Chapter 16
Super Learning for Right-Censored Data

Eric C. Polley, Mark J. van der Laan

The super learner was introduced in Chap. 3 as a loss-based estimator of a parameter of the data-generating distribution, defined as the minimizer of the expectation (risk) of a loss function over all candidate parameter values in the parameter space. This chapter demonstrates how the super learner framework can also be applied to estimate parameters such as conditional hazards or survival functions of a failure time, given a vector of baseline covariates, based on right-censored data. We strongly advise readers to familiarize themselves with the concepts presented in Chap. 3 before reading this chapter, as we assume the reader has a firm grasp of that material.

If the outcome of interest is time-to-event, then one is often interested in the survival function of the time-to-event. This allows one to answer questions such as, "What is the probability of having a recurrence of cancer within 5 years?" A survival function at a time point, such as 5 years, is defined as the probability that the survival time exceeds 5 years. Thus, the survival function is a monotone decreasing function in time, starting at 1 at time 0 ($t = 0$), and typically ending at 0 (assuming that every subject will eventually experience the event). Since survival functions will change as a function of characteristics of the subject, it is often of interest to understand the effect of treatment and baseline covariates on the conditional survival function at one particular time point, or on the whole conditional survival curve.

The hazard is defined as the instantaneous probability of the event occurring at time t, given the event has not occurred yet by time t. One inescapable feature of survival time data is that the time-to-event is almost always subject to right censoring: some subjects will drop out before the event can occur, or, at the endpoint of the study, a subject has not failed yet. The (conditional) hazard function provides the (conditional) survival function, and the conditional hazard can be estimated in the same manner as if there was no censoring. These two reasons provide an important motivation for the construction of estimators of the (conditional) hazard. As we will

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_16,

see, indeed, we can provide convenient loss functions for the hazard that naturally handle right censoring, without the requirement of incorporating an estimate of a censoring mechanism.

Since many target parameters, such as causal effects of a treatment on a survival time, are functions of the conditional hazard, the TMLE requires an initial estimator of the conditional hazard, and a corresponding targeted update of this conditional hazard, before mapping it into the desired target parameter. As a consequence, in order to obtain a TMLE of causal effects based on right-censored data structures, data-adaptive estimation of the conditional hazard is needed. In some applications, one might be interested in the density or hazard itself as well, as in this chapter.

In fact, the probability distribution of any longitudinal data structure can be expressed in terms of a product over time t of conditional probabilities of binary outcomes/indicators at time t, given a past string of events. The TMLE of a causal effect of a multiple-time-point intervention requires estimation of such conditional probabilities (for the nonintervention nodes of the SCM). Super learner for the conditional hazard presented here applies more generally to the estimation of such conditional probability functions for binary indicators. Longitudinal data are handled rigorously in Part VIII.

This provides more than enough motivation to devote a chapter to super learning of the conditional hazard in this book. As we learned in Chap. 3, the super learner requires defining a valid loss function, building a collection of candidate estimators, proposing a parametric family consisting of weighted combinations of the estimators in the collection, and computing the optimal weighted combination by minimizing the cross-validated risk of the loss function over all candidate weighted combinations of estimators.

16.1 Data Structure

Let T be the survival time and $W = (W_1, W_2, \ldots, W_p)$ a set of p baseline covariates. The full data structure is defined as $X = (T, W)$, and let $P_{X,0}$ denote its probability distribution known to belong to the statistical model $\mathcal{M}^F$. The survival time is possibly right censored by the censoring time C. The observed data structure is defined as

$$O = \left(W, \tilde{T} = \min(T, C), \Delta = \mathrm{I}(\tilde{T} = T)\right).$$

Let P_0 denote the true probability distribution of O. Note that the probability distribution of O is determined by the distribution $P_{X,0}$ of $X = (T, W)$ and the conditional distribution of C, given X.

