A framework for leveraging machine learning tools to estimate personalized survival curves

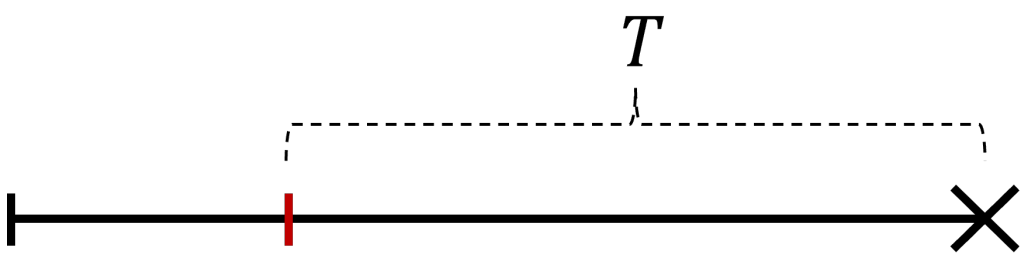
Charlie Wolock, Peter Gilbert, Noah Simon, Marco Carone

March 22, 2023

Background

In this project, we study T, the time between an initiating event and a terminating event.

- Example: time from disease onset to death



T may be right-censored due to loss to follow-up, end of study, etc.

Full and observed data

The full-data world:

X	covariates
T	time to event
C	time to censoring

The observed-data world:

X	covariates
$Y := \min\{T, C\}$	observed follow-up time
$\Delta := \mathbb{1}(T \leq C)$	event indicator

Machine learning for conditional survival curves

Goal: Use machine learning tools to estimate the conditional survival function of T subject to right censoring.

Desired properties:

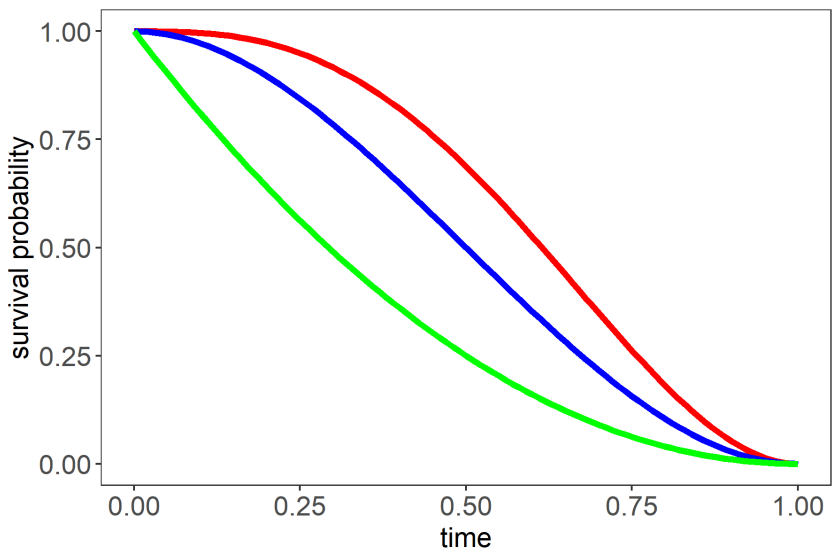
1. Incorporate off-the-shelf machine learning methods (not adapted for censoring)

2. Estimate the entire survival function over some interval (not just at a single time-point)

The conditional survival function

Our goal is to estimate $S(\tau \mid x) := P(T > \tau \mid X = x)$. Why is this quantity of interest?

Personalized predictions



Nuisance parameter in non- and semiparametric problems

$$\mathbb{E}\left[S(\tau \mid X)\right]$$

Loss functions

Machine learning methods rely on loss functions for

1. Optimization: e.g., gradient boosting, neural nets
2. Tuning parameter selection: comparing predictions to observations

For estimating $S(\tau \mid x)$, we might use familiar squared-error loss:

$$L(x, t, \theta) = \{\mathbb{1}(t > \tau) - \theta(x)\}^2$$

In fact, the minimizer of this loss is $\theta(x) = S(\tau \mid x)$. Unfortunately:

- We can't evaluate this because we don't observe T.
- This only targets a single time τ, rather than an entire survival curve.

