PhD Mini-Task

Ryan Lucas

January 2023

Introduction: SurVITE

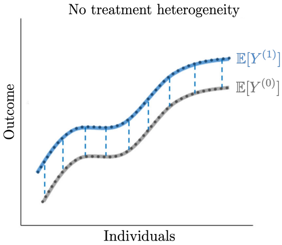
- SurVITE (Curth, Lee, van der Schaar, 2021) is a new framework for estimating **Heterogenous Treatment Effects (HTEs)** in **time-to-event (TTE)** data.
- **HTEs** measure variability in a treatment response for *individuals* within a population.
- **TTE** data records, among other things, the length of time until the occurrence of a particular end-point of interest (e.g. death in a medical study).

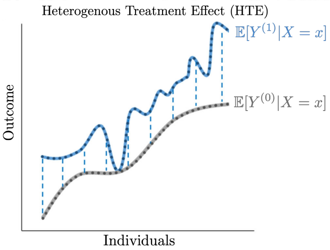
A primer on regular HTE estimation

- For HTE estimation in standard observational data, not dependent on time, the goal is often to measure the conditional average treatment effect (CATE).

$$\mathrm{HTE}_{\mathrm{CATE}}(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]$$

Knowing the patient's characteristics, should we prescribe treatment $a = 1$ or treatment $a = 0$? Importantly, the CATE depends on the unique patient covariates $X = x$:





HTE estimation in TTE data

- In a TTE setting, we are instead interested in several *time-dependent* HTEs, for example the difference in expected survival probability under treatment $a = 1$ and $a = 0$:

$$\mathrm{HTE}_{\mathrm{surv}}\,(\tau \mid x) = S^{(1)}(\tau \mid x) - S^{(0)}(\tau \mid x)$$

- Here $S^{(a)}$ is computable using the treatment-specific hazard function:

$$\underbrace{\lambda^{(a)}(\tau \mid x)}_{\substack{\text{hazard function} \\ \text{for treatment group } a}} = \underbrace{\mathbb{P}(T = \tau \mid T \geq \tau, \mathrm{do}(A = a, C \geq \tau), X = x)}_{\substack{\text{Probability event occurs at time } \tau \\ \text{given still at-risk \& patient characteristics}}}$$

$$\implies S^{(a)}(\tau \mid x) = \prod_{t \leq \tau} (1 - \lambda^{(a)}(t \mid x))$$

- For this reason, $\lambda^{(a)}(\tau \mid x)$ is the main quantity we seek to estimate.

Isn't that just a standard classification problem?

- Notice that we want to make this prediction of the hazard $\hat{\lambda}^{(a)}(\tau \mid x)$ at baseline (the outset).

- When a new patient comes along, we can then easily compute $\hat{S}^{(1)}(\tau \mid x) - \hat{S}^{(0)}(\tau \mid x)$ and make a decision to administer $a = 1$ or $a = 0$ based on the difference in survival probability.

- To find $\hat{\lambda}^{(a)}(\tau \mid x)$, we would like to solve the target problem:

$$\hat{\lambda}^{(a)}(\tau \mid x) \in \underbrace{\arg\min_{h \in \mathcal{H}} \mathbb{E}_{X, Y(\tau) \sim \mathbb{P}_0(X, Y(\tau)),} \left[\ell\left(Y(\tau), h(X)\right)\right]}_{\text{Target Problem}}$$

  which would be a standard classification problem with $Y(\tau) = \mathbb{1}[T = \tau]$ as the target and $\mathbb{P}_0(\cdot)$ the joint distribution of patients at baseline. Think of $\mathbb{P}_0(\cdot)$ as the overall population distribution. [1]

1

- If in addition the assigned treatment are independent of the covariates, i.e. $\mathbb{P}(A = 1 \mid X = x) = \mathbb{P}(A = 0 \mid X = x)$, then this target problem would be correctly specified.

What makes HTE estimation in TTE data special?

