CONFORMALIZED SURVIVAL ANALYSIS: A REVIEW

COURSE PROJECT: STAT 2261- SURVIVAL ANALYSIS

Shubha Sankar Banerjee*
Department of Statistics
University of Pittsburgh

December 19, 2023

ABSTRACT

In this project we review discuss the methodology involving inference using conformal prediction proposed by Candès et al. (2023) which can be used with any survival prediction methods to produce calibrated, covariate-dependent lower predictive bounds on survival times. Since conformal prediction is a powerful but relatively new tool for inference, we discuss the basic concepts related to conformal prediction for more rudimentary models before motivating ourselves to apply similar techiques in the survival analysis regime. In this project we talk about generation of lower predictive bounds with almost exact coverage and discuss an algorithm adapted from equivalent regression model approach and its application in the survival model setup.

1 Introduction and Motivation

1.1 Introduction to Conformal Prediction

The goal of any prediction algorithm is to generate prediction sets for unknown responses based on observed covariates with a proper pre-determined level of coverage. More precisely, let $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R},\ i = 1, \ldots, n$ denote training data. We further assume that the $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. from an unspecified distribution $\mathcal{P}$. For a pre-determined coverage level $1 - \alpha \in (0, 1)$, we are motivated to construct a band $\hat{\Gamma}_n(.)$, based on the training data such that, for a new i.i.d. point (test point) (X_{n+1}, Y_{n+1}), we have,

$$\mathbb{P}\left[Y_{n+1} \in \hat{\Gamma}(X_{n+1})\right] \geq 1 - \alpha \tag{1}$$

where, the probability is taken over the $n + 1$ i.i.d. points $\{(X_i, Y_i)\}_{i=1}^{n+1} \sim \mathcal{P}$ and for a point $x \in \mathbb{R}^d$, and $\hat{\Gamma} : \mathbb{R}^d \to \{\text{intervals in } \mathbb{R}\}$. The aim behind conformal prediction (Vovk (2005)) is for (1) to hold without any assumptions on $\mathcal{P}$ and for constructing $\hat{\Gamma}$ which has finite-sample (non-asymptotic) *validity*. A confidence predictor is valid if in the long run, the relative frequency of errors does not exceed α.

An important property of conformal prediction is that it dynamically adjusts the prediction intervals for a new test point based on the observations in hand sequentially, which makes the problem valid and also prevents overfitting.

To highlight the requiement of a revised prediction interval generation method, we take a look at the following example from Lei et al. (2018). In the regression problem, we consider a naive method for prediction interval construction for Y_{n+1} based on X_{n+1}. Recall $\{(X_i, Y_i)\}_{i=1}^{n+1} \overset{i.i.d.}{\sim} \mathcal{P}$. If $\hat{\mu}$ is the estimator for the population regression function, a naive prediction interval can be as follows:

$$\Gamma_{naive}(X_{n+1}) = \left[\hat{\mu}(X_{n+1}) - \hat{F}_n^{-1}(1-\alpha), \hat{\mu}(X_{n+1}) + \hat{F}_n^{-1}(1-\alpha)\right] \tag{2}$$

where, $\hat{F}_n$ is the empirical distribution function of the fitted residual $|Y_i - \hat{\mu}(X_i)|,\ i = 1, \ldots, n$ and $\hat{F}_n^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile for $\hat{F}_n$. This procedure yields prediction intervals which are approximately valid but requires the

*Email: shubha.banerjee@pitt.edu

estimator $\hat{\mu}$ to be accurate enough for the estimated $\hat{F}_n^{-1}(1-\alpha)$ to be close to the $(1-\alpha)$-quantile of the population residuals $|Y_i - \mu(X_i)|$. As can be seen in literature, ensuring such a property requires appropriate regularity conditions on underlying data distribution $\mathcal{P}$ and $\hat{\mu}$, for example, a model which has been specified correctly and in case we are looking at model selection with constraints, the choice of a tuning parameter.

The naive method (2) generally yields narrower prediction intervals (Morzuch (2000)) and can lead to serious undercoverage problems. This happens since often, the fitted residual distribution is biased downwards. Conformal prediction intervals, by construction, overcomes the deficiencies of such intervals and are guaranteed to deliver proper finite-sample coverage without any assumptions on $\mathcal{P}$ and $\hat{\mu}$.