Estimation at a single time point

What if there were no censoring? Then $S(\tau \mid x)$ can be viewed as a binary regression (or classification) problem.

$$\left(\begin{array}{cc} X & T \\ \hline X_1 & T_1 \\ X_2 & T_2 \\ X_3 & T_3 \\ \vdots & \vdots \\ X_n & T_n \end{array}\right) \longrightarrow \left(\begin{array}{cc} X & \text{outcome} \\ \hline X_1 & \mathbb{1}(T_1 > \tau) \\ X_2 & \mathbb{1}(T_2 > \tau) \\ X_3 & \mathbb{1}(T_3 > \tau) \\ \vdots & \vdots \\ X_n & \mathbb{1}(T_n > \tau) \end{array}\right)$$

Estimation on a grid (global stacking)

Choose a time grid $\mathcal{C} := \{\tau_1, \ldots, \tau_J\}$.

$$\begin{pmatrix} X & T \\ \hline X_1 & T_1 \\ X_2 & T_2 \\ X_3 & T_3 \\ \vdots & \vdots \\ X_n & T_n \end{pmatrix} \xrightarrow{\text{stack}} \begin{pmatrix} X & \text{time} & \text{outcome} \\ \hline X_1 & t_1 & \mathbb{1}(T_1 > \tau_1) \\ X_2 & t_1 & \mathbb{1}(T_2 > \tau_1) \\ \vdots & \vdots & \vdots \\ X_n & t_1 & \mathbb{1}(T_n > \tau_1) \\ X_1 & t_2 & \mathbb{1}(T_1 > \tau_2) \\ X_2 & t_2 & \mathbb{1}(T_2 > \tau_2) \\ \vdots & \vdots & \vdots \\ X_n & t_2 & \mathbb{1}(T_n > \tau_2) \\ X_1 & t_3 & \mathbb{1}(T_1 > \tau_3) \\ X_2 & t_3 & \mathbb{1}(T_2 > \tau_3) \\ \vdots & \vdots & \vdots \end{pmatrix}$$

A discrete approach (local stacking)

Alternatively, we could treat T as discrete, such that it can only take values in $\mathcal{C}$.

$$P(T > \tau \mid X = x) = \prod_{\tau_i < \tau} \{1 - P(T = \tau_i \mid T > \tau_{i-1}, X = x)\}$$

Each of these probabilities can be estimated using binary regression, or can be estimated jointly by stacking the data matrices.

Unlike in the previous approach, the choice of $\mathcal{C}$ determines the number of events in each time bin.

- Coarse grid: more events in each bin, but potential loss of information since all events in same bin are treated equally
- Fine grid: fewer events in each bin, more difficult estimation problem

Loss functions under censoring

Possible solutions to the censoring problem:

- Adapt the loss function to account for censoring.

$$L(x, y, \delta, \theta) = \frac{\delta}{P(C > y \mid X = x)} \{\mathbb{1}(y > \tau) - \theta(x)\}^2$$

- Use an loss that doesn't depend on actual event times (e.g., the negative Cox partial likelihood).

Hazards

The conditional hazard function $\lambda(\tau \mid x)$ is the instantaneous event rate at time τ conditional on $X = x$.

$$\lambda(\tau \mid x) = \lim_{\epsilon \to 0} \frac{P(\tau \leq T \leq \tau + \epsilon \mid T \geq \tau, X = x)}{\epsilon}$$

The cumulative hazard is $\Lambda(\tau \mid x) = \int_0^\tau \lambda(u \mid x) du$.

The hazard and survival functions are linked via the product integral:

$$S(\tau \mid x) = \prod_{u \in (0,\tau]} \{1 - \Lambda(du \mid x)\}$$

Hazards

Hazards allow us to identify the survival function in the presence of conditionally independent right censoring.

In the discrete case,

$$P(T = \tau \mid T \geq \tau, X = x) = P(Y = \tau, \Delta = 1 \mid Y \geq \tau, X = x).$$

Therefore, the discrete local stacking approach is still valid, even with censoring.