- However, due to the nature of TTE data, certain patients drop out of the at-risk population over time (recall that $\hat{\lambda}^{(a)}(\cdot)$ estimates the hazard *conditional* on $T, C \geq \tau$).
- Hence we must instead solve the observational problem:

$$\hat{\lambda}^{(a)}(\tau \mid x) = \underbrace{\arg\min_{h_{a,\tau} \in \mathcal{H}} \mathbb{E}_{X, Y(\tau) \sim \mathbb{P}_{a,\tau}(\cdot)} \left[\ell\left(Y(\tau), h_{a,\tau}(X)\right) \right]}_{\text{Observational Problem}}$$

  where $\mathbb{P}_{a,\tau}(\cdot)$ is the joint distribution of patients who are still at-risk.
- This would closely approximate the target problem when $\mathbb{P}_0(\cdot) \approx \mathbb{P}_{a,\tau}(\cdot)$. However, this is often not the case.

Why is $\mathbb{P}_0 \neq \mathbb{P}_{a,\tau}$? Covariate shift

- If the treatment assigned depends on the covariates, then $\mathbb{P}_{a=1,\tau=0}(X) \neq \mathbb{P}_{a=0,\tau=0}(X)$ at the outset $\implies \mathbb{P}_{a,\tau}(X) \neq \mathbb{P}_0(X)$.
- The event itself or censoring (T or C) can be dependent on the covariates. If only 'healthy' individuals survive to be at-risk past a certain time $\tau' > 0$, then clearly these patients have a different covariate distribution than individuals at baseline: $\mathbb{P}_{a,\tau=0}(X) \neq \mathbb{P}_{a,\tau'}(X) \implies \mathbb{P}_{a,\tau'}(X) \neq \mathbb{P}_0(X)$.

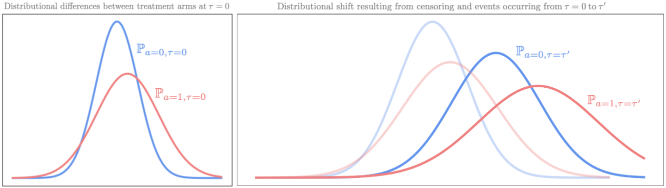


Figure: (**Left**): A classic problem in observational studies arises in the difference between populations across treatment arms. These differences are present upon the assignment of treatments. (**Right**): In TTE studies, new issues arise due to censoring and event realizations that take place over time.

Balancing $\mathbb{P}_0$ and $\mathbb{P}_{a,\tau}$ over time

- SurVITE proposes a resolution to the unique sources of covariate shift that arise in TTE data, using representation learning to balance the target and observational distributions.

- For fixed a and τ, the goal is to learn a representation $\Phi : \mathcal{X} \to \mathcal{R}$ outputting a set of latent features that are similar between the baseline distribution $\mathbb{P}_0^{\Phi}$ and the at-risk distribution $\mathbb{P}_{a,\tau}^{\Phi,w}$.

- The distance between them can be approximated using the Integrated Probability Metric:

$$\mathrm{IPM}\left(\mathbb{P}_0^{\Phi}, \mathbb{P}_{a,\tau}^{w,\Phi}\right) = \sup_{h \in \mathcal{H}} \left| \int h(\phi)(\mathbb{P}_0^{\Phi}(\phi) - \mathbb{P}_{a,\tau}^{w,\Phi}(\phi)) d\phi \right|$$

where $\phi = \Phi(X = x)$

SurVITE bound

- (Curth, Lee, van der Schaar, 2021) provide a bound on the target risk, summing (i) the observational risk, (ii) the distance between the latent baseline and observational distributions and (iii) the information lost in the representation.

$$\underbrace{\mathbb{E}_{X\sim\mathbb{P}_0}\left[\ell_{h,\mathbb{P}}(X;a,\tau)\right]}_{\text{Target Risk}} \leq \underbrace{\mathbb{E}_{X\sim\mathbb{P}_{a,\tau}}\left[w_{a,\tau}(X)\ell_{h,\mathbb{P}}(X;a,\tau)\right]}_{\text{Weighted observational risk}}$$
$$+ C_\Phi \underbrace{\mathrm{IPM}_{\mathcal{H}}\left(\mathbb{P}_0^{\Phi}, \mathbb{P}_{a,\tau}^{w,\Phi}\right)}_{\text{Distance in }\Phi\text{-space}}$$
$$+ \underbrace{\eta_{\Phi}^{\ell}(h)}_{\text{Info loss}}$$

- Since the info loss is included, this bound does not rely on the invertibility of Φ! Many others in the literature do.