We will briefly talk about the important general steps in conformal prediction before moving on to the topic of our focus. First we denote $Quantile(\beta; F)$ to be the level β quantile of a distribution F, i.e., $Quantile(\beta; F) = \inf\{z : \mathbb{P}[Z \leq z] \geq \beta\}$ for $Z \sim F$. We allow the quantiles to be defined for distributions F on the augmented real line $\mathbb{R} \cup \{\infty\}$. For a dataset $v_{1:n} = \{v_1, \ldots, v_n\}$, we denote quantiles of the empirical distribution as: $Quantile(\beta; v_{1:n}) = Quantile(\beta; n^{-1}\sum_{i=1}^{n} \delta_{v_i})$, where δ_a denotes a point mass at a. Tibshirani et al. (2019) showed that if $V_1, \ldots, V_{n+1}$ are exchangeable random variables with no ties, then for any $\beta \in (0,1)$, we have,

$$\beta \leq \mathbb{P}\{V_{n+1} \leq Quantile(\beta; V_{1:n} \cup \{\infty\})\} \leq \beta + \frac{1}{n+1} \tag{3}$$

For sake of simplicity of notation, we define the following, $Z_i = (X_i, Y_i),\ i = 1, \ldots, n$ and $Z = \{Z_i\}_{i=1}^n$. We choose a score function $\mathcal{S}$ which depends on a chosen point (x, y) and the dataset Z. A low value of $\mathcal{S}((x,y), Z)$ indicates that the point (x, y) *conforms* to Z, with high value indicating that (x, y) is atypical relative to the points in Z. In the regression problem, $\mathcal{S}$ can be as follows: $\mathcal{S}((x,y), Z) = |y - \hat{\mu}(x)|$, where we implement a prediction algorithm on Z to get $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$, the fitted regression function. Now, given $x \in \mathbb{R}^d$, we construct the conformal prediction interval $\hat{\Gamma}_n(x)$ by repeating the following procedure $\forall y \in \mathbb{R}$. We calculate the *nonconformity scores*

$$V_i^{(x,y)} = \mathcal{S}(Z_i, Z_{1:n} \cup \{(x,y)\}),\ i = 1, \ldots, n,\ \&\ V_{n+1}^{(x,y)} = \mathcal{S}((x,y), Z_{1:n} \cup \{(x,y)\}) \tag{4}$$

we include y in $\hat{\Gamma}(x)$ if $V_{n+1}^{(x,y)} \leq Quantile(1-\alpha; V_{1:n}^{(x,y)} \cup \{\infty\})$, where, $V_{1:n}^{(x,y)} = \{V_1^{(x,y)}, \ldots, V_n^{(x,y)}\}$.

We get the following result from Vovk (2005) and Lei et al. (2018).

Theorem 1. *Assume that $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R},\ i = 1, \ldots, n+1$are exchangeable. For any score function $\mathcal{S}$ and any $\alpha \in (0,1)$, define the conformal band based on the first n samples) at $x \in \mathbb{R}^d$ by*

$$\hat{\Gamma}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq Quantile(1-\alpha; V_{1:n}^{(x,y)} \cup \{\infty\}) \right\} \tag{5}$$

where $V_i^{(x,y)}, i = 1, \ldots, n+1$ are defined as in (4), then $\hat{\Gamma}_n$ satisfies

$$\mathbb{P}\left\{ Y_{n+1} \in \hat{\Gamma}_n(X_{n+1}) \right\} \geq 1 - \alpha$$

Furthermore, if ties between $V_1^{(X_{n+1}, Y_{n+1})}, \ldots, V_{n+1}^{(X_{n+1}, Y_{n+1})}$ occur with probability 0, then this probability is upper bounded by $1 - \alpha + 1/(n+1)$.

This establishes the validity and exact coverage properties of conformal prediction which makes the method so powerful. Motivated by this, we shift our focus to using this method to deal with survival data.

1.2 Survival Analysis

In time-sensitive data on diagnosis of any disease to an event time which is generally fatality from the particular disease, it is of crucial importance to have a proper prediction of survival times based on a set of covariates which might be useful in alloction of resources to mitigate the fatal effect of the disease. One of the motivation behind survival analysis is to infer on survival function i.e. the probability that a patient will survive beyond any specified time. The survival times are often censored and thus it is an of consequence to come up with valid predictions for survival functions from censored data. The Kaplan-Meier curve can be used to generate estimates for survival function when population under study is group of patients with certain characteristics. While Kaplan-Meier curve does not make any distributional assumptions on the survival times, it requires suffiiently many events in each subgroups and thus can only be applied to a handful of subpopulations.

While point-predictions for survival times are of important, for decision-making in sensitive and uncertain environments, it is perhaps more useful to to have guaranteed coverages with prediction intervals for *uncensored* survival times. If

the intervals are wide, it reflects on an inherent lack of knowledge and a conservative coverage. Again, as stated in subsection 1.1, it is of interest to generate valid prediction bands for any survival analysis procedure with valid marginal coverage (which can also be extended to approximately valid conditional coverage), i.e., prediction bands which will contain the true value of the survival time on an average across all covariate configurations in the long run.

The paper by Candès et al. (2023), on which this project is based on, extends conformal inference to handle right-censored outcomes in the setting of Type-I censoring, where it is assumed that censoring time is observed for every unit while outcome is observed only for uncensored units. We learn about generation of a covariate-dependent lower prediction bound (LPB) on uncensored survival time (can be interpretated as a one-sided $(1-\alpha)$-prediction interval). It is worthwhille to note that we only focus on a LPB since we want to be more cautious with an earlier survival time prediction which might be beneficial when dealing with critical care, even though the process is a conservative analysis of survival time.

2 Prediction Intervals for survival times

For $i = 1, \ldots, n$, let X_i be the vector of covariates, C_i be the censoring time and T_i be the survival time for the i^{th} unit or patient. We further assume $\{(X_i, C_i, T_i)\}_{i=1}^n \overset{i.i.d.}{\sim} (X, C, T)$. We consider Type-I right-censoring, i.e., for each unit i, we observe the vector X_i, censoring time C_i and the censored survival time $\tilde{T}_i$, defined as follows:

$$\tilde{T}_i = \min\{T_i, C_i\}$$

For example, T_i measures time lapse between the admission into hospital and death, C_i measires time lapse betwen admission into hospital and the day data analysis is conducted and $\tilde{T}_i = T_i$ if the i-th patient died before the day of data analysis and $\tilde{T} = C_i$ if the patient survives beyond that day.

The problem with trying to perform an exact analysis of T is that in our observed data, some of the information is censored and all we observe is a censored time and thus, we proceed by imposing constraints on the relationship between T and C. We focus on two assumptions of conditionally independence censoring (Kalbfleisch and Prentice (2011)) and completely independent censoring.

Assumption 1 (Conditionally Independent Censoring)**.**

$$T \perp\!\!\!\perp C \mid X \tag{6}$$

This assumes that there exists no unmeasured confounders which affects the survival and censoring time. We may also assume a more stronger case where both survival time and covariates are independent of censoring time.

Assumption 2 (Completely Independent Censoring)**.**

$$(T, X) \perp\!\!\!\perp C \tag{7}$$

2.1 Naive Lower Prediction Bounds

We note that our main goal is to generate a covariate-dependent LPB for a conservative prediction of the uncensored survival time T. We first focus on the procedure of naively generating a LPB. Let $\hat{L}$ be a generic LPB estimated using th observed data $\{(X_i, C_i, T_i)\}_{i=1}^n$. We say that a LPB is *calibrated* if it satisfies the following coverage condition:

$$\mathbb{P}[T \geq \hat{L}(X)] \geq 1 - \alpha \tag{8}$$

where $\alpha \in (0, 1)$ and the probability is calculated both over $\hat{L}$ and future unit (X, C, T) which is independent of $\{(X_i, C_i, T_i)\}_{i=1}^n$.

The interesting part is to notice that since $\tilde{T} \leq T$, any calibrated LPB on the censored time $\tilde{T}$ is also a calibrated LPB o the uncensored survival time T. Thus, a naive LPB could be constructed using only the censored survival time $\tilde{T}$ (thereby disregarding the individual censoring times completely). There are many procedures in literature which rely on the i.i.d. property of the samples $(X_i, \tilde{T}_i)$ to construct distribution-free calibrated LPB on $\tilde{T}$ (see Vovk (2005), Lei et al. (2018) and the references therein).

The first theorem in the paper establishes that all distribution-free calibrated LPBs on T must by LPBs on $\tilde{T}$.