Hazards

However, we don't need to artificially discretize time. It turns out we can decompose the cumulative hazard conveniently as

$$\Lambda(du \,|\, x) = \frac{\pi(x) F_{Y,1}(du \,|\, x)}{\pi(x)\{1 - F_{Y,1}(u \,|\, x)\} + \{1 - \pi(x)\}\{1 - F_{Y,0}(u \,|\, x)\}}$$

Three components to estimate:

- $\pi(x) := P(\Delta = 1 \,|\, X = x)$ conditional event probability
- $F_1(u \,|\, x) := P(Y \leq u \,|\, \Delta = 1, X = x)$ conditional CDF of Y among the uncensored
- $F_0(u \,|\, x) := P(Y \leq u \,|\, \Delta = 0, X = x)$ conditional CDF of Y among the censored

CDF estimation

Same as before, but this time we stratify on Δ:

$$\begin{pmatrix} X & \Delta & Y \\ \hline X_1 & 1 & Y_1 \\ X_2 & 0 & Y_2 \\ X_3 & 1 & Y_3 \\ \vdots & \vdots & \vdots \\ X_n & 1 & Y_n \end{pmatrix} \xrightarrow[\Delta=1]{\text{filter on}} \begin{pmatrix} X & \Delta & Y \\ \hline X_1 & 1 & Y_1 \\ X_3 & 1 & Y_3 \\ \vdots & \vdots & \vdots \\ X_n & 1 & Y_n \end{pmatrix} \xrightarrow{\text{stack}} \begin{pmatrix} X & \text{time} & \text{outcome} \\ \hline X_1 & t_1 & \mathbb{1}(Y_1 \leq t_1) \\ X_3 & t_1 & \mathbb{1}(Y_3 \leq t_1) \\ \vdots & \vdots & \vdots \\ X_n & t_1 & \mathbb{1}(Y_n \leq t_1) \\ X_1 & t_2 & \mathbb{1}(Y_1 \leq t_2) \\ X_3 & t_2 & \mathbb{1}(Y_3 \leq t_2) \\ \vdots & \vdots & \vdots \\ X_n & t_2 & \mathbb{1}(Y_n \leq t_2) \\ X_1 & t_3 & \mathbb{1}(Y_1 \leq t_3) \\ X_3 & t_3 & \mathbb{1}(Y_3 \leq t_3) \\ \vdots & \vdots & \vdots \end{pmatrix}$$

Computational concerns

The stacked matrix for CDF estimation can be quite large, with dimension depending on sample size and the time grid $\mathcal{C}$.

- Time and memory usage are potential issues.

Solution: Adopt stochastic gradient descent.

1. Mini-batch over sample indices $\{1, \ldots, n\}$.
2. Mini-batch over times in the grid $\mathcal{C} = \{\tau_1, \ldots, \tau_J\}$.

Extra information

- Global survival stacking implemented in `R` package `survML`: `https://github.com/cwolock/survML`.
- Manuscript available on arXiv [Wolock et al., 2022].

References

Craig, E., Zhong, C., and Tibshirani, R. (2021).
Survival stacking: casting survival analysis as a classification problem.
arXiv:2107.13480.

van der Laan, M. J. and Rose, S. (2011).
Targeted Learning: Causal Inference for Observational Data.
Springer.

Westling, T., Luedtke, A., Gilbert, P. B., and Carone, M. (2021).
Inference for treatment-specific survival curves using machine learning.
arXiv:2106.06602.

Wolock, C. J., Gilbert, P. B., Simon, N., and Carone, M. (2022).
A framework for leveraging machine learning tools to estimate personalized survival curves.
arXiv:2211.03031.

