KSP: Kolmogorov-Smirnov metric-based Post-Hoc Calibration for Survival Analysis

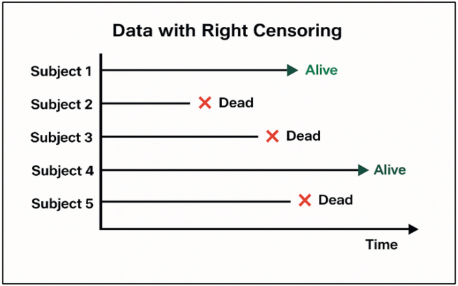
Jeongho Park[1], Daheen Kim[1], Cheoljun Kim[2], Hyungbin Park[2], Sangwook Kang[1], Gwangsu Kim[2,*]

[1]Department of Statistics and Data Science, Yonsei University
[2]Department of Statistics, Jeonbuk National University

39th Conference on Neural Information Processing Systems

Introduction

- Survival analysis aims to estimate the probability that an event (e.g., death) occurs after a given time.
- Utilizing DNN has become an essential part of the survival analysis
- A key challenge is handling censoring and balancing discrimination with calibration.



Notation

- T: event time, C: (right) censoring time
- $Y = \min\{T, C\}$: observed time
- $\delta = \mathbb{I}(T \leq C)$: censoring indicator
- $\boldsymbol{z}$: vector of covariates
- $F(\cdot \mid \boldsymbol{z})$: conditional CDF (cumulative distribution function) of T
- $S(\cdot \mid \boldsymbol{z}) = 1 - F(\cdot \mid \boldsymbol{z})$: survival function

D-calibration (Distributional calibration)

- How can we measure whether our estimated survival function is calibrated or not?
  - Use the property of CDF
  - If $T \sim F$, then $F(T) \sim \text{Unif}[0,1]$.
  - For $T > C$, we get $F(T \mid \boldsymbol{z}) \sim \text{Unif}\left[F(C \mid \boldsymbol{z}), 1\right]$
- $\mathbb{E}_{Y,\delta,\boldsymbol{z}}\left[\mathbb{I}\left(F\left(Y \mid \boldsymbol{z}\right) \leq x\right)\left\{\delta + (1-\delta)\frac{x - F(Y|\boldsymbol{z})}{1 - F(Y|\boldsymbol{z})}\right\}\right] = x, \ \forall x \in [0,1]$
- A key is how to measure the difference between both sides of the equation.

Kolmogorov-Smirnov metric

$$\mathbb{E}_{Y,\delta,\boldsymbol{z}}\left[\mathbb{I}\left(F\left(Y \mid \boldsymbol{z}\right) \leq x\right)\left\{\delta + (1-\delta)\frac{x - F\left(Y \mid \boldsymbol{z}\right)}{1 - F\left(Y \mid \boldsymbol{z}\right)}\right\}\right] = x$$

- Previous approaches used bin-based difference.
- We adopt the Kolmogorov-Smirnov (KS) metric.
- Let

  $$\tilde{F}\left(x\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\hat{F}_{\boldsymbol{\theta}}\left(Y_i \mid \boldsymbol{z}_i\right) \leq x\right)\left\{\delta_i + (1-\delta_i)\frac{x - \hat{F}_{\boldsymbol{\theta}}\left(Y_i \mid \boldsymbol{z}_i\right)}{1 - \hat{F}_{\boldsymbol{\theta}}\left(Y_i \mid \boldsymbol{z}_i\right)}\right\}$$

  - $\hat{F}_{\boldsymbol{\theta}}$: estimated CDF
- KS-cal$= \sup_{x \in [0,1]} \left|\tilde{F}\left(x\right) - x\right|$

Convergence of KS-cal

Theorem

Under the regularity conditions,
$\hat{F}_{\boldsymbol{\theta}} = F$ *if and only if* $\sup_{x \in [0,1]} \left| \tilde{F}(x) - x \right| = o_p(1)$ *as* $N \to \infty$.

- Minimizing KS-cal guarantees the model is to be calibrated.
- We propose KS-cal based Post-hoc calibration (KSP).

KSP

Algorithm. KSP

Input: Estimated CDFs $\hat{F}_{\boldsymbol{\theta}}$, strictly monotone increasing link function $G : [0,1] \to (-\infty, \infty)$
Initialize parameters $a\ (>0)$, b, $\alpha\ (>0)$
Sort $\hat{F}_{\boldsymbol{\theta}}$ for computational efficiency
while KS-cal not improved **do**
    Compute transformed CDF: $\hat{F}^*_{\boldsymbol{\theta}} = \left\{G^{-1}\left(a \cdot G(\hat{F}_{\boldsymbol{\theta}}) + b\right)\right\}^{\alpha}$
    Compute KS-cal on validation set: $\max_{1 \le j \le N} D^*_j$, where D^*_j denotes D_j evaluated using $\hat{F}^*_{\boldsymbol{\theta}}$
    Update (a, b, α) via gradient descent (ADAM) to minimize the KS-cal
end while
Apply final calibrated transformation to the test set using optimized (a, b, α)
Output: Calibrated CDF $\hat{F}^*_{\boldsymbol{\theta}}$

- No
  - surrogate loss
  - additional nonparametric estimator
  - quantile estimation
  - sampling procedure
- Yes
  - intuitive and easy to implement
  - preserve time-dependent C-index

Result

Table 1: Summary of pairwise comparisons between post-processing methods. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
KSP	20 (12)	**46 (45)**	**46 (43)**	**47 (45)**	**47 (35)**	**44 (25)**
Non-calibrated	18 (1)	13 (7)	14 (6)	13 (5)	13 (10)	16 (1)
Ties	**22**	1	0	0	0	0
KSP	13 (2)	**36 (29)**	**48 (45)**	**51 (42)**	**37 (32)**	**42 (25)**
CSD	**34 (2)**	24 (19)	12 (10)	9 (8)	23 (19)	18 (10)
Ties	13	0	0	0	0	0
KSP	21 (0)	**32 (21)**	**46 (39)**	**44 (29)**	**45 (36)**	**38 (9)**
CSD-iPOT	**25 (1)**	28 (19)	14 (13)	16 (11)	15 (10)	22 (10)
Ties	14	0	0	0	0	0

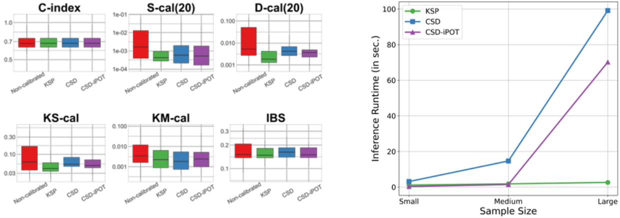


Figure 2: Boxplots of metric values (left) and inference runtime by sample size (right), aggregated across all datasets and models.

Conclusion

- Stength
  - Capture local discrepancies more than quantile-based methods
  - Scalable
- Weakness
  - Sensitive to discretized models
  - Less robust to tied times
- Future research
  - Extend KSP to conditional calibration