Phd Mini-Task

Ryan Lucas

1 Introduction

SurVITE ([1]) is a new framework for estimating population-level hazard functions in time-to-event (TTE) data. The method targets unique sources of covariate shift that arise in the TTE setting, given how the characteristics of the at-risk population change from the outset of the study up to the eventual time horizon. The authors target two main sources of time-induced covariate shift: censoring bias, which occurs as censored individuals are gradually lost from the at-risk population, and event-based shift, which occurs as individuals experiencing the event hazard drop out of the at-risk population.

The proposed approach addresses this covariate shift by estimating representations that balance the baseline distribution with the at-risk distribution over time, by mapping from the (increasingly) imbalanced covariates to a balanced latent feature distribution. The framework builds on the success of representation learning in standard treatment effect estimation (e.g. [2], [3]). However, while the resulting SurvITE model successfully addresses time-induced covariate shift by extending these representations to the TTE setting, the method may suffer from excess information loss since the representation is not guaranteed to be invertible. By introducing importance weights and regularizing the representation against information loss, I propose several modifications to tighten the upper bound on the SurVITE risk function.

2 Notation and problem setup

I borrow the notation from [1] and present a summary of key notation in Table 1. Here we assume access to a TTE dataset $\mathcal{D} = \{(a_i, x_i, \tau_i, \delta_i)\}_{i=1}^n$ respectively corresponding to realizations of random variables of assigned treatment $A \in \{0, 1\}$, the patient covariates $X \in \mathcal{X}$, the observed time-to-event $\tilde{T} \in \mathcal{T}$ and an indication of censoring $\Delta \in \{0, 1\}$. Expanded to long form, this gives $Y(\tau) \in \{0, 1\}$ as the target variable, an indication of whether the event of interest occurs at time τ or not. We seek a probabilistic classifier $\lambda^a : \mathcal{X} \times \mathcal{T} \to [0, 1]$ to predict at any time τ, the probability that the event will occur for a patient belonging to treatment subgroup a. In [1] and elsewhere in survival analysis, this is achieved using treatment-specific hazard functions:

$$\underbrace{\lambda^a(\tau|, x)}_{\substack{\text{hazard function} \\ \text{for treatment group } a}} = \underbrace{\mathbb{P}(T = \tau \mid T \geq \tau, \text{do}(A = a, C \geq \tau), X = x)}_{\substack{\text{Probability event occurs at time } \tau \\ \text{given no event so far \& patient characteristics}}} \tag{1}$$

where do$(\cdot)$ is the same operator used in [1]. As annotated in Equation 1, treatment-specific hazard functions find, for each treatment subgroup a, the probability that the event (e.g. death in a medical setting) occurs at time τ, conditioned on the patient still being at-risk and known patient characteristics. While [1] also extends the SurVITE framework beyond hazard estimates to other target parameters that are generally of interest in survival analysis and HTE, they will not be investigated in this research task.

	Name	Random Variable	Data Realization
Short data	Assigned Treatment	$A \in \{0, 1\}$	$\{a_i\}_{i=1}^n$
	Covariates	$X \in \mathcal{X}$	$\{x_i\}_{i=1}^n$
	Time period of interest	$\mathcal{T} = \{1, ..., t_{\max}\}$	—
	Time-to-event	$T \in \mathcal{T}$	—
	Time-to-censoring	$C \in \mathcal{T}$	—
	Observed time-to-event	$\tilde{T} = \min(T, C)$	$\{\tilde{\tau}_i\}_{i=1}^n$
	Censoring (Y/N)	$\Delta = \mathbb{1}(T \leq C)$	$\{\delta_i\}_{i=1}^n$
	Basline Dataset	$\mathcal{D} = (A, X, \tilde{T}, \Delta)$	$\{(a_i, x_i, \tau_i, \delta_i)\}_{i=1}^n$
Long data	Censoring before time t (Y/N)	$N_T(t) = \mathbb{1}(\tilde{T} \leq t, \Delta = 1)$	—
	Event before time t (Y/N)	$N_C(t) = \mathbb{1}(\tilde{T} \leq t, \Delta = 0)$	—
	Event at time t (Y/N)	$Y(t) = \mathbb{1}(\tilde{T} = \tau, \Delta = 1)$	$\{y_i(t)\}_{t \in \mathcal{T}, i \in [n]}$
	At-risk dataset	$\mathcal{D}_{a,\tau} = (X, \Delta, Y(\tau))$	$\{(x_i, \delta_{i,t}, y_i(t)\}_{t \in \mathcal{T}, i \in [n_{a,\tau}]}$

Table 1: Notations used in [1], assumed throughout.

3 Covariate shift in TTE data

As outlined in [1], covariate shift arises when the training distribution $X, Y \sim \mathbb{Q}_0$ and the target distribution $X, Y \sim \mathbb{Q}_1$ do not match. Given a hypothesis class $\mathcal{H}$, we are thus left in a situation where $\arg\min_{h\in\mathcal{H}} \mathbb{E}_{X,Y\sim\mathbb{Q}_1(\cdot)}[\ell(Y, h(X))] \neq \arg\min_{h\in\mathcal{H}} \mathbb{E}_{X,Y\sim\mathbb{Q}_0(\cdot)}[\ell(Y, h(X))]$, where ℓ is some loss function.[1]

In this research task, the goal is to estimate hazard functions $\lambda^a(\tau \mid x)$ for the entire population. Thus our target distribution is the general population (or baseline) distribution $X \sim \mathbb{P}(X)$, which I will refer to as $\mathbb{P}_0$. However, given the nature of TTE data, we are forced to train on data from an observational at-risk distribution $\mathbb{P}_{a,\tau} = \mathbb{P}_{a,\tau}(X, Y(\tau)) = \lambda^a_T(\tau|X)\mathbb{P}_{a,\tau}(X)$ with $\mathbb{P}_{a,\tau}(X) = \mathbb{P}(X|\tilde{T} \geq \tau, A = a)$. This gives rise to a separation between our target problem, which we ultimately *want to solve*, and our observational problem which we *can solve* given TTE data. In each respective case, we find our estimate of the hazard function $\hat{\lambda}^a(\tau|x)$ as the minimizer in:[2]

$$\underbrace{\arg\min_{h\in\mathcal{H}} \mathbb{E}_{X,Y(\tau)\sim\mathbb{P}_0(\cdot)}\left[\ell\left(Y(\tau), h(X)\right)\right]}_{\text{Target Problem}} \qquad \underbrace{\arg\min_{h_{a,\tau}\in\mathcal{H}} \mathbb{E}_{X,Y(\tau)\sim\mathbb{P}_{a,\tau}(\cdot)}\left[\ell\left(Y(\tau), h_{a,\tau}(X)\right)\right]}_{\text{Observational Problem}}$$

Covariate shift in this scenario thus corresponds to differences between the marginal distribution $\mathbb{P}_0$ and the at-risk distribution $\mathbb{P}_{a,\tau}$, which can bias our estimation of the target problem. The first source of covariate shift, which arises naturally when using observational data to estimate HTEs, is treatment selection bias. If sicker individuals, for example, are given treatment $a = 0$, while less sick individuals are given treatment $a = 1$, then in general $\mathbb{P}_{a=0,\tau=0} \neq \mathbb{P}_{a=1,\tau=0}$ and hence $\mathbb{P}_{a,\tau=0} \neq \mathbb{P}_0$ for a given a. Notice that this source of shift is present upon the assignment of treatments ($\tau = 0$, left panel of Figure 1).

This is not the case for the two other sources of shift in TTE data. Censoring bias occurs part-way through the study in situations where the censoring hazard is dependent on the covariates. For example, if healthy individuals are censored (discontinue the study) more often, then it is clear that the distribution at some time $\tau' > 0$ changes relative to the outset ($\mathbb{P}_{a,\tau=0} \neq \mathbb{P}_{a,\tau=\tau'}$). Similarly, if the event-hazard (for instance, death) depends on the covariates then once again $\mathbb{P}_{a,\tau=0} \neq \mathbb{P}_{a,\tau=\tau'}$ (right panel of Figure 1). More generally, it can be said that $\mathbb{P}_{a,\tau=\tau_1} \neq \mathbb{P}_{a,\tau=\tau_2}$ for $\tau_1 \neq \tau_2$, since in principle the distribution will shift each time either censoring or the event-hazard occur.

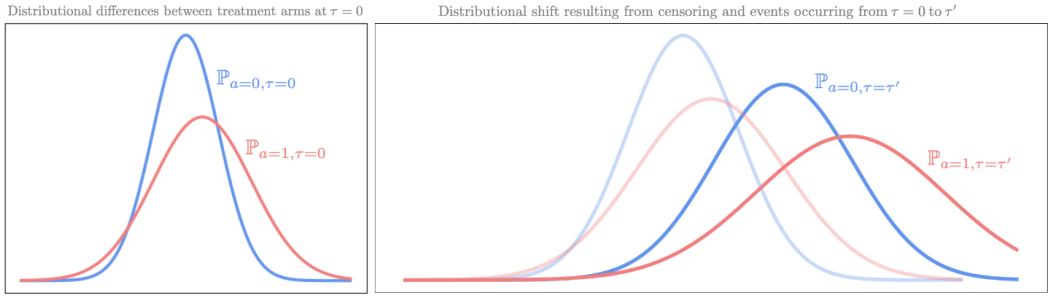


Figure 1: (**Left**): A classic problem in observational studies arises in the difference between populations across treatment arms; $\mathbb{P}_{a=0,\tau=0} \neq \mathbb{P}_{a=1,\tau=0} \neq \mathbb{P}_0$. These differences are present upon the assignment of treatments ($\tau = 0$), and result from so-called treatment selection bias. (**Right**): In TTE studies, new issues arise due to censoring and event realizations that take place over time. Hence $\mathbb{P}_{a,\tau=\tau_1} \neq \mathbb{P}_{a,\tau=\tau_2}$ for a given $a \in \{0, 1\}$.

4 Representation learning approach

Proposed in [1] is a resolution to the unique sources of covariate shift that arise in TTE data, relying on representation learning. In the general setting of domain shift, the motivation for representation learning is to *only* exploit information that is common to both the training and target distributions [4]. A representation achieves this by mapping from the observational covariates to a distribution of latent features that are similar between the training and target distributions. In [1] the authors are concerned with balancing the baseline distribution $\mathbb{P}_0$, for which we would ultimately like to obtain hazard estimates, with the observational at-risk distribution $\mathbb{P}_{a,\tau}$. The observational distribution has been contaminated with censoring and event bias and, if used directly, would give biased hazard estimates in the target problem. For fixed a and τ, the goal is thus to learn a representation $\Phi : \mathcal{X} \to \mathcal{R}$ outputting a set of latent features that are similar between the baseline distribution and the at-risk distribution. For a given hypothesis class $\mathcal{H}$, the similarity between distributions $\mathbb{P}$ and $\mathbb{Q}$ can be measured by the Integrated Probability Metric (IPM):

[1] At the same time, under covariate shift we assume the conditional distributions remain the same, i.e. $\mathbb{Q}_0(Y|X) = \mathbb{Q}_1(Y|X)$. That is, even though the distributions themselves differ, the conditional relationship between X and Y we assume does not.

[2] Here ℓ is given as the log-loss, hence both the target and observational problems are likelihood optimization problems.

$$\text{IPM}_{\mathcal{H}}(\mathbb{P},\mathbb{Q}) = \sup_{h\in\mathcal{H}} \left| \int h(x)(\mathbb{P}(x) - \mathbb{Q}(x))dx \right|$$

In our case, the goal is to minimize $\text{IPM}_{\mathcal{H}}\left(\mathbb{P}_0^{\Phi}, \mathbb{P}_{a,\tau}^{w,\Phi}\right)$, the distance between the latent baseline distribution and the latent observational at-risk distribution. The IPM quantifies the distance between our training and target distributions in the latent features we seek to find. The theoretical result proposed in [1] and given in Proposition 1, then shows that it is possible to bound the risk of the target problem based on a summation of (i) the weighted observational risk of a classification model $h \in \mathcal{H}$ making estimates $h(\Phi(X))$ based on latent features, (ii) The IPM distance between the baseline distribution and at-risk distribution in representation space and (iii) the information lost in the representation.

This bound differs from others proposed in the literature in that it does not rely on invertibility to hold. Where traditional bounds rely on the fact that when Φ is invertible, the information loss $\eta_{\Phi}^{\ell}(h) = \xi_{\mathbb{Q}^{\Phi},\mathbb{Q}}(x) \overset{\text{def}}{=} \ell_{h,\mathbb{Q}^{\Phi}}(\phi;a,\tau) - \ell_{h,\mathbb{Q}}(x;a,\tau) = 0$, Proposition 1 is guaranteed to hold regardless of the value $\eta_{\Phi}^{\ell}(h)$ takes. However, a criticism of their proposal is that while the bound does not rely on the assumption of invertibility, the *tightness* of the upper bound does. Namely, for some non-invertible function Φ, the target risk $\mathbb{E}_{X\sim\mathbb{P}_0}\left[\ell_{h,\mathbb{P}}(X;a,\tau)\right]$ does not have a tight upper bound. As I will discuss in the upcoming sections, this has important consequences for their empirical estimator, where no explicit dependence on information loss is made.

Proposition 1: (Curth, Lee, van der Schaar, 2021)

For fixed a, τ and representation $\Phi : \mathcal{X} \to \mathcal{R}$, let $\mathbb{P}_0^{\Phi}, \mathbb{P}_{a,\tau}^{\Phi}$ and $\mathbb{P}_{a,\tau}^{w,\Phi}$ denote the target, observational, and weighted observational distribution of the representation Φ. Define the pointwise losses

$$\ell_{h,\mathbb{Q}}(x;a,\tau) \overset{\text{def}}{=} \mathbb{E}_{Y(\tau)|x,a\sim\mathbb{Q}}[\ell(Y(\tau), h(\Phi(X))) \mid X = x, A = a]$$
$$\ell_{h,\mathbb{Q}^{\Phi}}(\phi;a,\tau) \overset{\text{def}}{=} \mathbb{E}_{Y(\tau)|\phi,a\sim\mathbb{Q}^{\Phi}}[\ell(Y(\tau), h(\Phi)) \mid \Phi = \phi, A = a]$$

of (hazard) hypothesis $h \equiv h_{a,\tau} : \mathcal{R} \to [0,1]$ w.r.t. distributions in covariate and representation space, respectively. Assume there exists a constant $C_{\Phi} > 0$ s.t. $C_{\Phi}^{-1}\ell_{h,\mathbb{P}_{a,\tau}^{w,\Phi}}(\phi,a,\tau) \in \mathcal{G}$ for some family of functions $\mathcal{G}$. Then we have that

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

where $\text{IPM}\ (\mathbb{P},\mathbb{Q}) = \sup_{g\in\mathcal{G}}\left|\int g(x)(\mathbb{P}(x) - \mathbb{Q}(x))dx\right|$ and we define the excess target information loss $\xi_{\mathbb{Q}^{\Phi},\mathbb{Q}}(x) = \ell_{h,\mathbb{Q}^{\Phi}}(\phi;a,\tau) - \ell_{h,\mathbb{Q}}(x;a,\tau) = \ell_{h,\Phi}(\Phi;a,\tau) - \ell_{h,\mathbb{Q}}(x;a,\tau)$. For invertible Φ, $\eta_{\Phi}^{\ell}(h) = \xi_{\mathbb{Q}^{\Phi},\mathbb{Q}}(x) = 0$.

Moving from Proposition 1 to its empirical analogue, the SurVITE estimator finds representation $\Phi : \mathcal{X} \to \mathcal{R}$ and hazard function $h : \mathcal{R} \to [0,1]$ by parameterizing Φ with θ_{ϕ} and h with $\theta_{h_{a,\tau}}$, implemented as fully-connected neural networks. The goal is then to minimize the following approximator of the target risk:

$$\mathcal{L}_{\text{target}}(\theta_{\phi},\theta_{a,\tau}) = \mathcal{L}_{\text{risk}}(\theta_{\phi},\theta_h) + \beta\mathcal{L}_{\text{IPM}}(\theta_{\phi}) \tag{2}$$

where

$$\mathcal{L}_{\text{risk}}\ (\theta_{\phi},\theta_h) = \frac{1}{t_{\max}}\sum_{t=1}^{t_{\max}}\sum_{i:\tilde{\tau}_i\geq t}\underbrace{n_{1,t}^{-1}a_i\ell\left(y_i(t), h_{1,t}\left(\Phi\left(x_i\right)\right)\right)}_{\text{ERM for patient sub-group } a=1} + \underbrace{n_{0,t}^{-1}\left(1-a_i\right)\ell\left(y_i(t), h_{0,t}\left(\Phi\left(x_i\right)\right)\right)}_{\text{ERM for patient subgroup } a=0}, \tag{3}$$

$$\mathcal{L}_{\text{IPM}}\left(\theta_{\phi}\right) = \sum_{a\in\{0,1\}}\sum_{t=1}^{t_{\max}}\underbrace{\text{Wass}\left(\left\{\Phi\left(x_i\right)\right\}_{i=1}^{n}, \left\{\Phi\left(x_i\right)\right\}_{i:\tilde{\tau}_i\geq t, a_i=a}\right)}_{\text{finite-sample Wasserstein distance}}, \tag{4}$$

where $n_{a,t}$ is the number of individuals belonging to treatment sub-group a at time t, implemented to ensure each a, τ-combination weighs equally in the loss.

The estimator $\mathcal{L}_{\text{target}}$ has many strong properties. $\mathcal{L}_{\text{risk}}$ punishes deviation between predictions of the hazard and the at-risk labels, while $\mathcal{L}_{\text{IPM}}$ penalises the distance between the baseline distribution and *each* at-risk distribution in representation space. Minimizing Equation 2 thus ensures prediction accuracy at the level of the observational distribution, while also moving the observational distribution closer to the target distribution we truly seek to minimize over. However, the issue remains that there is no explicit dependence on $\eta_{\Phi}^{\ell}(h)$ in the target risk, despite

its important purpose within the bound of Proposition 1. Moreover, the authors assume that the importance weights $w_{a,\tau}(X)$ are equal to 1 for all X.[3] I demonstrate in the following section that these assumptions together can cause an unnecessary relaxation of the upper bound and thus a poorer approximation of the target risk.

5 My contributions

5.1 Investigating SurVITE assumptions

The authors make several assumptions, both implicit and explicit, when converting from the theoretical bound in Proposition 1 to its estimating counterpart in Equation 2. Firstly, the importance weights $w_{a,\tau}(X)$ within the weighted observational risk $\mathbb{E}_{X\sim\mathbb{P}_{a,\tau}}\left[w_{a,\tau}(X)\ell_{h,\mathbb{P}}(X;a,\tau)\right]$ are explicitly assumed to be equal to 1 for all X. This is a departure from standard HTE estimation, where importance weightings (most often, propensity weights) are a popular resolution to covariate shift [5] and have been shown to minimize the asymptotic variance of the estimated treatment effect among the class of balancing weights [6].

A separate assumption made by the authors lies in removing the dependence on information loss in the empirical risk, relying on the prospect that the information lost in estimating $\Phi(X)$ is not excessive. This is related to the invertibility assumption made in other work (e.g. [7], [8]), where Φ is assumed invertible and hence $\eta^{\ell}_{\Phi}(h)$ is taken to be zero. However, it is well understood that most neural networks are not invertible, and imposing them to be invertible requires restrictive constraints on the architecture [9]. Thus while the bound in Proposition 1 does not rely on the assumption of invertibility, the bound would also be of significantly lower quality in the absense of that condition.