Theorem 2. *Take $X \in \mathbb{R}^p$ and $C \geq 0, T \geq 0$. Assume that $\hat{L}(.)$ is a calibrated LPB on T for all joint distributions of (X, C, T) obeying the conditionally independent censoring assumption (6) with X being continuous and (T, C) being continuous or discrete. Then for any such distribution,*

$$\mathbb{P}[\tilde{T} \geq \hat{L}(X)] \geq 1 - \alpha.$$

It can be intuitively understood that even if we have a calibrated LPB, a LPB constructed on $\tilde{T}$ must be somewhat conservative because of the censoring mechanism. To demonstrate this behaviour, we identify that the oracle LPB on $\tilde{T}$ is the α-th conditional quantile of $\tilde{T}|X$, which we denote by $\tilde{q}_\alpha(X)$. Similarly, let $q_\alpha(X)$ be the oracle LPB on T. Under conditionally independence censoring mechanism, we see the following:

$$\begin{aligned}\mathbb{P}[T \geq q_\alpha(x)|X=x] = 1-\alpha &= \mathbb{P}[\tilde{T} \geq \tilde{q}_\alpha(x)|X=x] \\ &= \mathbb{P}[T \geq \tilde{q}_\alpha(x)|X=x] \cdot \mathbb{P}[C \geq \tilde{q}_\alpha(x)|X=x] \qquad [\because \tilde{T} = \min\{T, C\}]\end{aligned}$$

The censoring mechanism arbitrarily affects the coverage of the naive LPB. In a very simple example, assume $T \sim Exp(1)$ and $C \sim Exp(\lambda)$, then, it is very easy to verify that $q_\alpha = -\log(1-\alpha)$ and $\tilde{q}_\alpha(X) = -\log(1-\alpha)/(1+\lambda)$. Thus, if the censoring times are small, there is a wider gap between $\tilde{q}_\alpha(x)$ and $q_\alpha(x)$. Thus, if we do not take into account the censoring times, we might end up with a very conservative LPB.

It is worthwhile to think that if we have smaller censoring time, more often than now we will have less observed survival time which will limit us to analyse the target variable properly and thus we end up having more conservative LPB to make up for the issue of more censored observations.

Theorem 2 states that only under conditionally independent censoring assumption is the calibrated LPB on T also a calibrated LPB on $\tilde{T}$. We proceed by making more assumptions on the distributions to try overcoming the just explained limitations of the naive LPB.

2.2 Leveraging the Censoring Mechanism

Since we now know that the problem with the naive approach is that with smaller censoring times it becomes more conservative, a plausible way to prevent this issue might be to discard units with small values of C. We consider a threshold c_0 and use this to extract a subpopulation on which $C \geq c_0$. We note that this selection leads to a distributional shift between the subpopulation and the whole population as in $(X, C, T) \overset{d}{\neq} (X, C, T)|C \geq c_0$.

We now study this distributional shift in more detail. The joint distribution of $(X, \tilde{T})$ on the whole population is $\mathcal{P}_X \times \mathcal{P}_{\tilde{T}|X}$ and that of the subpopulation is $\mathcal{P}_{(X,\tilde{T})|X\geq c_0} = \mathcal{P}_{X|C\geq c_0} \times \mathcal{P}_{\tilde{T}|X,C\geq c_0}$. We further observe that $\mathcal{P}_{\tilde{T}|X} \neq \mathcal{P}_{\tilde{T}|X,C\geq c_0}$ even under completely independent censoring (7) ($(T, X) \perp\!\!\!\perp C \not\Rightarrow \tilde{T} \perp\!\!\!\perp C\,|X$). What it means is that both covariate distribution and the conditional distribution of $\tilde{T}|X$ is different in the two populations.

We consider a modified secondary censoring scheme where the outcome is $\tilde{T} \wedge c_0$ ($a \wedge b = \min\{a, b\}$). It is very clear to see the following:

$$\begin{aligned}\mathcal{P}_{(X,\tilde{T}\wedge c_0)|C\geq c_0} = \mathcal{P}_{X|C\geq c_0} \times \mathcal{P}_{\tilde{T}\wedge c_0|X,C\geq c_0} &= \mathcal{P}_{X|C\geq c_0} \times \mathcal{P}_{T\wedge c_0|X,C\geq c_0} \quad [\because T \wedge c_0 = \tilde{T} \wedge c_0, \text{ if } C \geq c_0] \\ &= \mathcal{P}_{X|C\geq c_0} \times \mathcal{P}_{T\wedge c_0|X} \qquad [\text{Assumption (6)}] \end{aligned} \tag{9}$$

Further we can see the on the whole population the distribution of $(X, T \wedge c_0)$ can be written as,

$$\mathcal{P}_{(X,T\wedge c_0)} = \mathcal{P}_X \times \mathcal{P}_{T\wedge c_0|X} \tag{10}$$

Thus clearly from Equations (10) and (9), introduction of a secondary censoring scheme on the subpopulation only leads to a *covariate shift*.