We denote the conditional survival function of C by $G_0(t \mid X) = P_0(C > t \mid X)$ and its conditional density by $g_0(t \mid X) = P_0(C = t \mid X)$. We assume that the time

scale is discretized and we denote the time points by $t = 1, \ldots, \tau$. Let $N(t) = I(\tilde{T} \leq t, \Delta = 1)$ and $A(t) = I(\tilde{T} \leq t, \Delta = 0)$ be the counting processes that indicate if a failure time event and censoring event is observed at time t, respectively. Note that $N(t)$ is a process that starts at 0 at time $t = 0$, and jumps to 1 at time $\tilde{T}$ if a failure is observed (i.e., if $\Delta = 1$). Similarly, $A(t)$ jumps to 1 at time $\tilde{T}$ if a censoring event is observed (i.e., if $\Delta = 0$).

The likelihood of the data can be represented as a product over time t of the conditional probability of events one observes at time t, conditional on all the data observed up till time t. Thus, such a likelihood involves conditional probabilities of observing a jump of N (or A) at time t, conditional on the history up till time t, which we often refer to as the parents of $N(t)$ (or parents of $A(t)$). We will denote such an indicator of N (or A) jumping from 0 to 1 at time t by $dN(t)$ (or $dA(t)$) and its parents by $Pa(N(t))$ (or $Pa(A(t))$).

Let $Q_{dN(t)}$ denote the conditional distribution of $dN(t)$, given its parents $Pa(N(t)) = (\bar{N}(t-1), \bar{A}(t-1), W)$, and let $g_{dA(t)}$ denote the conditional distribution of $dA(t)$, given its parents $Pa(A(t)) = (\bar{N}(t), \bar{A}(t-1), W)$. The true probability distribution of O factorizes as

$$P_0(O = o) = P_{W,0}(W) \prod_{t=1}^{\tau} Q_{dN(t),0}(dN(t) \mid Pa(N(t))) \prod_{t=1}^{\tau} g_{dA(t),0}(dA(t) \mid Pa(A(t))).$$

The conditional distribution $Q_{dN(t),0}$ of the binary indicator $dN(t)$ is determined by $P_0(dN(t) = 1 \mid Pa(N(t))) = E_0(dN(t) \mid Pa(N(t)))$.

In the counting process literature, for counting processes that jump in continuous time, one refers to the instantaneous conditional probabilities $E(dN(t) \mid Pa(N(t))$, conditional on the history right before $N(t)$, as an intensity. Therefore, we will refer to $E_0(dN(t) \mid Pa(N(t))$ as a discrete intensity, since we assumed that events only occur on a discrete time scale. Note that these discrete intensities equal zero if $Pa(N(t))$ imply that $N(t)$ cannot jump anymore, i.e., if $\tilde{T} < t$. Let $\bar{Q}_0(t \mid W) = E_0(dN(t) \mid W, \tilde{T} \geq t)$, and $\bar{g}_0(t \mid W) = E_0(dA(t) \mid N(t) = A(t-1) = 0, W)$ denote the discrete intensities (conditioning on histories for which the counting process is at risk of jumping) of these two counting processes $N(t)$ and $A(t)$.

The statistical model for P_0 is implied by the statistical model $\mathcal{M}^F$ and a statistical model $\mathcal{G}$ for g_0. We assume coarsening at random (CAR) for the conditional distribution of C, given $X = (T, W)$, which will be referred to as the censoring mechanism. CAR is implied by the assumption that C is independent of T, given W. We note that, under CAR, these discrete intensities equal the conditional hazard of T, given W, and C, given W, respectively:

$$\bar{Q}_0(t \mid W) = P_0(T = t \mid T \geq t, W),$$
$$\bar{g}_0(t \mid W) = P_0(C = t \mid C \geq t, W).$$

Thus, under CAR we can also refer to these intensities as conditional hazards.

16.2 Parameters of Interest

For the remainder of this chapter, we consider the case that prediction is the target parameter of interest, not a causal effect; thus we use Ψ notation to refer to this function. We consider two common parameter of interests for survival outcomes.

