Proximal Survival Analysis for Dependent Left Truncation

Yuyao Wang

Department of Mathematics, UC San Diego
yuw079@ucsd.edu

Joint work with:

Andrew Ying, Google Inc.

Ronghui (Lily) Xu, UC San Diego

Left truncation – selection due to delayed entry

- Example: aging studies.
  - Age is the time scale of interest.
  - Subjects enrolled at various ages instead of at the time origin (time at birth).

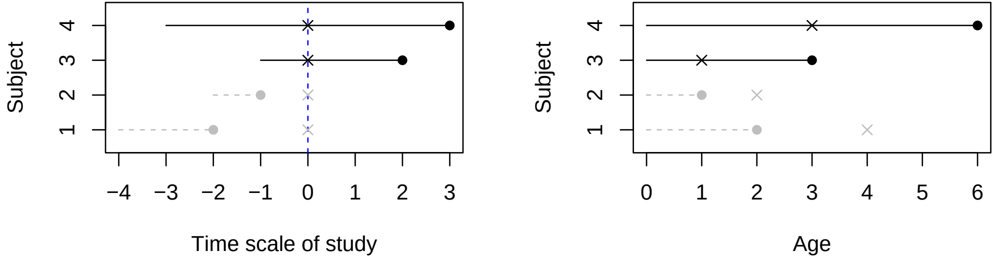


Figure: A toy example for aging study; '×' - enrollment times; dots - times to events.

Left truncation – mathematical formulation

- Time-to-event: T^*
- Left truncation time: Q^* – usually the study enrollment time

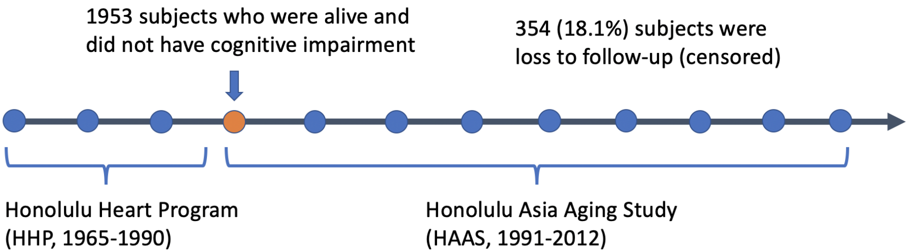
- T^* is **left truncated** by Q^* if only subjects with $T^* > Q^*$ are included in the data.

- Subjects with early event times tend not to be captured
  $\rightarrow$ **biased sample** $\rightarrow$ **selection bias**

Examples:

- Aging studies – age is the time scale of interest
- Pregnancy studies
- Some cancer survivorship studies, e.g., SJLIFE.

HAAS data



- T^*: age to moderate cognitive impairment or death;
- Q^*: age at entry of HAAS.

Literature on left truncation

Conventional methods on left truncation

- Under random left truncation/quasi-independence assumption
- Extended to conditional (quasi-)independence assumption under regression settings
  - when the dependence-inducing covariates are included as regressors

Y. Wang, A. Ying, and R. Xu (2024a). Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika* 111: 789–808.

Y. Wang, A. Ying, and R. Xu (2024b). A liberating framework from truncation and censoring, with application to learning treatment effects. *arXiv:2411.18879*.

Literature on left truncation

Conventional methods on left truncation

- Under random left truncation/quasi-independence assumption
- Extended to conditional (quasi-)independence assumption under regression settings
  - when the dependence-inducing covariates are included as regressors

For marginal estimands under covariate dependent left truncation

- Inverse probability of truncation weighting (Vakulenko-Lagun et al., 2022)
- Efficient influence function-based doubly robust (DR) approaches (Wang et al., 2024a)
- A general Neyman orthogonal and doubly robust framework for handling covariate dependent LTRC. (Wang et al., 2024b)

Y. Wang, A. Ying, and R. Xu (2024a). Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika* 111: 789–808.

Y. Wang, A. Ying, and R. Xu (2024b). A liberating framework from truncation and censoring, with application to learning treatment effects. *arXiv:2411.18879*.

Literature on left truncation

Conventional methods on left truncation

- Under random left truncation/quasi-independence assumption
- Extended to conditional (quasi-)independence assumption under regression settings
  - when the dependence-inducing covariates are included as regressors

For marginal estimands under covariate dependent left truncation

- Inverse probability of truncation weighting (Vakulenko-Lagun et al., 2022)
- Efficient influence function-based doubly robust (DR) approaches (Wang et al., 2024a)
- A general Neyman orthogonal and doubly robust framework for handling covariate dependent LTRC. (Wang et al., 2024b)

! There may be unmeasured latent factors that induce the dependence.
- e.g. HAAS: overall health status, socioeconomic status, health seeking behavior.