Following Tibshirani et al. (2019), we take a look at the likelihood ratio between the two covariate distribution and use it to carefully reweight the samples to adjust for the bias induced by the distribution shift between the selected samples and the target population.

$$\frac{d\mathcal{P}_X}{d\mathcal{P}_{X|C\geq c_0}}(x) = \frac{\mathbb{P}[X=x]}{\mathbb{P}[X|C\geq c_0]} = \frac{\mathbb{P}[C\geq c_0]}{\mathbb{P}[C\geq c_0|X=x]} \tag{11}$$

We will apply one-sided version of weighted conformal inference (Tibshirani et al. (2019)) which gives a calibrated LBP on $T \wedge c_0$ and thereby a calibrated LPB on T. This effectively allows us to choose a large threshold c_0 if we have sufficiently many units with large values of C to reduce loss of power caused by censoring.

From here on, we will refer to the denominator of (11), i.e., $\mathbb{P}[C \geq c_0|X = x]$ as the *censoring mechanism* and denote it by $c(x; c_0)$. This is the conditional survival function of C evaluated at c_0. It is easy to see that under Type-I censoring, the C_i's are fully observed while T_i's are partially observed and thus estimation of the conditional survival function of C is a relatively easier task.

3 Conformal Inference for censored outcomes

3.1 Weighted Conformal Inference

Understanding the relationship between Eq. (9) and (10), we now want to construct LPB $\hat{L}(.)$ on $T \wedge c_0$ from the training samples $\{(X_i, \tilde{T}_i \wedge c_0)_{C_i \geq c_0}\} = \{(X_i, T_i \wedge c_0)_{C_i \geq c_0}\}$ such that $\mathbb{P}[T \wedge c_0 \geq \hat{L}(X)] \geq 1 - \alpha$. Note that since $T \wedge c_0 \leq T$, $\hat{L}(.)$ is also a calibrated LPB on T. We consider c_0 to be fixed for now, later the paper discusses a data-adaptive approach to choosing an optimal c_0.

The interesting part of this paper is the way of dealing with covariate shifts using technique introduced in Tibshirani et al. (2019). It follows from an intuitive idea: Assume $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{P}_X \times \mathcal{P}_{Y|X}$ and our objective is to construct prediction intervals for test points drawn from the target distribution $\mathcal{Q}_X \times \mathcal{P}_{Y|X}$, then using weighted conformal inference, we get prediction intervals $\hat{\Gamma}(.)$ such that $\mathbb{P}_{(X,Y)\sim\mathcal{Q}_X \times \mathcal{P}_{Y|X}}[Y \in \hat{\Gamma}] \geq 1 - \alpha$, with probability taken over both training set and test point (X, Y) and assuming that we know $w(x) = d\mathcal{Q}_X(x)/d\mathcal{P}_X(x)$. In our case, the target is $T \wedge c_0$ and the covariate shift is $w(x) = \mathbb{P}[C \geq c_0]/c(x, c_0)$ as we get in (11).

Algorithm 1 Weighted Conformalized Survival Analysis

Input: Level α; Data $\mathcal{Z} = (X_i, \tilde{T}_i, C_i)_{i \in \mathcal{I}}$; Testing point x;

Function $V(x, y; \mathcal{D})$ to compute the conformity score between (x, y) and data $\mathcal{D}$;
Function $\hat{w}(x, \mathcal{D})$ to fit the weight function at x using $\mathcal{D}$;
Function $\mathcal{C}(\mathcal{D})$ to select the threshold c_0 using $\mathcal{D}$.