Empirical SurVITE risk

- An empirical analogue of the SurVITE bound is then proposed. This estimator parameterizes Φ with θ_ϕ and h with $\theta_{h_{a,\tau}}$, implemented as fully-connected neural networks. The goal is to minimize the following approximator of the target risk:

$$\mathcal{L}_{\text{target}}(\theta_\phi, \theta_{a,\tau}) = \mathcal{L}_{\text{risk}}(\theta_\phi, \theta_h) + \beta \mathcal{L}_{\text{IPM}}(\theta_\phi)$$

where

$$\mathcal{L}_{\text{risk}}\ (\theta_\phi, \theta_h) = \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} \sum_{i:\tilde{\tau}_i \geq t} \Big(\underbrace{n_{1,t}^{-1} a_i \ell\left(y_i(t), h_{1,t}\left(\Phi\left(x_i\right)\right)\right)}_{\text{ERM for patient sub-group } a=1} + \underbrace{n_{0,t}^{-1}\left(1-a_i\right) \ell\left(y_i(t), h_{0,t}\left(\Phi\left(x_i\right)\right)\right)}_{\text{ERM for patient subgroup } a=0} \Big)$$

$$\mathcal{L}_{\text{IPM}}\left(\theta_\phi\right) = \sum_{a \in \{0,1\}} \sum_{t=1}^{t_{\max}} \underbrace{\text{Wass}\left(\left\{\Phi\left(x_i\right)\right\}_{i=1}^{n}, \left\{\Phi\left(x_i\right)\right\}_{i:\tilde{\tau}_i \geq t, a_i=a}\right)}_{\text{finite-sample Wasserstein distance}}$$

My contribution: investigating SurVITE assumptions

- While the info loss $\eta_{\Phi}^{\ell}(h)$ serves an important purpose in the bound, it is notably absent in the empirical risk!

- Similarly, the importance weightings present in the bound ares assumed equal to one for all X.

- Taken together, these assumptions imply that:
  - The information lost in estimating Φ is not excessive.
  - True importance weights $w_{a,\tau}^{*}(x) = \frac{\mathbb{P}_0(x)}{\mathbb{P}_{a,\tau}(x)}$ can be replaced by the learned representation Φ.

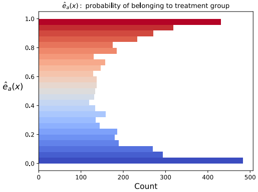
- These assumptions interact in a pernicious way: by assuming that importance weights can be replaced by a learned representation, we start with more imbalance and necessitate the need for a more aggressive (hence lossy) representation.

My contribution: proposed resolutions

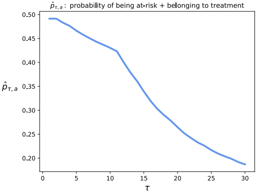
- Reintroduce importance weightings $\hat{\boldsymbol{w}}_i = \{\hat{w}_{a,\tau}(x_i)\}_{a \in A, \tau \in \mathcal{T}}$.

- $\hat{w}_{a,\tau}(x) = \frac{\hat{p}_{\tau,a}}{\hat{e}_a(x)\hat{r}^a(x,\tau)}$, where $\hat{e}_a(x) = \mathbb{P}(A = a | X = x)$,
  $\hat{p}_{\tau,a} = \mathbb{P}(\tilde{T} \geq \tau | A = a)$, and $\hat{r}^{(a)}(x,\tau) = \mathbb{P}(\tilde{T} \geq \tau | A = a, X = x)$.