Y. Wang, A. Ying, and R. Xu (2024a). Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika* 111: 789–808.

Y. Wang, A. Ying, and R. Xu (2024b). A liberating framework from truncation and censoring, with application to learning treatment effects. *arXiv:2411.18879*.

Proximal Causal Inference - Review

- Proximal causal inference for handling unmeasured confounding
  - Point exposure
  - Longitudinal studies
  - Mediation analysis
  - ...

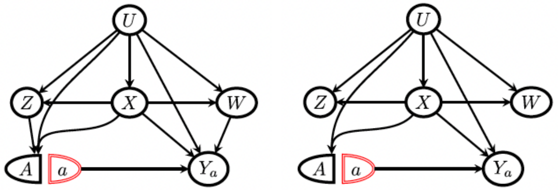


Figure: Single world intervention graphs with point exposure and proxies. (Tchetgen Tchetgen et al., 2024)

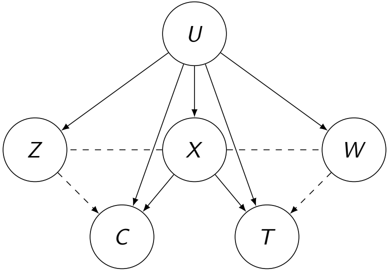
- Consistency, latent positivity
- $(W, Y_a) \perp\!\!\!\perp (A, Z) \mid (U, X)$

- Existence of outcome confounding bridge function + completeness $\rightarrow$ proximal g-formula
- Existence of treatment confounding bridge function + completeness $\rightarrow$ proximal IPW
- Doubly robust identification

E.J. Tchetgen Tchetgen, et al. (2024). An introduction to proximal causal inference. *Statistical Science* 39(3): 375-390.

Proximal Survival Analysis - Review

- Proximal survival analysis to handle dependent right censoring



- Proximal independence
- Latent positivity

- Existence of bridge processes
  + proxy relevance (completeness)

$\Longrightarrow$
- proximal event-inducing identification
- proximal censoring-inducing identification
- doubly robust identification

Ying, A. (2024). Proximal survival analysis to handle dependent right censoring. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86(5): 1414–1434.

Our contributions

- Propose a proximal weighting identification framework for dependent left truncation
  - measured covariates may only serve as proxies for explaining the dependence
- Extend the framework to also handling right censoring by incorporating IPCW.
- Construct estimators that are shown to be CAN when conditions are met.
- Enriches the proximal inference literature - selection bias (biased sampling)

Notation and Estimand

- With '*' – in truncation-free data; without '*' – in truncated data;
- T - Event time of interest; Q - left truncation time; C - right censoring time;
- U - unmeasured latent factor; W_1, W_2, Z – Measured covariates.
- Observe $O = (Q, X, \Delta, W_1, W_2, Z)$, where $X = \min(T, C)$ and $\Delta = \mathbb{1}(T < C)$.

- **Estimand:**

$$\theta = E\{\nu(T^*)\},$$

- ν: a known transformation satisfying $\nu(t) = \nu(t_0)$ for all $t \geq t_0$.

  - e.g., $\nu(t) = \mathbb{1}(t > t_0) \implies \theta = \mathbb{P}(T^* > t_0)$ (survival probability).
  - e.g., $\nu(t) = \min(t, t_0) \implies \theta = E\{\min(T^*, t_0)\}$ (restricted mean survival time, RMST).

Notation and Estimand

- With '*' – in truncation-free data; without '*' – in truncated data;
- T - Event time of interest; Q - left truncation time; C - right censoring time;
- U - unmeasured latent factor; W_1, W_2, Z – Measured covariates.
- Observe $O = (Q, X, \Delta, W_1, W_2, Z)$, where $X = \min(T, C)$ and $\Delta = \mathbb{1}(T < C)$.

- **Estimand:**

$$\theta = E\{\nu(T^*)\},$$

- ν: a known transformation satisfying $\nu(t) = \nu(t_0)$ for all $t \geq t_0$.
    - e.g., $\nu(t) = \mathbb{1}(t > t_0) \implies \theta = \mathbb{P}(T^* > t_0)$ (survival probability).
    - e.g., $\nu(t) = \min(t, t_0) \implies \theta = E\{\min(T^*, t_0)\}$ (restricted mean survival time, RMST).