Procedure:

1. Split $\mathcal{Z}$ into training fold $\mathcal{Z}_{tr} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{tr}}$ and a calibration fold $\mathcal{Z}_{ca} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{ca}}$.
2. Select $c_0 = \mathcal{C}(\mathcal{Z}_{tr})$ and let $\mathcal{I}'_{ca} = \{i \in \mathcal{I}_{ca} : C_i \geq c_0\}$.
3. For each $i \in \mathcal{I}'_{ca}$, compute the conformity score $V_i = V(X_i, \tilde{T}_i \wedge c_0; \mathcal{Z}_{tr})$.
4. For each $i \in \mathcal{I}'_{ca}$, compute the weight $W_i = \hat{w}(X_i; \mathcal{Z}_{tr}) \in [0, \infty)$.
5. Compute the weights $\hat{p}_i(x) = \frac{W_i}{\sum_{i \in \mathcal{I}'_{ca}} W_i + \hat{w}(x; \mathcal{Z}_{tr})}$ and $\hat{p}_\infty(x) = \frac{\hat{w}(x; \mathcal{Z}_{tr})}{\sum_{i \in \mathcal{I}'_{ca}} W_i + \hat{w}(x; \mathcal{Z}_{tr})}$.
6. Compute $\eta(x) = Quantile\left(1 - \alpha; \sum_{i \in \mathcal{I}'_{ca}} \hat{p}_i(x)\delta_{V_i} + \hat{p}_\infty(x)\delta_\infty\right)$.

Output: $\hat{L}(x) = \inf\{y : V(x, y; \mathcal{Z}_{tr}) \leq \eta(x)\} \wedge c_0$

In Algorithm 1, if the covariate shift $w(x)$ is unknown, it is estimated using the training fold. The algorithm also works in extreme cases, for example, if $\hat{w}(x; \mathcal{Z}_{tr}) = \infty$, then, by definition, $\hat{p}_i(x) = 0$ $(i \in Z_{ca})$ and $\hat{p}_\infty(x) = 1$, which yields the estimated LPB $\hat{L}(x) = -\infty$. By construction, we can also see that even though $\mathcal{Q}_X$ is not absolutely continuous with respect to $\mathcal{P}_X$ $(X_i \sim \mathcal{P}_X)$, we have $W_i \in [0, \infty)$.

It is also clear, from the construction of the weights $\hat{p}_i$ that $\eta(x)$ is invariant to positive rescalings of $\hat{w}(x)$. Thus, we can set $w(x) = 1/\hat{c}(x; c_0)$.

3.2 Choice of conformity scores

Although Algorithm 1 does not require a specific conformity score, we discuss three popular choices for $V(x, y; \mathcal{D})$ from literature:

- *Conformalized Mean Regression (CMR)*- The conformity scores are defined as $V(x, y; \mathcal{Z})_{tr} = \hat{m}(x) - y$. Here, $\hat{m}(.)$ is the estimate of the conditional mean of Y given X. The resulting LPB is $(\hat{m}(x) - \eta(x)) \wedge c_0$.
- *Conformalized Quantile Regression (CQR)*- The conformity scores are defined as $V(x, y; Z + tr) = \hat{q}_\alpha(x) - y$. Here, $\hat{q}_\alpha(.)$ is an estimate of the α-th conditional quantile of Y given X. The resulting LPB is $(\hat{q}_\alpha(x) - \eta(x)) \wedge c_0$. This is more adaptive and robust than CMR and has better conditional coverage.

- *Conformalized Distribution Regression (CDR)*- The conformity scores are defined by $V(x, y; Z_{tr}) = \alpha - \hat{F}_{Y|X=x}(y)$. Here, $\hat{F}_{Y|X=x}$ is the estimate of conditional distribution of Y given X. The resulting LPB is $\hat{F}^{-1}_{Y|X=x}(\alpha - \eta(x)) \wedge c_0$ (i.e., the $(\alpha - \eta)$-th quantile of Y given X).

Further, under completely censoring assumption, we can see that $\mathbb{P}[C \geq c_0|X] = \mathbb{P}[C \geq c_0]$ almost surely. Thus, based on the discussion above, we can easily set $\hat{w}(x) = w(x) \equiv 1$. This in turn yields calibrated LPB without any distributional assumption.