I now argue that if these assumptions do not hold, there may be unnecessary inflation of the SurVITE risk and hence a failure to obtain a tight upper bound on the target risk. I formalize and discuss these assumptions below:

1. In ignoring the importance weightings, the authors assume that the empirical representation function $\Phi(X)$ is capable of estimating balanced latent features which can replace the true importance weightings. In other words, given access to true importance weights $w^{*}_{a,\tau}(X)$ and a function $g:(\mathcal{X}\times\mathcal{T})\to[0,1]$ making classifications directly from observed covariates, the difference:

$$\mathbb{E}_{X\sim\mathbb{P}_{a,\tau}}\Bigg[\underbrace{w_{a,\tau}(X)}_{=1}\ \underbrace{\ell_{h,\mathbb{P}}(X;a,\tau)}_{\text{loss under representation}}\Bigg]-\mathbb{E}_{X\sim\mathbb{P}_{a,\tau}}\Bigg[\underbrace{w^{*}_{a,\tau}(X)}_{\text{true imp. weights}}\ \underbrace{\mathbb{E}_{Y(\tau)|x,a\sim\mathbb{P}_{a,\tau}}[\ell(Y(\tau),g(X))\mid X=x,A=a])}_{\text{loss under no covariate adjustment}}\Bigg]$$

   is assumed to be sufficiently close to zero in practice. The authors presumably know that true importance weights $w^{*}_{a,\tau}(X)$ could be far from 1. The assumption made is hence not that $w_{a,\tau}(X)=1$ is a valid importance weight, but rather that $\Phi(X)$ is an appropriate counter-measure achieving a similar rebalancing effect. Unfortunately, due to the non-linear nature and lack of guarantees concerning $\Phi(X)$, as well as our inability to access true importance weights in an observational setting, this assumption is difficult to prove nor falsify in practice. Still, while this assumption has the potential to be problematic, I argue that it may not be necessary. In subsequent sections, I investigate whether stable weights can be estimated using appropriate regularization.[4]

2. In removing all terms depending on the information loss $\eta^{\ell}_{\Phi}(h)$ in the empirical estimator, the authors implicitly assume there is not excessive information lost in the estimation of the representation. That is, the information loss:

$$\eta^{\ell}_{\Phi}(h)=\mathbb{E}_{X\sim\mathbb{P}_0}\left[\xi_{\mathbb{P}^{\Phi},\mathbb{P}}(X)-\xi_{\mathbb{P}^{w,\Phi}_{a,\tau},\mathbb{P}}(X)\right]$$

   is assumed to be sufficiently small in practice. This assumption would hold in situations where the risk based on latent features is close to the risk based on observational covariates. In this sense, the SurVITE risk already does punish information loss to an extent, since the $\mathcal{L}_{\text{risk}}$ term punishes Φ for producing latent features $\Phi(X)$ that are no longer useful for predicting $Y(\tau)$. However, the information loss may still be much larger than necessary, due to the fact that importance weights are ignored. By assuming that the importance weights can be effectively replaced by a learned representation, we necessitate the need for a more aggressive (and thereby lossy) representation. That is, by abandoning importance weights, we start with more imbalance between the at-risk distribution and the target, encouraging a more aggressive minimization of the IPM risk $\mathcal{L}_{\text{IPM}}$. In the following section, I show how regularization based on finite-sample mutual information can be integrated

[3] However, the authors do explain that while the importance weighting is assumed to be one for their paper, they leave $w_{a,\tau}(X)$ as placeholder that admits other importance weight estimates.

[4] As the authors point out in [1], estimated importance weightings can be prone to problems such as finite-sample bias and high variance that produce extreme weights [10].

into the SurVITE loss function to mitigate information loss. Taken together with the importance weightings, these adjustments encourage SurVITE to only estimate lossy representations when a similar effect can not be achieved with importance weighting.

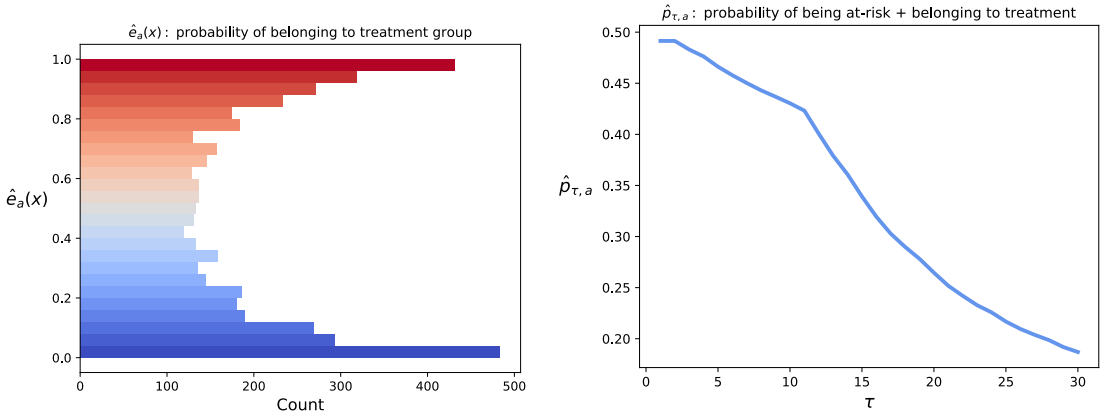
5.2 Proposed resolutions

In addressing the first assumption, I consider whether replacing the placeholder importance weightings $w_{a,\tau}(X) = 1$ with true weights $w^*_{a,\tau}(x) = \frac{\mathbb{P}_0(x)}{\mathbb{P}_{a,\tau}(x)}$ learned end-to-end could relieve information loss and tighten the bound of the SurVITE risk. I propose a framework that jointly learns these importance weights while regularizing against information loss.