1. The first is a conditional expectation of a user-supplied function of T, given the baseline covariates, $\psi_0(W) = \mathrm{E}_0(m(T) \mid W)$ for a user-supplied function $m(T)$. Here possible choices for $m(T)$ are given by $m(T) = T$, $m(T) = \log T$, and $m(T) = I(T > t_0)$ for some time point t_0. Since many distributions of a survival time T are skewed, and since T is often not observed in the tail of its distribution due to right censoring, it is often argued that the mean of T is not as much of interest as other location parameters such as the median of T or a truncated mean. A truncated mean can be obtained by defining $m(T)$ as a truncated version of T. One can also simply truncate T and focus on the mean of the truncated T. Since the density of a log-survival time T is often more symmetrically distributed, the mean of $\log T$ is often viewed as an interesting parameter, possibly transformed back to the T-scale. The choice $m(T) = I(T > t_0)$ is naturally of interest since it provides the conditional survival function.
2. The second parameter of interest we consider is the conditional hazard (or conditional density) $\psi_0(t \mid W) = \bar{Q}_0(t \mid W)$, even though its main application might be to map it into its corresponding survival function.

In both cases ψ_0 is a parameter of the full-data distribution $P_{X,0}$, and only through the Q_0-factor, so that $\Psi(P_0) = \Psi^F(Q_0)$. The parameter space for this parameter is implied by the full-data model $\mathcal{M}^F$: $\mathbf{\Psi} = \{\Psi(P) : P \in \mathcal{M}\} = \{\Psi^F(Q(P_X)) : P_X \in \mathcal{M}^F\}$.

16.3 Cross-Validation for Censored Data Structures

Estimator selection based on cross-validation for censored data structures is extensively examined in van der Laan and Dudoit (2003). Suppose the parameter ψ_0 of the full-data distribution can be defined as a minimizer of the expectation of a full-data loss function:

$$\psi_0 = \mathrm{argmin}_\psi \int L(x, \psi)\, dP_{X,0}(x).$$

In order to apply loss-based cross-validation and, in particular, the loss-based super learner, we need to construct an observed data loss function $L(O, \psi)$ of the observed data structure O so that $E_0 L(O, \psi) = E_0 L(X, \psi)$.

If the parameter of interest is the conditional mean, $\psi_0 = \mathrm{E}(m(T) \mid W)$, a commonly used loss function is the IPCW squared error loss function:

$$L(O, \psi) = \frac{\Delta}{\bar{G}_0(T \mid X)} \{m(T) - \psi(W)\}^2, \tag{16.1}$$

where $\bar{G}_0(T \mid W) = P_0(C > t \mid W)$ is the conditional survival function of censoring time C, given W. Since the censoring mechanism is often not known, $\bar{G}_0$ is an unknown nuisance parameter for the loss function. If the parameter of interest is the conditional hazard function, $\psi_0 = \bar{Q}_0(\cdot \mid W)$, we have the following two possible loss functions:

$$L_{loglik}(O, \psi) = \sum_t \mathrm{I}\left(\widetilde{T} \geq t\right) \log(\psi(t \mid W))^{dN(t)} \log(1 - \psi(t \mid W))^{1-dN(t)},$$

$$L_{L_2}(O, \psi) = \sum_t \mathrm{I}\left(\widetilde{T} \geq t\right) \{dN(t) - \psi(t \mid W)\}^2 .$$

In this case, neither loss function is indexed by an unknown nuisance parameter.

Given an observed data loss function, the cross-validation selector to select among candidate estimators $\hat{\Psi}_k$ of ψ_0, $k = 1, \ldots, K$, is defined as before in Chap. 3, with the only remark that, if the loss function depends on a nuisance parameter, then one needs to plug in an estimator of this nuisance parameter. If censoring is known to be independent, then one could estimate the marginal survivor function $\bar{G}_0(t) = P(C > t)$ with the Kaplan–Meier estimator defined as:

$$\bar{G}_{KM,n}(t) = \prod_{s \leq t} \left(1 - \frac{\sum_{i=1}^n I(\tilde{T}_i = s, \Delta_i = 0)}{\sum_{i=1}^n I(\tilde{T}_i \geq s)}\right).$$

If, on the other hand, such knowledge is not available, then one can use a machine learning algorithm, such as a super learner, to estimate the conditional hazard $\bar{g}_0(t \mid W)$ of C, given W, using one of the two loss functions $L_{loglik}(O, \bar{g}_0)$ or $L_{L_2}(O, \bar{g}_0)$, to construct a data-adaptive estimator of $\bar{g}_0(t \mid W)$.