- **First consider no right censoring**

Assumptions

1. **Proximal independence**: $(W_1^*, Q^*) \perp\!\!\!\perp (W_2^*, T^*) \mid Z^*, U^*$.

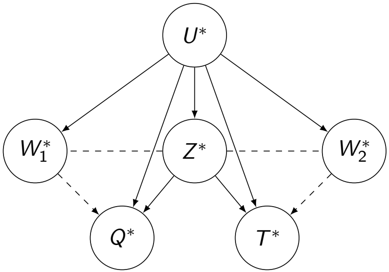


Figure: An example.

HAAS data:

- U^*: socioeconomic status, overall health status, health seeking behavior
- W_1^*: grip strength
- W_2^*: education, alcohol consumption, cigarettes consumption
- Z^*: *APOE* genotype, systolic blood pressure, heart rate

2. **Latent positivity**: $\mathbb{P}(Q^* < T^* \mid T^*, U^*, Z^*) > 0$ almost surely.

Assumptions

3. **Existence of an truncation-inducing bridge process**:

   There exists a bounded $b(t, W_1, Z)$ satisfying

   $$E\{db(t, W_1, Z) - d\bar{N}_Q(t)b(t, W_1, Z) \mid Q \leq t < T, W_2, Z\} = 0, \qquad (1)$$

   with initial condition $b(t, W_1, Z) = 1$ for all $t \geq \tau$ (the maximum of the support of Q^*), where $\bar{N}_Q(t) = I(t \leq Q < T)$.

   Note: (1) is equivalent to

   $$E\left[\int_Q^T \varphi(t, W_2, Z)\{db(t, W_1, Z) - d\bar{N}_Q(t)b(t, W_1, Z)\}\right] = 0$$

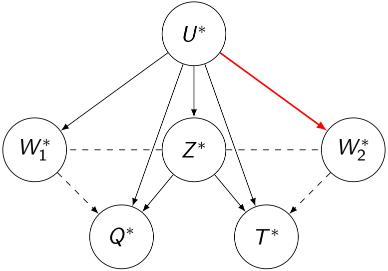
   for any integrable function $\varphi(t, W_2, Z)$.

Assumptions

4. **Completeness**: For any $t > 0$ and any integrable function ζ,

$$E\left[\zeta(t, U, Z) \mid Q \leq t < T, W_2, Z\right] = 0$$

if and only if $\zeta(t, U, Z) = 0$ a.s..



- W_2 has sufficient variability relative to U.
  - e.g., when both are categorical, the number of categories of $W_2 \geq$ that of U.
- Rules out conditional independence between U^* and W_2^* conditional on Z^* and being at risk.

Proximal truncation-inducing identification

- Under Assumptions 1 - 4, for any truncation-inducing bridge process $\{b(t, W_1, Z) : t \geq 0\}$ satisfying (1) and the initial condition, we have

$$E\{db(t, W_1, Z) - d\bar{N}_Q(t)b(t, W_1, Z) \mid Q \leq t < T, U, Z\} = 0,$$

  and

$$\theta = \frac{E\{b(T, W_1, Z)\nu(T)\}}{E\{b(T, W_1, Z)\}}.$$

Proximal truncation-inducing identification

- Under Assumptions 1 - 4, for any truncation-inducing bridge process $\{b(t, W_1, Z) : t \geq 0\}$ satisfying (1) and the initial condition, we have

$$E\{db(t, W_1, Z) - d\bar{N}_Q(t)b(t, W_1, Z) \mid Q \leq t < T, U, Z\} = 0,$$

and

$$\theta = \frac{E\{b(T, W_1, Z)\nu(T)\}}{E\{b(T, W_1, Z)\}}.$$

- A special case: Covariate-dependent left truncation $T^* \perp\!\!\!\perp Q^* \mid Z^*$ (i.e., $U^* = \varnothing$, $W_1^* = \varnothing$, $W_2^* = \varnothing$)
  $\rightarrow$ $b(t, Z) = 1/G(t|Z)$ satisfies (1), where $G(t|z) = P(Q^* \leq t|Z^* = z)$.
  $\rightarrow$ Inverse probability of truncation weighting.

Estimation

- Estimation for θ – With an i.i.d. sample of size n,

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{b}(T_i, W_{1i}, Z_i)\nu(T_i)}{\sum_{i=1}^{n} \hat{b}(T_i, W_{1i}, Z_i)}.$$

Estimation

- Estimation for θ – With an i.i.d. sample of size n,

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{b}(T_i, W_{1i}, Z_i)\nu(T_i)}{\sum_{i=1}^{n} \hat{b}(T_i, W_{1i}, Z_i)}.$$

- Estimation for b – Consider semiparametric model

$$b(t, W_1, Z; B(t)) = \exp\{B_0(t) + W_1 B_1(t) + ZB_z(t)\},$$

with the initial condition $B_0(\tau) = B_1(\tau) = B_z(\tau) = 0$, where
$B(t) = (B_0(t), B_1(t)^\top, B_z(t)^\top)^\top$.

$\rightarrow$ Closed form solution for $B(t)$ by solving the estimating equation backwards in time.

Estimation

Taking $\varphi(t, W_2, Z) = (1, W_2, Z)$ and plugging in the model for b into

$$E\left[\int_Q^T \varphi(t, W_2, Z)\{db(t, W_1, Z) - d\bar{N}_Q(t)b(t, W_1, Z)\}\right] = 0.$$

$\implies$ Closed-form solution for $B(t)$:

$$\hat{B}(t) = -\frac{1}{n}\int_t^\tau \mathbb{M}_B(t)^\dagger \sum_{i=1}^n \mathbb{1}(Q_i \leq t < T_i)\exp\{(1, W_{1i}, Z_i)\hat{B}(t+)\} \cdot (1, W_{2i}, Z_i)^\top d\bar{N}_{Qi}(t),$$

where $\dagger$ denote the Moore-Penrose inverse (pseudo-inverse) of a matrix, and

$$\mathbb{M}_B(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Q_i \leq t < T_i)\exp\{(1, W_{1i}, Z_i)\hat{B}(t+)\} \cdot (1, W_{2i}, Z_i)^\top (1, W_{1i}, Z_i).$$

Extension for right censoring

$D = C - Q$ residual censoring time.
Recall τ - maximum support of Q^*; $\nu(t) = \nu(t_0)$ for all $t \geq t_0$.
Denote $t^* = t_0 \vee \tau$.

Assumptions:

- Independent residual censoring: $D \perp\!\!\!\perp (Q, T, W_1, W_2, Z)$.
- Positivity: $S_D(t^*) > 0$.

Estimation:

Extension for right censoring

$D = C - Q$ residual censoring time.
Recall τ - maximum support of Q^*; $\nu(t) = \nu(t_0)$ for all $t \geq t_0$.
Denote $t^* = t_0 \vee \tau$.

Assumptions:

- Independent residual censoring: $D \perp\!\!\!\perp (Q, T, W_1, W_2, Z)$.
- Positivity: $S_D(t^*) > 0$.

Estimation:

- For θ: incorporate IPCW weights:
$$\hat{\theta}_c = E\left\{\frac{\Delta(t^*)b(X, W_1, Z)\nu(X)}{S_D(X \wedge t^* - Q)}\right\} \Bigg/ E\left\{\frac{\Delta(t^*)b(X, W_1, Z)}{S_D(X \wedge t^* - Q)}\right\},$$
where $\Delta(t) = I(\Delta = 1 \text{ or } X > t) = I(T \wedge t < C)$.

- For b: under no censoring, the estimating equation involves $\mathbb{1}(Q \leq t < T)$
$\rightarrow$ incorporate time-varying IPCW weights:
$$\Delta(t)/\hat{S}_D(X \wedge t - Q).$$

Asymptotics

Assumptions on $\hat{b}$:

- Consistency: $\|\hat{b}(T, W_1, Z) - b_0(T, W_1, Z)\|_1 = o(1)$.
- Asymptotically linearity: the estimator $\hat{b}(T, W_1, Z)$ is asymptotically linear.

$$\left\| \hat{b}(T, W_1, Z) - b_0(T, W_1, Z) - \frac{1}{n}\sum_{i=1}^{n} \xi(T, W_1, Z; O_i) \right\|_1 = o(n^{-1/2}).$$

- Regularity conditions: $\hat{b}(T, W_1, Z)$ bounded a.s.; $\hat{b}(t, W_1, Z) = \hat{b}(\tau, W_1, Z)$ for all $t \geq \tau$.

Then

- Consistency: $\hat{\theta}_c \xrightarrow{p} \theta_0$;
- Asymptotic normality: $\sqrt{n}(\hat{\theta}_c - \theta_0) \xrightarrow{d} N(0, \sigma^2)$.

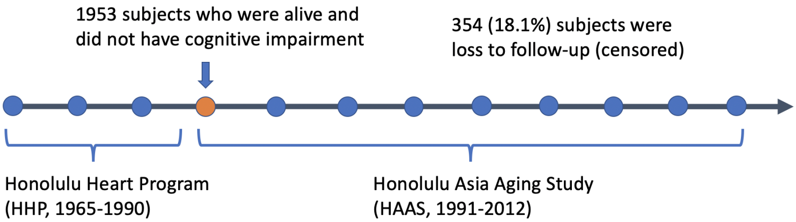
Simulation

Table: Simulation results for different estimators under the case with right censoring. Each observed data set has sample size 500 or 1000, and 500 data sets are simulated. Truncation rate 47%; censoring rate 37%. Estimand $\theta = \mathbb{P}(T^* > 1) = 0.4632$.

	$n = 500$				$n = 1000$			
Method	Bias	SD	bootSE	CP	Bias	SD	bootSE	CP
PQB	**-0.0124**	0.0745	0.0605	**0.926**	**-0.0062**	0.0395	0.0393	**0.944**
IPQW	0.0168	0.0426	0.0391	0.906	0.0172	0.0284	0.0278	0.882
PQB-cw	-0.0209	0.0858	0.0710	0.928	-0.0565	0.0623	0.0558	0.872
IPQW-cw	0.0116	0.0479	0.0435	0.906	-0.0325	0.0379	0.0355	0.832
PL	0.0710	0.0312	0.0293	0.324	0.0707	0.0208	0.0207	0.072
KM	0.2633	0.0222	0.0203	0.000	0.2631	0.0145	0.0144	0.000
naive	0.1889	0.0222	0.0212	0.000	0.1891	0.0157	0.0150	0.000

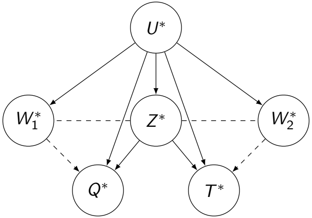
'-cw': Estimators that incorporate case weights $\Delta / S_D(X)$ for handling right censoring (require stronger positivity assumption)

HAAS data

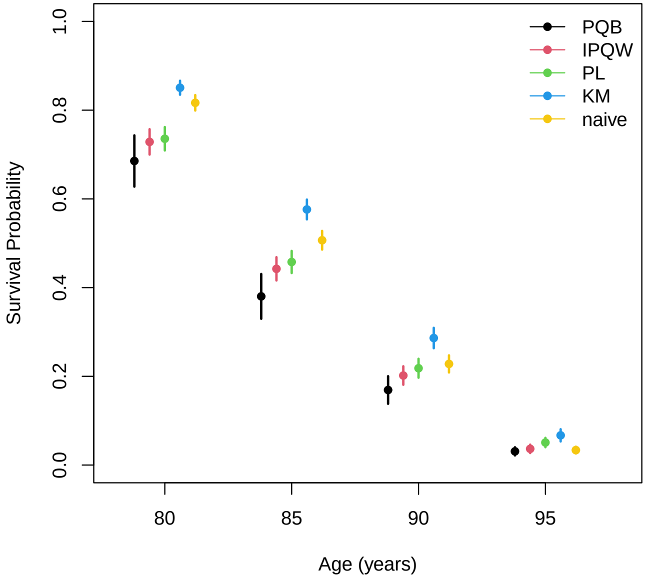


- T^*: age to moderate cognitive impairment or death
- Q^*: age at entry of HAAS – minimum Q_i's: 71.3
- Conditional Kendall's tau test: p-value $= 0.0036$.

- U^*: socioeconomic status, overall health status, health seeking behavior
- W_1^*: grip strength
- W_2^*: education (≤ 12 years/otherwise), alcohol consumption (heavy/non-heavy), cigarettes consumption (yes/no)
- Z^*: *APOE* genotype, systolic blood pressure, heart rate



HAAS data



PQB
IPQW
PL
KM
naive
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
80
85
90
95
Age (years)

Summary

- We have developed a proximal weighting identification framework for handling dependent left truncation, in the presence of unmeasured dependence-inducing covariates.

Future direction

- Extend the framework to handle dependent right censoring.
- Nonparametric approaches for estimating the bridge function.
- Explore event-inducing identification and doubly robust identification.

Reference

- Yuyao Wang, Andrew Ying, Ronghui Xu. Proximal survival analysis for dependent left truncation. *(In preparation)*

  A version in: Wang, Y. (2025), Towards robust and efficient estimation under dependent left truncation (Chapter 3), PhD thesis, University of California San Diego.