Adjustment for importance weighting. My own implementation firstly involves a modification to $\mathcal{L}_{\text{risk}}(\theta_\phi, \theta_h)$, introducing estimated importance weightings $\hat{\boldsymbol{w}}_i = \{\hat{w}_{a,\tau}(x_i)\}_{a\in A,\tau\in\mathcal{T}}$ for each patient i:

$$\mathcal{L}_{\text{risk}}(\theta_\phi, \theta_h, \hat{\boldsymbol{W}}) = \sum_{t=1}^{t_{\max}} \sum_{i:\tilde{\tau}_i \geq t} \left(\hat{w}_{1,t}(x_i) n_{1,t}^{-1} a_i \ell(y_i(t), h_{1,t}(\Phi(x_i))) + \hat{w}_{0,t}(x_i) n_{0,t}^{-1}(1-a_i)\ell(y_i(t), h_{0,t}(\Phi(x_i)))\right) + \lambda ||\hat{\boldsymbol{W}}||_2$$

with $\hat{w}_{a,\tau}(x) = \frac{\hat{p}_{\tau,a}}{\hat{e}_a(x)\hat{r}^a(x,\tau)}$, where $\hat{e}_a(x) = \mathbb{P}(A = a|X = x)$ is the propensity score, $\hat{p}_{\tau,a} = \mathbb{P}(\tilde{T} \geq \tau|A = a)$ is the overall probability of being at-risk belonging to treatment subgroup a, and $\hat{r}^a(x,\tau) = \mathbb{P}(\tilde{T} \geq \tau|A = a, X = x)$ is the probability of being at-risk given known patient covariates. $\hat{p}_{\tau,a}$ can be estimated explicitly via $\hat{p}_{\tau,a} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\tilde{\tau}_i \geq \tau] \times \mathbb{1}[a = a_i]$, while I estimate $\hat{e}_a(x)$ in advance of any network fitting via a logistic regression. I calculated these quantities for the simulated examples generated in [1] and stored at SurVITE, and present the results in Figure 2.



(a) Estimating $\mathbb{P}(A = 1|X = x)$ (b) Estimating $\mathbb{P}(\tilde{T} \geq \tau, A = 1)$

Figure 2: Components of the importance weighting that are estimated in advance.

Notice that the last component of the importance weighting, $\hat{r}^a(x,\tau)$, depends on the unknown target parameters. For this reason, it's not possible to estimate $\hat{r}^a(x,\tau)$ in advance of fitting the SurVITE network. However, it's value can be easily computed from the network as a function of the target parameters $\hat{\lambda}^a(\tau|x) = h(\Phi(x))$:

$$\begin{aligned}\hat{r}^a(x,\tau) &= \mathbb{P}(\tilde{T} \geq \tau|A = a, X = x) \\ &= \prod_{t<\tau}(1 - \hat{\lambda}^a(t|x))\end{aligned}$$

which is similar to the survival function defined in [1], except the product operator is not inclusive at $t = \tau$. Now, to summarize the estimation of the weighting function, we obtain $\hat{w}_{a,\tau}(x)$ as follows:

$$\hat{w}_{a,\tau}(x) = \underbrace{\hat{p}_{\tau,a}}_{\text{explicitly in advance}} \times \frac{1}{\underbrace{\hat{e}_a(x)}_{\text{in advance via log. regression}}} \times \underbrace{\frac{1}{\hat{r}^a(x,\tau)}}_{\text{end-to-end via } \hat{\lambda}^a(\cdot)}$$

Regularization of Φ based on information loss. The second modification I make to the SurVITE risk is to introduce a regularization term punishing the information lost in estimating the representation Φ. A tempting notion of information loss comes from mutual information (MI), which has been widely used in data compression [11]. For random variables Z and Y, MI has the following form:

$$\mathrm{MI}(Z;Y) = \int_{\mathcal{Y}} \int_{\mathcal{Z}} \mathbb{P}_{(Z,Y)}(z,y) \log \left(\frac{\mathbb{P}_{(Z,Y)}(z,y)}{\mathbb{P}_Z(z)\mathbb{P}_Y(y)} \right) dz\, dy$$

For a given a,τ-combination, we can thus use the MI between the original covariates x and the latent representation $\Phi(x) = \phi$ as a measure of the information lost in estimating the representation function:

$$\mathrm{MI}_{a,\tau}(\Phi;X) = \int_{\mathcal{X}} \int_{\mathcal{R}} \mathbb{P}^{\Phi,X}_{a,\tau}(x,\phi) \log \left(\frac{\mathbb{P}^{\Phi,X}_{a,\tau}(x,\phi)}{\mathbb{P}^{X}_{a,\tau}(x)\mathbb{P}^{\Phi}_{a,\tau}(\phi)} \right) d\phi\, dx \tag{5}$$