To construct the super learner to estimate our parameter of interest we require the following ingredients:

1. A collection of candidate estimators of the parameter of interest ψ_0,
2. A loss function $L(O, \psi)$,
3. A parametric statistical model for combining the estimators in the collection.

The candidate estimators are not restricted to being based on the loss function used in the cross-validation selector. However, for the cross-validation one wishes to use a loss function whose dissimilarity

$$d(\psi, \psi_0) = E_0 L(O, \psi) - E_0 L(O, \psi_0),$$

directly measures a discrepancy between the true target parameter ψ_0 and candidate ψ. For example, if one is concerned with estimation of the conditional survival function $\psi_0(W) = P_0(T > t_0 \mid W)$, then the IPCW loss function presented in (16.1) provides a more direct measure of fit of ψ_0 than the log-likelihood or squared error loss function for the conditional hazard, even though all three loss functions are valid loss functions.

This may seem like common sense, but researchers often use a hazard loss function when the interest is on the survival probability at a specific time point (e.g., 10-year survival). The loss function should be chosen with respect to the problem you are trying to solve.

The oracle inequality results for the cross-validation selector, and thereby for the super learner, assume the loss function is bounded. Therefore, the parametric statistical model for combining the estimators in the collection needs to be chosen to maintain a bounded loss function among all candidates that can be constructed as weighted combinations of the candidate estimators. We propose constraints on the parametric statistical model for combining the algorithms to maintain a bounded loss function.

For the L_2 loss function, we can use as parametric statistical model

$$\left\{\sum_k \alpha_k \psi_k : \sum_k \alpha_k = 1, \alpha_k \geq 0\right\}.$$

Here we constrained the space for the weight vector α to be $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$. In this case, it is easy to show that the loss function is uniformly bounded by the bound on the parameter space $\mathbf{\Psi}$ and the number of time points τ.

For the log-likelihood loss function we propose using the logit link function:

$$g(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right),$$

using the convex combination on the logit scale, and then transforming it back into a probability with the inverse logit function. That is, given a candidate estimator of the conditional hazard, $\psi_{k,n}$, we transform it into

$$g(\psi_{k,n}) = \log \frac{\psi_{k,n}}{1-\psi_{k,n}},$$

and we consider combinations $\sum_k \alpha_k g(\psi_{k,n})$, which corresponds with a hazard estimator $g^{-1}(\sum_k \alpha_k g(\psi_{k,n}))$. The advantage of this approach is that the logit of hazards, $g(\psi_{k,n})$, are not subject to any constraint, so that the α-vector does not need to be constrained to positive weights. In addition, it has numeric advantages since the minimizer of the cross-validated risk is now an interior point.

In our implementation of the super learner, enforcing $\sum_k \alpha_k = 1$ and $\alpha_k \geq 0$ is an option. In this case, the parametric family for combining the candidate estimators is given by:

$$\left\{g^{-1}\left(\sum_k \alpha_k g(\psi_k)\right): \quad \text{with } \sum_k \alpha_k = 1 \ \& \ \alpha_k \geq 0 \ \forall k\right\}.$$

A problem occurs when the conditional hazard ψ_k approaches either 0 or 1, in which case logit(ψ) approaches $\pm\infty$. To avoid this problem, we will use the symmetric truncated logit link,

$$g^*(x,c) = \begin{cases} g(c) & \text{if } x < c, \\ g(x) & \text{if } c \le x \le 1-c, \\ g(1-c) & \text{if } x > 1-c, \end{cases}$$

for a small constant c. It follows that, as long as c is selected such that $g(1-c) < M/\tau$, the loss will be bounded by M. In current implementations, we enforce small levels of truncations such as $c = 0.01$. It is of interest for future research to investigate data-adaptive strategies for setting such global bounds on the loss function.

16.4 Super Learner for Hazard Estimation in Lung Cancer

In this censored-data demonstration, we focus on the case where the full-data statistical model is nonparametric, and estimate the conditional hazard. We examined the North Central Cancer Treatment Group lung cancer data set (Loprinzi et al. 1994), available in the survival package in R (Therneau and Lumley 2009).

