Implementing Targeted Maximum Likelihood Estimation for Longitudinal Causal Inference based on Stitelman et al. (2012)

Alissa Gordon, Yilong Hou, Kaitlyn Lee, Sylvia Song

Berkeley
UNIVERSITY OF CALIFORNIA

Background

- **Context & Motivation**
  - Many clinical trials focus on time-to-event outcomes (e.g., time to death).
  - Standard methods like Cox models often ignore covariates or cannot handle time-dependent confounding.
  - Informative censoring and time-varying confounders introduce bias, even in randomized trials.
  - IPCW and A-IPCW estimators can be unstable and sensitive to model misspecification.

Background

Limitations of Existing Longitudinal TMLE

- Iterative conditional expectations (ICE) starts with estimating the last conditional expectation first - the hardest function to estimate, since we have the fewest people who followed the treatment through the last time points. Errors propagate as we continue to regress the predictions onto different parts of the past.
- Standard TMLE implementations require integration over all possible covariate histories, which grows exponentially.
- Clever covariates and TMLE updates require computing over entire supports of time-dependent processes.
- Computational cost becomes infeasible as the number of time points and covariates increases.

Background

Goals of the Paper

- The TMLE in this paper fits the entire likelihood, rather than fitting conditional expectation functions.
  - The authors propose discretizing the variables to rewrite the likelihood in terms of conditional hazards.
- Develop a computationally feasible TMLE for treatment-specific survival curves under informative censoring.
- Introduce Markov assumptions, iterative updates, and weighted logistic regression to reduce computation time from exponential to linear.

Algorithm Steps (3 Time Points)

DGP: O = (W,A,L1,C1,L2,C2,Y)

$$
\begin{aligned}
\Psi(P_{U,X}) &= \mathbb{E}_{U,X}[Y_1] \\
&= \sum_{\bar{l}} \left[\mathbb{E}_0 \left[E_0[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}] \right] \times \prod_{t=1}^{K} P_0 \left[L(t)|\bar{A}(t-1)=\bar{a}(t-1), \bar{L}(t-1)=\bar{l}(t-1) \right] \right] \\
&= \sum_{\bar{l}} \mathbb{E}_0 [\underbrace{P(Y=1|L_2=l_2, L_1=l_1, C_2=0, C_1=0, A=1, W)}_{f} \\
&\quad \times \underbrace{P(L_2=l_2|C_1=0, L_1=l_1, A=1, W)}_{g} \\
&\quad \times \underbrace{P(L_1=l_1|A=1, W)}_{h}]
\end{aligned} \tag{1}
$$

Unlike ICE, we update each part f, g, and h separately and then plug them into the g-computation formula above.

Clever Covariate

$$C_{tji}(Q) = \Big\{P\big[L_{a,0}(K+1,1,1) = 1 \mid L(t,j,l) = 1, \mathrm{Pa}(L(t,j,l))\big] - P\big[L_{a,0}(K+1,1,1) = 1 \mid L(t,j,l) = 0, \mathrm{Pa}(L(t,j,l))\big]\Big\} \tag{2}$$

$$C_{tji}(g) = \frac{I(A = a)I(C > t_-)}{g_{A(0)}(1|L(0)) \prod_{s=1}^{t-1} g_{A^c(s)}(0|\mathrm{Pa}(A^c(s)))}. \tag{3}$$

So for our last time point (Y):

- $C_q = P(Y = 1|Y = 1, Pa(Y)) - P(Y = 1|Y = 0, Pa(Y)) = 1$
- $C_g = \frac{I(A=1)I(C_1=0)I(C_2=0)}{g(A)_{pred}*(1-g(C_1)_{pred})*(1-g(C2)_{pred})}$

We fit g_A, g_{C_1}, g_{C_2} using SuperLearner (library SL.glm.interaction, SL.mean) with uncensored cases and predict on our observed data.