Notice that here I consider information loss between the at-risk covariate distribution and the at-risk latent distribution. This is done because to the extent that information is lost in the representation ϕ, this is measurable only at a contemporary level. This is true because of the time-induced covariate shift outlined in section 3; namely that the MI estimator will too suffer from these shifts. We can represent the mutual information between the at-risk covariate distribution and the at-risk latent distribution using the Kullback-Leibler (KL) divergence between the joint at-risk distribution of Φ and X and the product of the at-risk marginals. This is given below in Equation 6 and visualized in Figure 3.[5]

$$\mathrm{MI}_{a,\tau}(\Phi;X) = H_{a,\tau}(\Phi) - H_{a,\tau}(\Phi|X) = \mathrm{KL}(\mathbb{P}^{\Phi,X}_{a,\tau} || \mathbb{P}^{\Phi}_{a,\tau} \otimes \mathbb{P}^{X}_{a,\tau}) \tag{6}$$

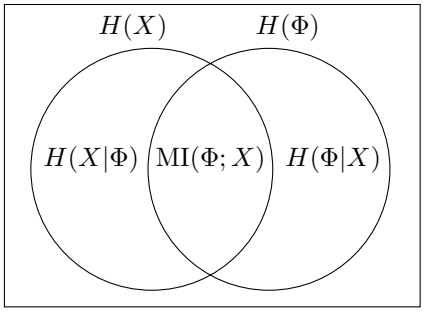


Figure 3: Mutual information between latent representation Φ and original covariates X. Here $H(\cdot)$ and $H(\cdot|\cdot)$ are the entropy and conditional entropy respectively, defined in subsection 6.1. $\mathrm{MI}(\Phi;X)$ represents the shared information gain of Φ and X. The a,τ are omitted for space but in reality MI is taken for each a,τ-combination.

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 [11][12]:

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

where for simplicity f is a set of unconstrained functions. Known as the Mutual Information Neural Estimator (MINE), [13] provides a suitable neural network approximation and algorithm for DV lower-bound minimization in a much more general setting. Given random variables Y and Z, and a function f parameterised by θ which in their case is not dependent on treatment a or time τ, MINE proposes to minimize the empirical analogue of the KL-bound:

$$\widehat{\mathrm{MI}}(Y,Z)_{(n)} = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}^{Y,Z}_{(n)}}[f(y,z)] - \log \mathbb{E}_{\mathbb{P}^{Y}_{(n)} \otimes \mathbb{P}^{Z}_{(n)}} \left[e^{f(y,z)} \right] \tag{7}$$

The expectations in Equation 7 are estimated using empirical samples from joint distribution $\mathbb{P}^{Y,Z}_{(n)}$ and the product of marginals $\mathbb{P}^{Y}_{(n)} \otimes \mathbb{P}^{Z}_{(n)}$. In practice, this involves drawing n samples from the joint distribution: $\left(y_{(1)}, z_{(1)}\right), \ldots, \left(y_{(n)}, z_{(n)}\right) \sim \mathbb{P}^{XZ}$ and *separately* drawing n samples from the marginal distribution of Z: $\tilde{z}_{(1)}, \ldots, \tilde{z}_{(n)} \sim \mathbb{P}_Z$ to give the following empirical loss:

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

[5] For the proof, see subsection 6.2

I adapt MINE for use in the context of HTE estimation and in particular to TTE data. Similarly to [1], this is done using a fully-connected neural network $f_{a,\tau} : \mathcal{X} \times \Phi \to \mathbb{R}$ parameterised by $\theta_{f_{a,\tau}}$. Presented in Equation 8 and Equation 9 are the empirical estimators for $a = 0$ and $a = 1$ with generic t:

$$\widehat{\mathrm{MI}}_{1,t}(\theta_{f_{1,t}}) = n_{1,t}^{-1} \sum_{i \in \mathcal{B}} a_i \left(f_{1,t}\left(x_i, \phi_i\right)\right) - \log \left(n_{1,t}^{-1} \sum_{i \in \mathcal{B}} a_i e^{f_{1,t}\left(x_i, \tilde{\phi}_i\right)} \right) \tag{8}$$

$$\widehat{\mathrm{MI}}_{0,t}(\theta_{f_{0,t}}) = n_{0,t}^{-1} \sum_{i \in \mathcal{B}} (1 - a_i) \left(f_{0,t}\left(x_i, \phi_i\right)\right) - \log \left(n_{0,t}^{-1} \sum_{i \in \mathcal{B}} (1 - a_i) e^{f_{0,t}\left(x_i, \tilde{\phi}_i\right)} \right) \tag{9}$$

where $\mathcal{B} = \{i : \tilde{\tau}_i \geq t\}$ and $\tilde{\phi}_i$ is a randomly sampled element of the set $\mathcal{B}$, chosen randomly so as to be a draw from the marginal distribution $\mathbb{P}^{\Phi}_{a,\tau}$. Similar to the technique employed by [1] for hazard function estimation, in practice I use just one function for $f_{a,\tau}$, to ensure information sharing between a, τ-combinations. Taken over all time and each treatment group, we then arrive at the desired mutual information risk:

$$\mathcal{L}_{\mathrm{MI}}(\theta_{f_{a,t}}) = \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} - \left(\widehat{\mathrm{MI}}_{1,t}(\theta_{f_{1,t}}) + \widehat{\mathrm{MI}}_{0,t}(\theta_{f_{0,t}}) \right) \tag{10}$$