- $\hat{p}_{\tau,a} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\tilde{\tau}_i \geq \tau] \times \mathbb{1}[a = a_i]$ can be estimated explicitly, while I estimate $\hat{e}_a(x)$ in advance of any network fitting via a logistic regression.



(a) $\mathbb{P}(A = 1 | X = x)$



(b) $\mathbb{P}(\tilde{T} \geq \tau, A = 1)$

- However $\hat{r}^{(a)}(x,\tau)$ depends on $\lambda^{(a)}(\cdot)$ and must be learned end-to-end! The weighted loss becomes $\mathcal{L}_{\text{risk}}(\theta_\phi, \theta_h, \hat{\boldsymbol{W}})$.

My contribution: regularizing against information loss

- Mutual Information (MI) can be used to quantify the information lost in the representation:

$$\mathrm{MI}_{a,\tau}(\Phi; X) = \int_{\mathcal{X}} \int_{\mathcal{R}} \mathbb{P}^{\Phi,X}_{a,\tau}(x,\phi) \log\left(\frac{\mathbb{P}^{\Phi,X}_{a,\tau}(x,\phi)}{\mathbb{P}^{X}_{a,\tau}(x)\mathbb{P}^{\Phi}_{a,\tau}(\phi)}\right) d\phi\, dx$$

  Here the goal is to punish the information lost when converting from the at-risk covariates to the at-risk latent distribution.

- MI can be viewed in terms of information theory:

$$\mathrm{MI}_{a,\tau}(\Phi; X) = H_{a,\tau}(\Phi) - H_{a,\tau}(\Phi|X) = \mathrm{KL}(\mathbb{P}^{\Phi,X}_{a,\tau} \| \mathbb{P}^{\Phi}_{a,\tau} \otimes \mathbb{P}^{X}_{a,\tau})$$

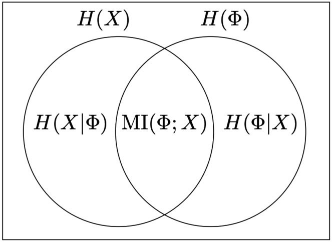


Figure: MI expressed in terms of information gain.

My contribution: regularizing against information loss

- While it is difficult to compute $\mathrm{KL}(\mathbb{P}^{\Phi,X}_{a,\tau} || \mathbb{P}^{\Phi}_{a,\tau} \otimes \mathbb{P}^{X}_{a,\tau})$, we can obtain its lower bound via the Donsker-Varadhan (DV) reformulation of KL-divergence:

$$\begin{aligned} \mathrm{KL}(\mathbb{P}^{\Phi,X}_{a,\tau} || \mathbb{P}^{\Phi}_{a,\tau} \otimes \mathbb{P}^{X}_{a,\tau}) &= \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}^{\Phi,X}_{a,\tau}}[f(x,\phi)] - \log \mathbb{E}_{\mathbb{P}^{X}_{a,\tau} \otimes \mathbb{P}^{\Phi}_{a,\tau}}\left[e^{f(x,\phi)}\right] \\ &\geq \mathbb{E}_{\mathbb{P}^{\Phi,X}_{a,\tau}}[f(x,\phi)] - \log \mathbb{E}_{\mathbb{P}^{X}_{a,\tau} \otimes \mathbb{P}^{\Phi}_{a,\tau}}\left[e^{f(x,\phi)}\right] \end{aligned}$$

- MINE (Belghazi et. al) conveniently provides a neural network approximation and algorithm for DV lower-bound minimization with generic R.Vs.

$$\mathcal{L}_{\mathrm{MI}}(\theta_f) = \frac{1}{n}\sum_{i=1}^{n} f(y_i, z_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{f(y_i, \tilde{z}_i)}\right)$$

where $\tilde{z}_i$ is permuted relative to z_i so as to be a draw from the marginal distribution $\mathbb{P}_Z$.