Theorem 3. *Let c_0 be any threshold independent of $\mathcal{Z}_{ca}$. Consider Algorithm 1 with $Y_i = T_i \wedge c_0$ and $\hat{w}(x; \mathcal{D}) \equiv 1$. Under the completely independent censoring assumption, $\hat{L}(X)$ is calibrated.*

4 Discussion

4.1 Doubly Robust Prediction Bounds

As we can see, the estimation of the censoring mechamism is not an issue under the more difficult assumption of independent complete censoring; however it does needs to be estimated if we work under the conditionally independent censoring regime. We can apply any technique in literature for estimation of $c(x; c_0) = \mathbb{P}[C \geq c_0|X = x]$ (for eg. kernel methods or distribution boosting, see papers cited in Candès et al. (2023)).

An important result based on the prediction problem which requires estimation of both of the above mentioned quantities is one stated for two-sided weighted split-CQR prediction bounds in Lei and Candès (2021). It states that in this case, the intervals satisfy a *doubly robust property* which states the following: the average coverage is guaranteed if either the covariate shift or the conditional quantiles are estimated well, and conditional coverage is guaranteed if the latter is true.

In this paper, the authors state asymptotic results (with elaborate extensions in Supplementary materials) for CQR-LPB and CDR-LPB under model oriented setups. Under proper regularity conditions on the censoring mechanism and the conditional quantiles (and conditional distribution in case of CDR-LPB), both CQR-LPB and CDR-LPB satisfy the doubly robust property.

This is an important result since now a researcher can deal with well enough estimation of conditional survival function and censoring mechanism without having concern for their relative accuracy. The resulting LPB will be calibrated (even if model, for eg. Cox model, is misspecified).

4.2 Choice of Threshold

As we can clearly see, the threshold c_0 is used to decrease the bias induced by low censoring times but at the same time also controls the size of data we are left with in our study. A large c_0 will reduce the gap between the target outcome T and $T \wedge c_0$ while reducing the sample size required to estimate the censoring mechanism and the conditional survival function.

We can, in principle, select any value for c_0 and get a calibrated LPB. Since we want to be more rigorous, we select a data-driven threshold to get accurate results.

The idea is to select c_0 based on the training fold $\mathcal{Z}_{tr}$ to make it independent of the calibration fold. Then, we can choose c_0 applying the following steps:

1. Set a grid of values for c_0.
2. Randomly sample a holdout set from $\mathcal{Z}_{tr}$
3. Apply Algorithm 1 on the rest of $\mathcal{Z}_{tr}$ for each value of c_0 to generate LPBs for each unit in the holdout set.
4. Select c_0 which maximizes the average LPBs on the holdout set.

Under suitable conditions, we can choose c_0 by using calibration fold $\mathcal{Z}_{ca}$ and the resulting LPBs will still be calibrated. To be specific, given a candidate set $\mathcal{C}$ for c_0, we get:

$$\hat{c}_0 = \arg\max_{c_0 \in \mathcal{C}} \frac{1}{|\mathcal{I}_{ca}|} \sum_{i \in \mathcal{I}_{ca}} \hat{L}_{c_0}(X_i)$$

Note that we choose the c_0 which maximizes the above function since we want a less conservative LPB which is a step towards generation of a more exactly calibrated LPB (one with almost exact coverage). This ia very computationally intensive task.

4.3 Simulation Results

This paper takes into consideration diverse setups to demonstrate the empirical validation of the theoretical results. CQR- and CDR-LPB is compared with the following alternatives: Cox model, Accelerated Failure Time (AFT) model, censored quantile regression, censored quantile regression forest and naive CQR. In each experiment, 200 independent datasets were generated, with training and test sizes being $n = 3000$ in each dataset.

The covariate vector $X \in \mathbf{R}^p$ is generated from $\mathcal{P}_X$, the survival time T is generated from an AFT model with Gaussian noise, i.e.,

$$\log T|X \sim \mathcal{N}(\mu(X), \sigma^2(X))$$

The following parameters were set up for the study.

	Dimension p	$\mathcal{P}_X$	$\mathcal{P}_{C\|X}$	$\mu(x)$	$\sigma(x)$
Uvt. + Homosc.	1	$\mathcal{U}(0,4)$	$\mathcal{E}(0.4)$	$2+0.37\sqrt{x}$	1.5
Uvt. + Heterosc.	1	$\mathcal{U}(0,4)$	$\mathcal{E}(0.4)$	$2+0.37\sqrt{x}$	$1+x/5$
Mvt. + Homosc.	100	$\mathcal{U}([-1,1]^p)$	$\mathcal{E}(0.4)$	$\log 2+1+0.55(x_1^2-x_3x_5)$	1
Mvt. + Heterosc.	100	$\mathcal{U}([-1,1]^p)$	$\mathcal{E}(0.4)$	$\log 2+1+0.55(x_1^2-x_3x_5)$	$\|x_{10}\|+1$

Here, 'Homosc.' and 'Heterosc.' are short for homoscedastic and heteroscedastic; 'Uvt.' and 'Mvt.' are short for univariate and multivariate. $\mathcal{U}(a,b)$ denotes uniform distribution on $[a,b]$; $\mathcal{E}(\lambda)$ denotes exponential distribution with rate λ.

For each dataset, the empirical coverage of the LPBs $(1/n_{test})\sum_{i=1}^{n_{test}} \mathbf{1}\{T_i \geq \hat{L}(X_i)\}$ is reported. $c(x)$ is estimated using distribution boosting and the target coverage level is $1-\alpha = 90\%$. We start by noting that in Figure 1, 'CQR-cRF'

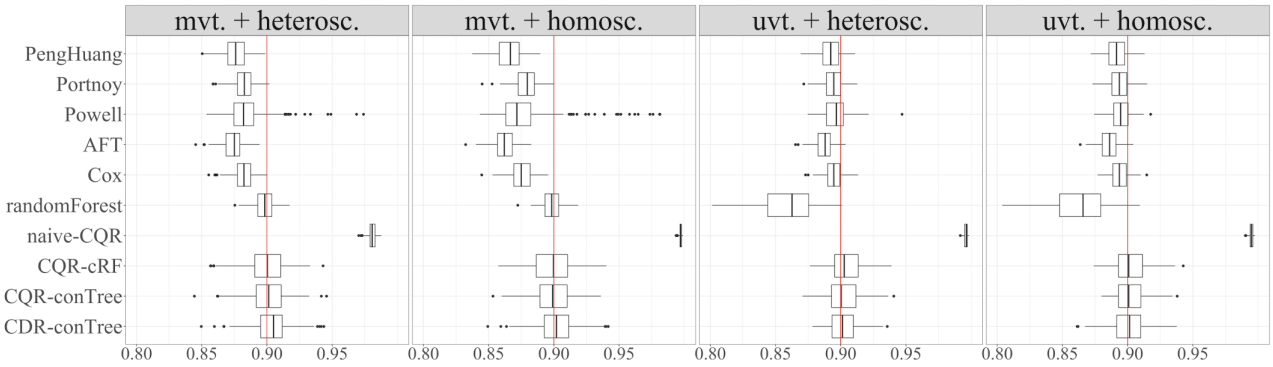

Figure 1: Empirical 90% coverage of uncensored survival time.

is short for CQR-LPB with censored quantile regression forest; 'CQR-conTree' and 'CDR-conTree' are short for CQR- and CDR-LPB with distribution boosting.

We can clearly see that naive CQR is overly conservative, but, both CQR- and CDR-LPB achieve near-exact marginal coverage. Standard models like Cox model and AFT suffer from undercoverage problems.

5 Conclusion

Conformalized Survival Analysis serves as a powerful tool for generation of prediction bands with near exact coverage. This is specially useful when dealing with events (typically death due to some disease) which require preventive action which are time-sensitive in nature. This method can extended to consider both end-of-study censoring caused by trial termination and loss-to-follow-up censoring caused by unexpected attrition. It can be also extended in dealing with prediction of counterfactual survival times had the cohort been exposed to a different condition. The process yields LPBs with near exact empirical coverage as seen in the simulation study and in case of critical events is useful in yielding LPBs which are not highly conservative which would lead to underestimating the true LPB, which is a problem, as seen, which affects standard techniques. The consideration of right-censoring is a constrained problem. An area of interest might be to extend this work to deal with random censoring and generation of prediction bands for uncensored survival times which are censored in this regime.

References

Candès, E., Lei, L., and Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45.

Kalbfleisch, J. and Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, L. and Candès, E. J. (2021). Conformal Inference of Counterfactuals and Individual Treatment Effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938.

Morzuch, B. (2000). Narrowness of prediction intervals revisited. *20th Annual Symposium on Forecasting, Lisbon, Portugal*.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Neural Information Processing Systems*.

Vovk, V. (2005). *Conformal prediction*, pages 17–51. Springer US, Boston, MA.