Berkeley
UNIVERSITY OF CALIFORNIA

t = 3 (Estimating f)

1. Regress Y on $\{W, A, L_1, L_2\}$ using SuperLearner with uncensored data
2. Predict outcomes under desired histories (with type=link) ($C_1 = 0, C_2 = 0, A = 1, L_1 = (0,1), L_2 = (0,1)$). Since we sum over $\bar{l}$, we predict:
   - $f_{11} = P(Y = 1 | L_2 = 1, L_1 = 1, C_2 = 0, C_1 = 0, A = 1, W)$
   - $f_{10} = P(Y = 1 | L_2 = 1, L_1 = 0, C_2 = 0, C_1 = 0, A = 1, W)$
   - $f_{01} = P(Y = 1 | L_2 = 0, L_1 = 1, C_2 = 0, C_1 = 0, A = 1, W)$
   - $f_{00} = P(Y = 1 | L_2 = 0, L_1 = 0, C_2 = 0, C_1 = 0, A = 1, W)$
3. Target the initial predictions for Y from $f_{\bar{l}}$ using weighted logistic regression with offsets as predictions from $f_{\bar{l}}$, $C_q = 1$ as a covariate and C_g as a weight (clever covariates from the previous slide).
4. Update and store the predictions for Y. There should be four sets of updated predictions: $f^*_{11} \ldots f^*_{00}$.

In R, targeting step looks like: qy.fit.L11.update = glm(Y$\sim$ offset(qy.fit.L11)+1, weights = C_g, family = "quasibinomial", data = L11)

Berkeley
UNIVERSITY OF CALIFORNIA

Update the Clever Covariate

The q factor clever covariate has a nice property in that as we move between time points, the q factor clever covariate for the next timepoint is calculated using the previously updated factors.

For t=2, the q factor of the clever covariate is:
$C_q = P(Y = 1 | L2 = 1, Pa(L2)) - P(Y = 1 | L2 = 0, Pa(L2))$
Note that we have already estimated and targeted $P(Y = 1 | L2 = \bar{l}_2, L1 = \bar{l}_1, A = 1, C1 = 0, C2 = 0)$. Therefore, we can use our fits from the previous steps to create our next clever covariate. We will have two separate q factor clever covariates conditioned on history of time varying covariates (L1):

- When L1=1: $f^*_{11} - f^*_{01}$
- When L1=0: $f^*_{10} - f^*_{00}$

The g-factor of the clever covariate updates to: $\frac{I(A=1)I(C_1=0)}{g(A)_{pred} * (1 - g(C_1)_{pred})}$

Berkeley
UNIVERSITY OF CALIFORNIA

t = 2 (estimating g)

We move on to the L2 part in the g-comp formula

1. Regress L2 on $\{W, A, L1\}$ using Superlearner with uncensored data
2. Predict L2 under desired histories (with type=link) $(C_1 = 0, A = 1, L1 = (0, 1))$. Since we sum over $\bar{l}$, we predict (conditioning on L1, similar to the clever covariates from previous slide):
   - $g_1 = P(L_2 = 1 | C_1 = 0, L_1 = 1, A = 1, W)$
   - $g_0 = P(L_2 = 1 | C_1 = 0, L_1 = 0, A = 1, W)$

   Note: if interested in $P(L_2 = 0 | \ldots)$, it is the complement of g_1 and g_0.
3. Target the initial predictions for L2 using weighted logistic regression with new clever covariates for t=2
4. Update and store the predictions for L2. There should be two sets of predictions from g_1^* and g_0^*.