My contribution: regularizing against information loss

- I adapt MINE for use in the context of HTE estimation and in particular to TTE data.

$$\mathrm{MI}(\theta_{f_{1,t}}) = n_{1,t}^{-1} \textstyle\sum_{i:\tilde{\tau}_i \geq t} a_i \left(f_{1,t}\left(x_i, \phi_i\right)\right) - \log\left(n_{1,t}^{-1} \textstyle\sum_{i:\tilde{\tau}_i \geq t} a_i e^{f_{1,t}(x_i, \tilde{\phi}_i)}\right)$$

$$\mathrm{MI}_{0,t}(\theta_{f_{0,t}}) = n_{0,t}^{-1} \textstyle\sum_{i:\tilde{\tau}_i \geq t} (1-a_i) \left(f_{0,t}\left(x_i, \phi_i\right)\right) - \log\left(n_{0,t}^{-1} \textstyle\sum_{i:\tilde{\tau}_i \geq t} (1-a_i) e^{f_{0,t}(x_i, \tilde{\phi}_i)}\right)$$

$$\mathcal{L}_{\mathrm{MI}}(\theta_{f_{a,t}}) = \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} -\left(\mathrm{MI}_{1,t}(\theta_{f_{1,t}}) + \mathrm{MI}_{0,t}(\theta_{f_{0,t}})\right)$$

- What we're left with is a computable loss function penalizing the information lost in the representation.
- Why is that useful for learning better representations?...

My contribution: Modifying the SurVITE risk function

- Estimating $\hat{\boldsymbol{W}}$ is 'cheap' w.r.t. info loss. We don't lose valuable information in the covariates. The same is not true of Φ.
- We would like to learn Φ and $\hat{\boldsymbol{W}}$ together end-to-end, forcing the model to only estimate lossy representations when the same effect cannot be achieved by importance weighting.
  $\implies$ weighting information loss by $\alpha\mathcal{L}_{\mathrm{MI}}$
- At the same time estimating $\hat{\boldsymbol{W}}$ is prone to outlying weights that cause high-variance in the estimator. We want to penalize the weights for deviating from 1.
  $\implies$ weighting deviation from 1 by $\lambda_w||\hat{\boldsymbol{W}} - 1||_2$
- With α, λ_w arbitrarily large, we recover the SurVITE estimator!

My contribution: Modifying the SurVITE risk function

- My modified SurVITE estimator achieving these properties is:

$$\mathcal{L}_{\text{target}}(\theta_\phi, \theta_{h_{a,\tau}}, \hat{\boldsymbol{W}}, \theta_{f_{a,t}}) = \mathcal{L}_{\text{risk}}(\theta_\phi, \theta_h, \hat{\boldsymbol{W}}, \lambda_w) + \beta\mathcal{L}_{\text{IPM}}(\theta_\phi) + \alpha\mathcal{L}_{\text{MI}}(\theta_{f_{a,t}})$$

β controls our sensitivity to covariate shift, while α controls our sensitivity to information loss.

Applications and Extensions

- I created a prototype of my MI network and hosted it here: SurVITE-MI
- With more time I would have liked to:
  - Investigate other proxies for information loss. Is there a more principled proxy based on excess targeted information loss?
  - Conduct experiments evaluating my estimator against SurVITE on semi-synthetic data!
- **Thank you!** ☺

Appendix: Integral Probability Metric

Given a class $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$, $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$:

$$\mathrm{IPM}_{\mathcal{H}}(\mathbb{P}, \mathbb{Q}) = \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{X \sim \mathbb{P}} [h(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} [h(Y)] \right|$$

The class of functions determines the finite-sample distance:

- When $\mathcal{H} = \{h : h \text{ is 1-Lipschitz }\}$, the IPM distance is Wasserstein, used here!
- When $\mathcal{H} = \{h : \|h\|_\infty \leq 1\}$, we get the total variation distance.
- $+$ Others!

Appendix: Do operator

Interventions and counterfactuals are defined through a mathematical operator called do(x), which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged.