X-CAL: Explicit Calibration for Survival Analysis

Mark Goldstein *
Courant Institute
New York University
New York, NY 10011
goldstein@nyu.edu

Xintian Han*
Center for Data Science
New York University
New York, NY 10011
xintian.han@nyu.edu

Aahlad M. Puli*
Courant Institute
New York University
New York, NY 10011
aahlad@nyu.edu

Adler J. Perotte
Biomedical Informatics
Columbia University Medical Center
New York, NY 10032
adler.perotte@columbia.edu

Rajesh Ranganath
Courant Institute
Center for Data Science
New York University
New York, NY 10011
rajeshr@cims.nyu.edu

Survival Analysis

- predict time until event
- time $\mathbf{t}$ drawn conditional on covariates $\mathbf{x}$.
- could be time until admission to hospital based on health records

Censoring

- for some data, we don't observe $\mathbf{t}$
- only observe right-censoring time $\mathbf{c} < \mathbf{t}$
- assume censoring-at-random

$$\mathbf{t} \perp\!\!\!\perp \mathbf{c} \mid \mathbf{x}$$

- e.g. someone leaves a heart study before they develop a heart condition

Calibration

- want accurate survival models, also want **calibration**
- predicted number of events within any time interval similar to observed number
- means probabilities can be interpreted as risk
- can be used for decisions

Intuition: Binary Classification

- covariates $\boldsymbol{x}$ and binary outcome $\boldsymbol{d}$.
- modeled risk $P_\theta(d = 1|x)$ denoted by $\mathtt{risk}_\theta(x)$.
- calibration: $P_{true}(\boldsymbol{d} = 1|\mathtt{risk}_\theta(\boldsymbol{x}) = r) \approx r$
- frequency of events is r among subjects whose modeled risks are r.

Evaluating Calibration

- usually evaluated post-hoc
- instead of checking all $r \in \mathbb{R}$, check some intervals of risk levels
- see tests like Lemeshow-Hosmer
- not differentiable due to checking set membership of data to risk groups

Calibration in Regression / Survival Analysis

- $(\boldsymbol{x}, \boldsymbol{t}) \sim P$ with conditional CDF F
- definition of risk needs to use **t**
- define $\texttt{risk}_\theta(t, x) = F_\theta(t|x)$
- For all sub-intervals $C = [a, b]$ of $[0, 1]$, calibration means

$$\mathbb{E}_{t,x\sim P}\Big[\mathbb{1}\left[\texttt{risk}_\theta(t, x) \in C\right]\Big] = \mathbb{E}_{t,x\sim P}\Big[\mathbb{1}\left[F_\theta(t|x) \in C\right]\Big] = |C|.$$

- Holds when $F_\theta(t|x) = F(t|x)$ or when $F_\theta(t|x) = F(t)$

D-Calibration: a way to measure

- Count proportion of points in set C

$$\phi_\theta(C) := \mathbb{E}_{t,x\sim P(\boldsymbol{t},\boldsymbol{x})} \mathbb{1}\left[F_\theta(t|\boldsymbol{x}=x) \in C\right]$$

- Pick disjoint sets $C \in \mathcal{C}$ that cover $[0,1]$ and measure:

$$\mathcal{R}(\theta) = \sum_{C\in\mathcal{C}} \left(\phi_\theta(C) - |C|\right)^2$$

Obtaining D-Calibration

- like binary classification, uses set membership
- also has some difficult expectations
- could we minimize this error in training?

Obtaining D-Calibration

- like binary classification, uses set membership
- also has some difficult expectations
- could we minimize this error in training?
- yes

X-Calibration

- use approximation of D-calibration as an objective alongside MLE
- improves calibration without sacrificing much likelihood/concordance,
- allows modeler to balance
- does this during training!

