{K_r+1} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \{\bar{Q}^{\bar{a}_r,*}_{t,n}(1) log(expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t))) \quad + \\ & \qquad (1 - \bar{Q}^{\bar{a}_r,*}_{t,n}(1)) log(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t)))\}.\end{aligned}$$

where E_n is the empirical distribution. In practice, this estimate is obtained by regressing $\bar{Q}_t^{\bar{a}_r,*}(1)$ on $\bar{a}_r$ and t according to the linear specification of the MSM.

3 Data generating algorithm for the simulation

$$L(0) \sim Bern[0.3]$$

$$L(t) = \begin{cases} \text{NA} & \text{if } Y(t) = 1 \\ Bern[expit(-0.75 + log(1.5)L(t-1) + log(0.6)A(t-1)+ & \\ log(1.25)L(t-2) + log(0.8)A(t-2)+ & \text{otherwise} \\ log(1.1)L(t-3) + log(0.9)A(t-3))] & \end{cases}$$

$$A(t) = \begin{cases} \text{NA} & \text{if } Y(t) = 1 \\ 1 & \text{if } A(t-1) = 1 \\ Bern[expit(-3.5 + log(1.75)L(t) + log(1.5)L(t-1) + & \\ log(1.25)L(t-2) + log(1.05)L(t-3))] & \text{otherwise} \end{cases}$$

$$Y(t+1) = \begin{cases} 1 & \text{if } Y(t) = 1 \\ Bern[expit(-4 + log(3.5)L(t) + log(0.75)A(t)+ & \\ log(2.5)L(t-1) + log(0.85)A(t-1) + log(1.5)L(t-2) & \text{otherwise} \\ +log(0.95)A(t-2))] & \end{cases}$$