The data set contains the survival time (in days) for 228 patients with advanced lung cancer. In addition to the survival time, information on the patient's age, sex, and three performance scores was included. The parameter of interest was the hazard function given the patient's age, sex, and performance scores. Five patients were removed from the analysis set due to incomplete information on the covariates. With the 223 patients in the analysis set, 63 were right censored and 160 had observed death times. We used the squared error loss function on the hazard for the super learner:

$$L_{L_2}(O,\psi) = \sum_t \mathrm{I}\left(\widetilde{T} \ge t\right) \{dN(t) - \psi(W)\}^2 .$$

The first step was to convert the right-censored data structure $(W, \Delta, \tilde{T})$ into a longitudinal data structure collecting at time t the change in counting processes, $dN(t)$, $dA(t)$: $(W, (dN(t), dA(t) : t))$. A grid of 30 time points was created using the quantiles of the observed death times, and then $dN(t)$ was defined as the number of observed failures in the window containing t, and, similarly, $dA(t)$ was defined as the number of observed censoring events in this window.

The collection of estimators consisted of logistic regression, random forests, generalized additive models (gam), polyclass, deletion/substitution/addition algorithm (DSA), neural networks (nnet) , and a null statistical model using only time and no covariates. For most hazard estimation problems, time is one of the most important variables in the estimator. We considered the variable time with a few approaches. One was to use an indicator for each time point. Logistic regression in a parametric statistical model with an indicator for each time point will approximate the Cox

proportional hazards model if one uses an increasingly finer grid of time points. Another approach added an additional step of smoothing in time using a generalized additive statistical model. These estimators in the collection involved a two-stage estimation procedure.

In the first step, one of the algorithms is applied to fit the hazard as a function of time t and covariates W. This results in an estimator $\bar{Q}^0_{k,n}(t \mid W)$ for the kth estimator of the conditional hazard $\bar{Q}_0(t \mid W)$. Subsequently, these predicted probabilities $\bar{Q}^0_{k,n}(t \mid W)$ were used as an offset in a generalized additive logistic regression statistical model with m degrees of freedom for time:

$$\text{logit}P(dN(t) \mid \tilde{T} \geq t, W) = \bar{Q}_{k,n}(t \mid W) + s(t, m).$$

The degrees-of-freedom tuning parameter m for the smoothing spline $s(t, m)$ is not known a priori, but different values of m simply represent different estimators in the collection of algorithms defining the super learner. The estimators in the super learner collection of algorithms are the candidate estimators for binary outcome repeated measures regression, coupled with the generalized additive statistical models estimator for the time trend.

Table 16.1 contains a list of the estimators used in the super learner. The first column is the algorithm used for the covariates (including time t as one covariate), and the second column indicates if any additional step was taken to estimate the effect of time. For the sake of comparison, we also report the results for a regression fit in a parametric statistical model for the conditional hazard according to the Weibull proportional hazards model:

$$P(T = t \mid T \geq t, W) = \alpha t^{\alpha-1} \exp(\beta^\top W).$$

This parametric model assumes, in particular, that the hazard function is monotone in time t, and it includes the exponential distribution (i.e., constant hazard) as a special case.

The honest cross-validated risks are provided for each of the estimators in the collection of algorithms, and for the super learner algorithms. The reported risk of the Weibull estimator of the hazard was also cross-validated. An estimate of the standard error of this honest V-fold cross-validated risk is also provided, based on the variance estimator

$$\frac{1}{n^2} \sum_{v=1}^{V} \sum_{i \in \text{Val}(v)} \left(L(O_i, \hat{\Psi}(P_{n,\text{Tr}(v)}) - \bar{L} \right)^2,$$

where $\text{Val}(v)$ and $\text{Tr}(v)$ are a partition of $\{1, \ldots, n\}$ indicating the observations in the validation sample and training sample, respectively, for the vth split, and

$$\bar{L} = \frac{1}{n} \sum_{v} \sum_{i \in \text{Val}(v)} L(O_i, \hat{\Psi}(P_{n,\text{Tr}(v)})$$

