Conformalized Survival Analysis[1]: A Review

Emmanuel J. Candès, Lihua Lei, Zhimei Ren

Shubha Sankar Banerjee

Department of Statistics
University of Pittsburgh

December 4, 2023

[1] Main Reference: Candès et al. (2023)

Contents

Prediction Intervals

- The goal of any prediction algorithm is to generate prediction sets for unknown responses based on observed covariates with a pre-determined level of coverage.
- Assume $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$, where P is an unspecified distribution and $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$.
- For a chosen coverage level $(1-\alpha) \in (0,1)$, we want to construct a band $\hat{\Gamma}$, based on training data such that for a new i.i.d. test point (X_{n+1}, Y_{n+1}), we have,

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

- We will call a confidence predictor $\hat{\Gamma}$ to be *valid* if Eq. (1) holds.

What is Conformal Inference?

- Conformal Inference is a procedure which is used to construct prediction bands as in Eq. (1) that have finite-sample (non-asymptotic) validity.
- It is a powerful method since it focuses on a distribution-free approach to the prediction problem.
- Under some assumptions, this method can yield prediction sets with *exact* validity.
- Used to dynamically adjust prediction intervals for a new test point based on observations in hand sequentially, making the problem immune to overfitting.

Naive Approach to Prediction Interval Construction

- In regression setting, let $\{(X_i, Y_i)\}_{i=1}^{n+1} \stackrel{i.i.d.}{\sim} P$ and let $\hat{\mu}$ be estimator of population regression function.
- A naive prediction interval:

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

- $\hat{F}_n$ is the empirical distribution function of the fitted residual $|Y_i - \hat{\mu}(X_i)|$ and $\hat{F}_n^{-1}(1-\alpha)$ is $(1-\alpha)$-quantile for $\hat{F}_n$.
- Approximately valid procedure but requires $\hat{\mu}$ to be accurate enough. Requires appropriate regularity conditions on underlying data distribution P and μ.
- Generally yields narrower prediction intervals and leads to undercoverage problems.

General Conformal Inference Procedure

- Let $Z_i = (X_i, Y_i)$ and $Z = \{Z_i\}_{i=1}^n$ be the data and (x, y) be a test point.
- Choose a score function S such that a low value of $S((x,y), Z)$ indicates that the point (x, y) *conforms* to Z.
- Calculate the *nonconfomity scores* $\forall y \in \mathbb{R}$:

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

- Include y in $\hat{\Gamma}(x)$ if $V_{n+1}^{(x,y)} \le Quantile(1-\alpha; V_{1:n}^{(x,y)} \cup \{\infty\})$.

Theorem

Assume $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$ and that ties between $V_j^{(X_{n+1}, Y_{n+1})}$, $j = 1, \ldots, n+1$ occur with probability 0. Define conformal band based on first n samples at $x \in \mathbb{R}^d$ by

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

then,

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

Survival Analysis

- In time-sensitive data on diagnosis of any disease to an event time, it is of crucial importance to have a proper prediction of survival times based on a set of covariates.
- Proper inference is difficult since often survival times are censored.
- It is of interest to have guaranteed coverages with prediction intervals for *uncensored* survival times.
- Candès et al. (2023) extends conformal inference to handle Type-I right censoring.
- The goal is to generate distribution-free covariate-dependent lower predictive bounds (LPB) on uncensored survival time.

Setting the Problem

- For $i = 1, \ldots, n$, let X_i be the vector of covariates, C_i the censoring time and T_i be the survival time for the i^{th} unit.
- Assume $\{(X_i, C_i, T_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} (X, C, T)$.
- In Type-I right censoring, we observe $X_i,\ C_i$ and censored survival time $\tilde{T}_i = \min\{T_i, C_i\}$.

Assumption (Conditionally Independent Censoring)

$$T \perp\!\!\!\perp C \,|X$$

Assumption (Completely Independent Censoring)

$$(T, X) \perp\!\!\!\perp C$$

Naive Lower Prediction Bound

- Let's look at the issues with generating naive LPB. A LPB $\hat{L}(.)$ is *calibrated* is the following holds:

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

- Note that, since $\tilde{T} \leq T$, a calibrated LPB on censored time $\tilde{T}$ is also a calibrated LPB on uncensored survival time T.
- Naive approach: Use any distribution-free prediction approach (eg. see Vovk et al. (2022), Lei et al. (2016)) and generate calibrated LPB on $\tilde{T}$.

Theorem

Assume that $\hat{L}(.)$ is a calibrated LPB on T, with (X, C, T) obeying conditionally independent censoring assumption, then

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

- Intuitively clear that LPB on $\tilde{T}$ will be conservative if our target is LPB on T.
- Note that under conditionally independent censoring regime,

$$\mathbb{P}[T \geq q_\alpha(x)|X = x] = 1 - \alpha = \mathbb{P}[T \geq \tilde{q}_\alpha(x)|X = x] \cdot \mathbb{P}[C \geq \tilde{q}_\alpha(x)|X = x]$$

- Distance between q_α and $\tilde{q}_\alpha$ increases with smaller censoring times.
- If data is dominated with units with small censoring times, the resulting LPB can be arbitrarily conservative.
- The previous theorem highlights the validity of the LPB only on conditionally independent censoring, which is a weaker assumption. We need to involve more assumptions to overcome the limitation of the naive LPB.

Leveraging Censoring Mechanism

- Since smaller censoring times make the LPB more conservative, maybe discard units with small censoring times.
- Extract a subpopulation on which $C \geq c_0$, for a selected c_0 noting that $(X, C, T) \overset{d}{\neq} (X, C, T) | C \geq c_0$.
- $P_{(X,\tilde{T})} = P_X \times P_{\tilde{T}|X}$ and $P_{(X,\tilde{T})|C \geq c_0} = P_{X|C \geq c_0} \times P_{\tilde{T}|X, C \geq c_0}$. Working in this setup is untractable even under completely independent censoring.
- Consider a secondary censoring scheme where outcome is $\tilde{T} \wedge c_0$.
- Under conditionally independent censoring,

$$P_{(X,\tilde{T} \wedge c_0)|C \geq c_0} = P_{X|C \geq c_0} \times P_{T \wedge c_0|X}$$

- On the whole population, distribution can be written as:

$$P_{X,T \wedge c_0} = P_X \times P_{T \wedge c_0|X}$$

- Clearly, the secondary censoring scheme on the subpopulation leads to a tractable problem of *covariate shift*.

- The likelihood ratio between the two covariate distributions is:

$$\frac{dP_X}{dP_{X|C\geq c_0}}(x) = \frac{\mathbb{P}[C \geq c_0]}{\mathbb{P}[C \geq c_0 | X = x]}$$

- This special form of the distribution shift allows us to adjust for the bias by carefully reweighting the samples.
- Following Tibshirani et al. (2019), we can get LPB on $T \wedge c_0$ which is also calibrated LPB on T.
- Referring $\mathbb{P}[C \geq c_0 | X = x] = c(x; c_0)$ as the *censoring mechanism*, we realize that it makes the overall problem easily estimable.
- With sufficiently many samples with large C, we can choose a larger threshold to reduce the loss of power induced by censoring.

Weighted conformal inference

- Based on the previous discussions, the goal now is to construct LPB $\hat{L}(.)$ on $T \wedge c_0$ from training samples $(X_i, \tilde{T}_i \wedge c_0)_{C \geq c_0} = (X_i, T_i \wedge c_0)_{C \geq c_0}$ such that
$$\mathbb{P}[T \wedge c_0 \geq \hat{L}(X)] \geq 1 - \alpha$$
- To deal with covariate shifts, following Tibshirani et al. (2019), we use weighted conformal inference.
- **Idea**: Suppose training samples $(X_i, Y_i)_{i=1}^n \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$ and we wish to construct prediction intervals for test points drawn from $Q_X \times P_{Y|X}$. Assuming $w(x) = dQ_X(x)/dP_X(x)$ is known, we generate $\hat{\Gamma}$ such that
$$\mathbb{P}_{(X,Y) \sim Q_X \times P_{Y|X}}[Y \in \hat{\Gamma}] \geq 1 - \alpha$$
- In our case, the outcome is $T \wedge c_0$ and covariate shift is $w(x) = [C \geq c_0 | X = x]/c(x; c_0)$.