In R, targeting step looks like: ql2.fit.L1.update = glm(L2$\sim$ offset(ql2.fit.L1) + C_q, weights = C_g, family="quasibinomial", data = L1)

Update the Clever Covariate

For t=1, the q factor of the clever covariate is:
$C_q = P(Y=1|L1=1, Pa(L1)) - P(Y=1|L1=0, Pa(L1))$ Again, we can use our targeted predictions. Using law of total probability and conditional distributions,

$$C_q = f_{11}^* g_1^* + f_{01}^*(1-g_1^*) - (f_{10}^* g_0^* + f_{00}^*(1-g_0^*))$$

At t=1, there is just one q factor clever covariate since there are no more time varying covariates to condition on.
The g-factor of the clever covariate updates to: $\frac{I(A=1)}{g(A)_{pred}}$

Berkeley
UNIVERSITY OF CALIFORNIA

t = 1 (estimate h)

We move on to the L1 part in the g-comp formula to estimate $P(L_1 = 1|A_1 = 1, W)$

1. Regress L1 on $\{W, A\}$ using Superlearner with observed data
2. Predict L1 under desired history (with type=link) (A=1). We predict:
   - $h = P(L_1 = 1|A = 1, W)$
3. Target the initial predictions for L1 using weighted logistic regression with new clever covariates for t=1
4. Update and store the predictions for L1. There should only be one set of predictions from h.

Berkeley
UNIVERSITY OF CALIFORNIA

Estimate of Ψ

Recall:

$$
\begin{aligned}
\Psi(P_{U,X}) &= \mathbb{E}_{U,X}[Y_1] \\
&= \sum_{\bar{l}} \Big[\mathbb{E}_0\left[E_0[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}]\right] \times \prod_{t=1}^{K} P_0\left[L(t)|\bar{A}(t-1)=\bar{a}(t-1), \bar{L}(t-1)=\bar{l}(t-1)\right]\Big] \\
&= \sum_{\bar{l}} \underbrace{P(Y=1|L_2=l_2, L_1=l_1, C_2=0, C_1=0, A=1, W)}_{f} \\
&\quad \times \underbrace{P(L_2=l_2|C_1=0, L_1=l_1, A=1, W)}_{g} \\
&\quad \times \underbrace{P(L_1=l_1|A=1, W)}_{h}
\end{aligned} \tag{4}
$$

Now we have all of the parts for the g-comp formula under all $\bar{l}$ and we can calculate the full density, and then take the expectation for our target parameter Y_1.

Specifically, we have:

$$\hat{\Psi} = E[f^*_{11}g^*_1h^* + f^*_{10}g^*_0(1-h^*) + f^*_{01}(1-g^*_1)h^* + f^*_{00}(1-g^*_0)(1-h^*)]$$

Berkeley
UNIVERSITY OF CALIFORNIA

Simulation Design

- Three data-generating processes (DGPs)
  1. **DGP 1**: low independent censoring (10% at each visit)
  2. **DGP 2**: high independent censoring (50% at each visit)
  3. **DGP 3**: informative censoring $\Pr(C_t = 1) = \text{logit}^{-1}(-3 + 0.8A + 0.3W + 0.25L_{1:t})$
- Target parameter: $\psi = \mathbb{E}[Y^{A=1}]$
- 200 simulations for each with **seeds <- 2025 + 1:200**
- Estimators compared
  - **G-computation**
  - **IPTW**
  - **TMLE**
  - **Ori**

Results – DGP 1

Low independent censoring

Estimator	Bias	Variance	MSE
G-comp	-0.0159	0.0005	0.0008
TMLE	-0.0045	0.0005	0.0005
IPTW	-0.0045	0.0005	0.0005
Ori	-0.0045	0.0005	0.0005

Results – DGP 2

High independent censoring

Estimator	Bias	Variance	MSE
G-comp	-0.0489	0.0024	0.0048
TMLE	-0.0094	0.0021	0.0022
IPTW	-0.0093	0.0020	0.0021
Ori	-0.0090	0.0020	0.0021

Results – DGP 3
Informative censoring

Estimator	Bias	Variance	MSE
G-comp	-0.0197	0.0006	0.0010
TMLE	-0.0071	0.0006	0.0007
IPTW	-0.0071	0.0006	0.0007
Ori	-0.0060	0.0006	0.0006