The code for my MI estimator can be found at SurVITE-MI. Finally, introducing α as a hyperparameter punishing information loss, we arrive at the modified SurVITE risk:

$$\mathcal{L}_{\mathrm{target}}(\theta_\phi, \theta_{h_{a,\tau}}, \hat{\boldsymbol{W}}, \theta_{f_{a,t}}) = \mathcal{L}_{\mathrm{risk}}(\theta_\phi, \theta_h, \hat{\boldsymbol{W}}) + \beta \mathcal{L}_{\mathrm{IPM}}(\theta_\phi) + \alpha \mathcal{L}_{\mathrm{MI}}(\theta_{f_{a,t}}) \tag{11}$$

Each new or modified term in this SurVITE risk addresses the assumptions and issues previously outlined: the modified $\mathcal{L}_{\mathrm{risk}}(\theta_\phi, \theta_h, \hat{\boldsymbol{W}})$ term trades-off between estimating imperfect balancing weights and lossy representations. The goal is to balance the distributions as much as possible using importance weighting, lessening the need for lossy rebalancing via representations. Only if balance cannot be achieved via importance weighting do we resort to representation learning. When a representation is used, $\mathcal{L}_{\mathrm{IPM}}$ serves its usual purpose of balancing the representation between the baseline and at-risk distributions. At the same time, $\mathcal{L}_{\mathrm{MI}}$ regularizes the information loss such that $\mathcal{L}_{\mathrm{IPM}}$ does not create overly lossy representations when performing this rebalancing.

6 Appendix

6.1 Entropy and conditional entropy

The entropy of the random variable X with discrete probability distribution $p(x)$ is simply:

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

If X and Y are discrete random variables and with discrete distributions $p(x)$ and $p(y)$ respectively, and $p(x, y)$ and $p(y \mid x)$ are the values of their joint and conditional probability distributions, then:

$$\mathrm{H}(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y \mid x)$$

is the conditional entropy of Y given X.

6.2 Finite-sample Mutual Information

For random variables X and Y, continuous form mutual information has the form:

$$\mathrm{MI}(X; Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} \mathbb{P}_{(X,Y)}(x, y) \log \left(\frac{\mathbb{P}_{(X,Y)}(x, y)}{\mathbb{P}_X(x) \mathbb{P}_Y(y)} \right) dx \, dy$$

This has the following discrete reformulation:

$$
\begin{aligned}
\mathrm{MI}(X;Y) &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{(X,Y)}(x,y)\log\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{(X,Y)}(x,y)\log\frac{p_{(X,Y)}(x,y)}{p_X(x)} - \sum_{x\in X,y\in\mathcal{Y}} p_{(X,Y)}(x,y)\log p_Y(y) \\
&= \sum_{x\in X,y\in\mathcal{Y}} p_X(x)p_{Y|X=x}(y)\log p_{Y|X=x}(y) - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{(X,Y)}(x,y)\log p_Y(y) \\
&= \sum_{x\in\mathcal{X}} p_X(x)\left(\sum_{y\in\mathcal{Y}} p_{Y|X=x}(y)\log p_{Y|X=x}(y)\right) - \sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}} p_{(X,Y)}(x,y)\right)\log p_Y(y) \\
&= -\sum_{x\in\mathcal{X}} p_X(x)\mathrm{H}(Y\mid X=x) - \sum_{y\in\mathcal{Y}} p_Y(y)\log p_Y(y) \\
&= -\mathrm{H}(Y\mid X) + \mathrm{H}(Y) \\
&= \mathrm{H}(Y) - \mathrm{H}(Y\mid X) \\
&= D_{\mathrm{KL}}\left(p_{(X,Y)}\|p_Xp_Y\right)
\end{aligned}
$$

References

[1] Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26740–26753. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e0eacd983971634327ae1819ea8b6214-Paper.pdf.

[2] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9363924.

[3] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf.

[4] Fredrik D. Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. 2018. doi: 10.48550/ARXIV.1802.08598. URL https://arxiv.org/abs/1802.08598.

[5] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011. doi: 10.1080/00273171.2011.568786. URL https://doi.org/10.1080/00273171.2011.568786. PMID: 21818162.

[6] Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, nov 2017. doi: 10.1080/01621459.2016.1260466. URL https://doi.org/10.1080%2F01621459.2016.1260466.

[7] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. 2016. doi: 10.48550/ARXIV.1606.03976. URL https://arxiv.org/abs/1606.03976.

[8] Fredrik D. Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. 2018. doi: 10.48550/ARXIV.1802.08598. URL https://arxiv.org/abs/1802.08598.

[9] Yang Song, Chenlin Meng, and Stefano Ermon. Mintnet: Building invertible neural networks with masked convolutions, 2019. URL https://arxiv.org/abs/1907.07945.

[10] Huzhang Mao, Liang Li, and Tom Greene. Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research*, 28:096228021878117, 06 2018. doi: 10.1177/0962280218781171.

[11] Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, may 2021. doi: 10.1007/s10618-021-00759-3. URL https://doi.org/10.1007%2Fs10618-021-00759-3.

[12] M. D. Donsker and S. R.S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, March 1983. ISSN 0010-3640. doi: 10.1002/cpa.3160360204.

[13] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation. 2018. doi: 10.48550/ARXIV.1801.04062. URL https://arxiv.org/abs/1801.04062.