Extra slides

Local survival stacking

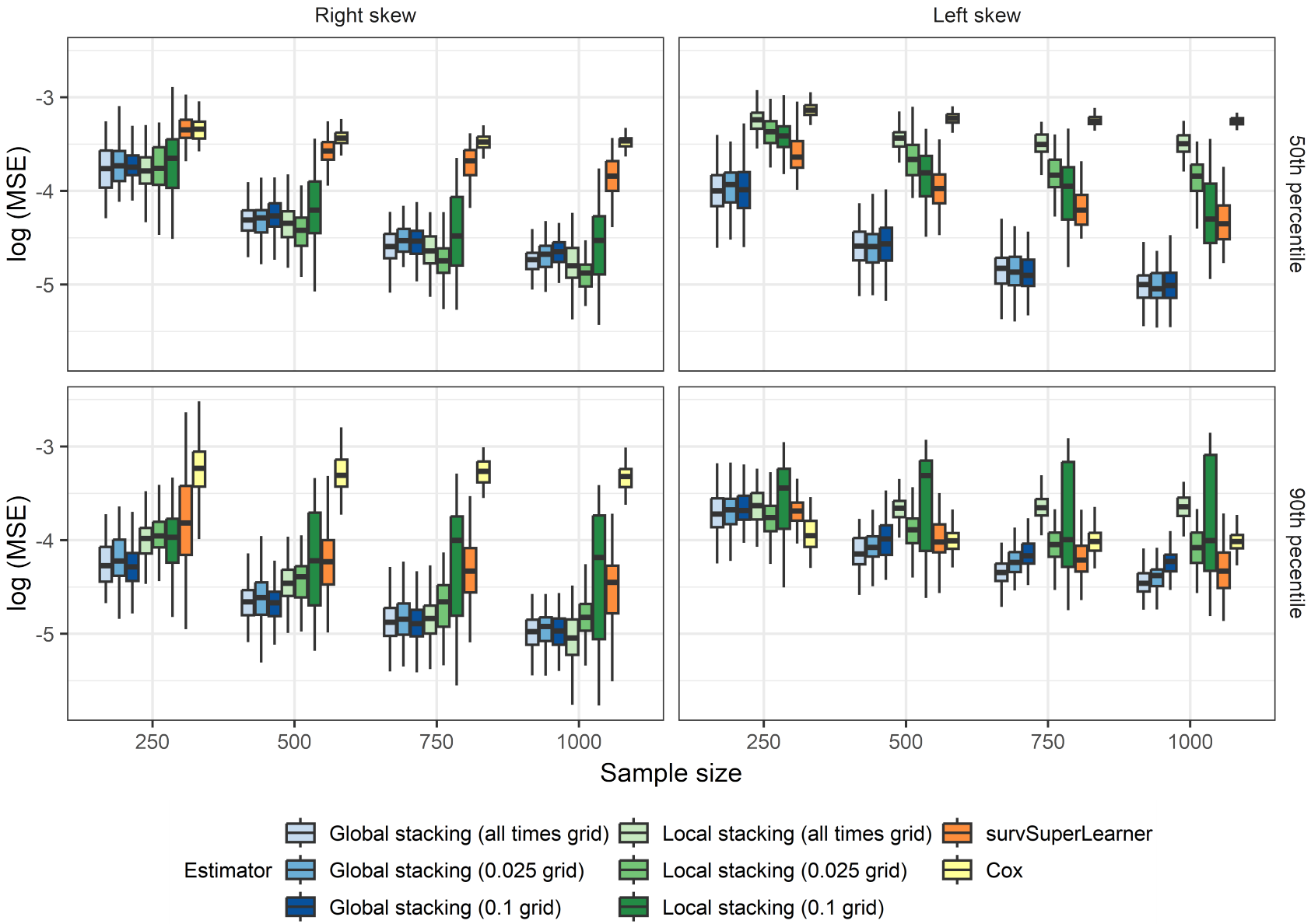
$$\begin{pmatrix} X & \Delta & Y \\ \hline X_1 & 1 & Y_1 \\ X_2 & 0 & Y_2 \\ X_3 & 1 & Y_3 \\ \vdots & \vdots & \vdots \\ X_n & 1 & Y_n \end{pmatrix} \xrightarrow{\text{stack}} \begin{pmatrix} X & \text{time} & \text{outcome} \\ \hline X_1 & Y_1 & 1 \\ X_2 & Y_1 & 0 \\ \vdots & \vdots & \vdots \\ X_n & Y_1 & 0 \\ X_3 & Y_3 & 1 \\ X_4 & Y_3 & 0 \\ \vdots & \vdots & \vdots \\ X_n & Y_3 & 0 \\ \vdots & \vdots & \vdots \end{pmatrix}$$

Simulations: setup

Compare the performance of the following:

1. Proposed method (global survival stacking): Using Super Learner for binary regression, three time grids (fine, medium, coarse)
2. Discrete hazards (local) survival stacking: [van der Laan and Rose, 2011, Craig et al., 2021] Using Super Learner for binary regression, three time grids (fine, medium, coarse)
3. `survSuperLearner`: [Westling et al., 2021] Ensemble method for survival-specific estimators (Cox, Kaplan-Meier, parametric regression, random survival forest)
4. Cox proportional hazards model

Simulation results



STEP trial

- Phase IIB trial for Ad5 HIV vaccine, 1,836 individuals assigned male sex at birth in Central and South America
- Some evidence of increased risk of infection among vaccine recipients, particularly among those who were (1) uncircumcised or (2) had baseline Ad5 neutralizing antibodies