Algorithm 1 Weighted Conformalized Survival Analysis

Input: Level α; Data $Z = (X_i, \tilde{T}_i, C_i)_{i \in \mathscr{I}}$; Testing point x;
Function $V(x, y; D)$: the conformity score between (x, y) and data D;
Function $\hat{w}(x, D)$ to fit the weight function at x using D;
Function $C(D)$ to select the threshold c_0 using D.

Procedure:

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

Output: $\hat{L}(x) = \inf\{t : V(x, t; Z_{tr}) \leq \eta(x)\} \wedge c_0$

- In Algorithm 1, unknown covariate shift $\hat{w}(x)$ can be estimated using training fold.
- $\hat{w}(x; Z_{tr}) = \infty \implies \hat{p}_\infty = 1 \implies \hat{L}(x) = -\infty$.
- $\eta(x)$ is invariant to positibe rescalings of $\hat{w}(x)$. We can easily set $\hat{w}(x) = 1/\hat{c}(x; c_0)$.
- **Choice of conformity score** $V(x, y; D)$:
  - *Conformalized Mean Regression* (CMR) scores defined via $V(x, y; Z_{tr}) = \hat{m}(x) - y$, where $\hat{m}(x)(.)$ is an estimate of mean of $Y|X$. The resulting LPB is then $\hat{m}(x) - \eta(x)) \wedge c_0$.
  - *Conformalized Quantile Regression* (CQR) scores defined via $V(x, y; Z_{tr}) = \hat{q}_\alpha(x) - y$, where $\hat{q}_\alpha(.)$ is estimate of α-quantile of $Y|X$. The resulting LPB is then $(\hat{q}_\alpha - \eta(x)) \wedge c_0$.
  - *Conformalized Distribution Regression* (CDR) scores defined via $V(x, y; Z_{tr}) = \alpha - \hat{F}_{Y|X=x}(y)$, where $\hat{F}_{Y|X=x}(y)$ is an estimate of distribution of $Y|X$. The resulting LPB is then $\hat{F}^{-1}_{Y|X=x}(\alpha - \eta(x)) \wedge c_0$.

Theorem

Let c_0 be independent of Z_{ca}, with target $T_i \wedge c_0$ in Algorithm 1 and $\hat{w}(w; D) \equiv 1$. Under completely independent censoring, $\hat{L}(X)$ is calibrated.

Doubly Robust Lower Prediction Bounds

- Under conditionally independent censoring assumption, $c(x; c_0)$ needs to be estimated.
- Instead of using procedures like Kernel estimation or bootstrapping, we can adapt the mechanism proposed in Lei and Candès (2021), i.e. *weighted split conformal* inference.
- If we estimate conditional quantile of the survival function and the censoring mechanism from two datasets which have been obtained by splitting the training set, we obtain a doubly robust LPB, i.e. we get guaranteed average coverage if either covariate shift or conditional quantiles are estimated well.
- The double robustness of weighted split conformal inference allows researchers to leverage the knowledge about both the conditional quantile and censoring mechanism with any concern for which is more accurate.

Choice of Threshold

- The threshold c_0 induces an estimation-censoring tradeoff: larger c_0 overcomes the censoring effect, closing gap between target T and the operating outcome $T \wedge c_0$, but reduces sample size for the estimation problem.
- Ideal steps to generate data-driven threshold:
  1. Set a grid of values for c_0.
  2. Randomly sample a holdout set from Z_{tr}
  3. Apply Algorithm 1 on rest of Z_{tr} for each value for c_0 to generate LPBs for each unit in the holdout set.
  4. Select c_0 which maximizes the average LPB's on the holdout set
- If the holdout set is the calibration set itself, resulting LPBs will be approximately calibrated. To be specific,

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

where $\mathscr{C}$ is a candidate set for c_0.

Concluding Remarks

- This method has been found to have better coverage in comparison to well established prediction techniques like cox-proportional hazard model, accelerated failure time model, censored quantile regression etc, most of which yield bands which are not valid.
- Under suitable assumptions, this method can be adapted to handle both end-of-study censoring caused by the trial termination as well as loss-to-follow-up censoring caused by unexpected attrition.
- The method can also be suitably extended to handle inference based on two or more subpopulation as well as in cohort analysis.

Reference

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.

Lei, J., G'Sell, M. G., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. A. (2016). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094 – 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.

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., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Second edition.