X-Calibration

- two main challenges
- relax indicator function with soft membership
- upper-bound square of expectation to derive stochastic estimator

Soft Membership

- Soft membership with temperature γ for set $C = [a, b]$:

$$\zeta(u; C, \gamma) = \text{Sigmoid}(\gamma(u-a)(b-u))$$

- Inexact for finite γ but has gradients
- Exact when $\gamma \to \infty$ but can't optimize

Soft Membership

Approximately check CDF value in C

$$\mathcal{R}(\theta) = \sum_{C \in \mathcal{C}} \Big(\phi_\theta(C) - |C| \Big)^2$$

$$\mathcal{R}(\theta) = \sum_{C \in \mathcal{C}} \Big(\mathbb{E}_{t,x} \mathbb{1} \left[F_\theta(t | \boldsymbol{x} = x) \in C \right] - |C| \Big)^2$$

with

$$\hat{\mathcal{R}}_\gamma(\theta) = \sum_{C \in \mathcal{C}} \left(\mathbb{E}_{t,x} \zeta(F_\theta(t | \boldsymbol{x} = x); C, \gamma) - |C| \right)^2 .$$

Bad Square

- gradient is product of two expectations (bad)
- move square into each term of expectation

$$\hat{\mathcal{R}}_\gamma(\theta) = \sum_{C\in\mathcal{C}} \left(\mathbb{E}_{t,x}\zeta(F_\theta(t|\boldsymbol{x}=x); C,\gamma) - |C|\right)^2 .$$

less than

$$\hat{\mathcal{R}}^+_\gamma(\theta) = \mathbb{E}_{S\sim P(\boldsymbol{t},\boldsymbol{x})^M} \sum_{C\in\mathcal{C}} \frac{1}{M^2} \sum_{t,x\in S} \left[\zeta(F_\theta(t|\boldsymbol{x}=x); C,\gamma) - |C|\right]^2$$

for batch S of size M by Jensen's

Example Objective: MLE + XCAL

$$\min_{\theta} \quad \mathbb{E}_{t,x\sim P} - \log P_{\theta}(t|\boldsymbol{x} = x) + \lambda \hat{\mathcal{R}}_{\gamma}^{+}(\theta)$$

Gamma Simulation

For the gamma simulation, we draw $\mathbf{x}$ from a $D = 32$ multivariate Normal with $\mathbf{0}$ mean and diagonal covariance with $\sigma^2 = 10.0$. We draw failure times $\mathbf{t}$ conditionally on $\mathbf{x}$ from a gamma distribution with mean $\boldsymbol{\mu}$ log-linear in $\mathbf{x}$. The weights of the linear function are drawn uniformly. The gamma distribution has constant variance 1e-3. This is achieved by setting $\alpha = \boldsymbol{\mu}_i^2/1e\text{-}3$ and $\beta = \boldsymbol{\mu}_i/1e\text{-}3$.

$$\mathbf{x}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad \mathbf{w}_d \sim \texttt{Unif}(-0.1, 0.1), \quad \boldsymbol{\mu}_i = \exp[\mathbf{w}^\top \mathbf{x}_i], \quad \mathbf{t}_i \sim \texttt{Gamma}(\alpha, \beta).$$

Censoring times are drawn like failure times but with a different set of weights for the linear function. This means $\mathbf{t} \perp\!\!\!\perp \mathbf{c} \mid \mathbf{x}$.

Results

Table 1: Gamma simulation, censored

	λ	0.0	1.0	10.0	100.0	500.0	1000.0
Log-Norm NLL	Test NLL	-0.059	-0.049	0.004	0.138	0.191	0.215
	Test D-CAL	0.0292	0.0195	0.0045	0.0002	6e-5	7e-5
	Test Conc.	0.981	0.969	0.942	0.916	0.914	0.897
Log-Norm S-CRPS	Test NLL	0.038	0.084	0.143	0.201	0.343	0.436
	Test D-CAL	0.0174	0.0071	0.0014	0.0001	5e-5	8e-5
	Test Conc.	0.982	0.978	0.963	0.950	0.850	0.855
Cat-NI	Test NLL	0.797	0.799	0.822	1.149	1.665	1.920
	Test D-CAL	0.0091	0.0064	0.0015	0.0002	6e-5	6e-5
	Test Conc.	0.987	0.987	0.987	0.976	0.922	0.861