Table 16.1 Honest 10-fold cross-validated risk estimates for the super learner, each algorithm in the collection, and the Weibull proportional hazards statistical model. All algorithms included time as a covariate. The second column (Time) denotes if any additional smoothing for time was part of the given estimator, and the value for the degrees of freedom, if used

Algorithm	Time	CV Risk	SE
Super learner		0.6548	0.0258
Discrete SL		0.6589	0.0261
glm	No smoothing	0.6534	0.0260
glm	df = 1	0.6534	0.0260
glm	df = 2	0.6541	0.0260
glm	df = 3	0.6548	0.0261
glm	df = 4	0.6556	0.0261
glm	df = 5	0.6564	0.0261
glm	Indicator	0.6700	0.0266
glm (2-way interactions)	df = 5	0.6569	0.0261
randomForest	No smoothing	0.7628	0.0313
randomForest	df = 2	1.0323	0.0607
randomForest	df = 3	1.0364	0.0627
randomForest	df = 4	1.0483	0.0628
randomForest	df = 5	1.0362	0.0608
gam	df = 2	0.6558	0.0260
gam	df = 3	0.6563	0.0260
gam	df = 4	0.6570	0.0261
gam	df = 5	0.6577	0.0261
gam(df = 2)	No smoothing	0.6554	0.0260
gam(df = 3)	No smoothing	0.6579	0.0261
gam(df = 4)	No smoothing	0.6619	0.0263
gam(df = 5)	No smoothing	0.6554	0.0260
gam (only time, df = 3)	No smoothing	0.6548	0.0257
gam (only time, df = 4)	No smoothing	0.6556	0.0257
gam (only time, df = 5)	No smoothing	0.6541	0.0256
polyclass	No smoothing	0.6570	0.0258
DSA	No smoothing	0.6671	0.0270
DSA	df = 5	0.6669	0.0269
nnet	No smoothing	0.7175	0.0302
Weibull PH model		0.7131	0.0300

denotes the cross-validated risk of the estimator $\hat{\Psi}$ (Dudoit and van der Laan 2005, Theorem 3).

The estimated coefficients for the super learner were

$$\begin{aligned}\Psi_{SL,n} = {} & 0.182\Psi_{n,\text{glm, no}} + 0.182\Psi_{n,\text{gam only time,df = 5}} \\ & +0.581\Psi_{n,\text{polyclass, no}} + 0.056\Psi_{n,\text{glm 2-way, df = 5}},\end{aligned}$$

where $\Psi_{n,a,b}$ represents the fit of algorithm a using the smoothing in time method b. The only estimators to receive nonzero weight in the final super learner fit were logistic regression using main terms and no smoothing, a gam statistical model us-

ing only time and 5 degrees of freedom, polyclass with no additional smoothing, and a logistic regression in a parametric statistical model with all two-way interaction (including time) combined with smoothing time using df = 5. Thus, of the 29 estimators in the library, only 4 received a nonzero weight.

16.5 Notes and Further Reading

Selection among candidate estimators of a target parameter for survival outcomes has received less attention compared to estimator selection for continuous and categorical (noncensored) outcomes. Notable examples of estimator selection for censored data include the cross-validation selector based on a double robust IPCW full-data loss function (van der Laan and Dudoit 2003). In van der Laan et al. (2004), the cross-validation selector based on the IPCW loss function is analyzed in detail.

A variety of methods have been proposed for nonparametric estimation of a conditional hazard based on right-censored data, involving likelihood-based cross-validation or penalized log-likelihood [e.g., LASSO in Tibshirani (1997) and Zhang and Lu (2007), or other penalties such as Akaike's information criterion in Akaike (1973)] to select fine-tuning parameters. For example, Hastie and Tibshirani (1990) proposed using additive Cox proportional hazards models with smoothing splines for the covariates. Kooperberg et al. (1995) similarly used polynomial splines to approximate the conditional hazard. Tree-based approximations of the conditional hazard have also been proposed, often referred to as survival trees or survival forests (LeBlanc and Crowley 1992; Segal 1988; Hothorn et al. 2006; Ishwaran et al. 2008). Cross-validated Cox regression is described in the context of penalized partial likelihoods in van Houwelingen et al. (2006). All these algorithms can be included in the library of the super learner to maximize its performance.