Deep Survival Analysis

Chapter 1: An Introduction to Conditional Survival Analysis

Yao Zhang

Intelligent Information Processing Research Group,

Faculty of Electrical Engineering and Computer Science,

Ningbo University

The audience is presumed to have a basic understanding of calculus, linear algebra, probability, statistics, and machine learning.

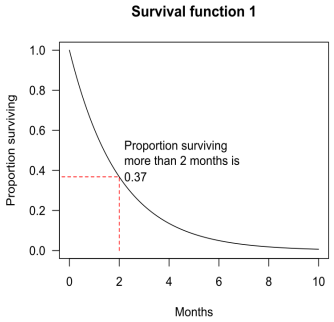
We may not always find an answer, and since we're not very familiar with (deep) survival analysis, we will need to dedicate more time to this topic.

Jan 6, 2025

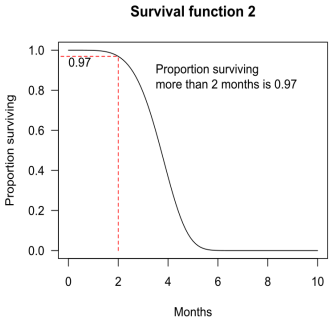
Overview

Survival Analysis

Survival analysis, or more generally, **time-to-event analysis**, refers to a set of methods for analyzing the length of time until the occurrence of a well-defined end point of interest.



(a)



(b)

Figure 1: Survival Function[1]

[1] Wikipedia. *Survival Function*. URL: https://en.wikipedia.org/wiki/Survival_function.

Statistical Framework

Problem Setting

Assume there are n training points, denoted as $(X_1, Y_1, \Delta_1), \ldots, (X_n, Y_n, \Delta_n)$, where each training point $i \in [1, n]$ is represented by:

- $X_i \in \mathcal{X}$ is the raw input (e.g., a fixed-length feature vector, an image, a text document, etc.),
- $Y_i \in [0, \infty)$ is the observed time (either the true survival time or the censoring time),
- $\Delta_i \in \{0, 1\}$ is the event indicator:
  - If $\Delta_i = 1$, the event occurred, and Y_i is the true survival time.
  - If $\Delta_i = 0$, the event did not occur, and Y_i is the censoring time (the last observed time).

The goal is to model the relationship between the raw input X and the survival or censoring time.

Statistical Framework

The notation for the observed data can be simplified as follows:

$$Y_i = \min(T_i, C_i), \tag{1}$$

and the event indicator is:

$$\Delta_i = 1\{T_i \leq C_i\}, \tag{2}$$

where $1\{\cdot\}$ is the indicator function, equal to 1 if its argument is true and 0 otherwise.

Thus, for any generic raw input X with true survival time T and true censoring time C, we have the following observations:

$$Y = \min(T, C), \tag{3}$$

and the event indicator:

$$\Delta = 1\{T \leq C\}. \tag{4}$$

The statistical framework described above is known as right-censored data, where for censored data (when $\Delta_i = 0$), the true survival time T_i is greater than the observed censoring time C_i. **This framework assumes independent censoring, meaning that the survival and censoring times are conditionally independent given the raw input X.**

Survival Analysis in Continuous Time

Key Assumptions A

For a test raw input $x \in \mathcal{X} \subseteq \mathbb{R}^d$, we assume that the survival time T conditioned on $X = x$ is a continuous random variable with probability density function (PDF) $f(t|x)$ and cumulative distribution function (CDF) $F(t|x) = \int_0^t f(u|x)\, du$; either of these functions fully characterizes the distribution $P_{T|X}(\cdot|x)$.

Survival Function A

The conditional survival function is given by

$$\begin{aligned} S(t \mid x) :=& P(\text{survive beyond time } t \mid \text{raw input is } x) \\ =& P(T > t \mid X = x) \\ =& 1 - P(T \leq t \mid X = x) \\ =& 1 - F(t|x), \end{aligned} \tag{5}$$

where $t \geq 0$ and $x \in \mathcal{X}$.

In this talk, we refer to the conditional survival function $S(\cdot|x)$ simply as the survival function since our notation already indicates that we are conditioning on x.

Predicting $S(\cdot|x)$ means estimating an entire function (i.e., a curve)—not just a single number (survival time)—for test raw input x.

Properties of the Survival Function A

1. $S(\cdot|x) = 1 - F(\cdot|x)$ monotonically decreases from 1 to 0 since any CDF monotonically increases from 0 to 1.

2. Estimating the function $S(\cdot|x)$ is equivalent to estimating the CDF $F(\cdot|x)$, which means that we aim to estimate the conditional survival time distribution $P_{T|X}(\cdot|x)$.

Different time-to-event prediction models make different assumptions on $S(\cdot|x)$ and often predict transformed variants of $S(\cdot|x)$ rather than predicting $S(\cdot|x)$ directly.

Hazard Function A

The hazard function is given as below:

$$\begin{aligned} h(t \mid x) &:= -\frac{d}{dt} \log S(t|x) \\ &= -\frac{\frac{d}{dt} S(t|x)}{S(t|x)} = -\frac{\frac{d}{dt}[1 - F(t|x)]}{S(t|x)} = \frac{f(t|x)}{S(t|x)}, \end{aligned} \tag{6}$$

where, $f(\cdot|x)$ is the PDF of distribution $P_{T|X}(\cdot|x)$. The hazard function is only nonnegative and could have arbitrarily large positive values.

If $h(\cdot|x)$ is known, $S(\cdot|x)$ can be recovered as follows:

$$\begin{aligned} h(t|x) = -\frac{d}{dt} \log S(t|x) \quad &\Leftrightarrow \quad \int_0^t h(u|x)\, du = -\log S(t|x) \\ &\Leftrightarrow \quad S(t|x) = \exp\left(-\int_0^t h(u|x)\, du\right). \end{aligned} \tag{7}$$

Cumulative Hazard Function A

The cumulative hazard function is defined as:

$$H(t \mid x) := \int_0^t h(u|x)\, du. \tag{8}$$

From equation (7), we observe that $S(t|x) = \exp(-H(t|x))$. Therefore, if we know $H(t|x)$, we can recover $S(t|x)$. Furthermore, from equation (8), we see that $h(t|x) = \frac{d}{dt}H(t|x)$. Hence, if we know $H(t|x)$, we can also recover $h(t|x)$.

It is important to note that while $S(t|x)$ and $H(t|x)$ are monotonic functions, $h(t|x)$ is not necessarily monotonic.

Key Relationships A

The following equations show how the PDF $f(\cdot|x)$, the CDF $F(\cdot|x)$, the survival function $S(\cdot|x)$, the hazard function $h(\cdot|x)$, and the cumulative hazard function $H(\cdot|x)$ are related:

$$
\begin{aligned}
&\text{KRA1}: \quad f(t|x) = \frac{d}{dt}F(t|x) = \frac{d}{dt}(1 - S(t|x)) = h(t|x)S(t|x), \\
&\text{KRA2}: \quad F(t|x) = \int_0^t f(u|x)\,du = 1 - S(t|x), \\
&\text{KRA3}: \quad S(t|x) = 1 - F(t|x) = \int_t^\infty f(u|x)\,du = e^{-H(t|x)} = e^{-\int_0^t h(u|x)\,du}, \\
&\text{KRA4}: \quad h(t|x) = \frac{d}{dt}H(t|x) = -\frac{d}{dt}\log S(t|x) = \frac{f(t|x)}{S(t|x)}, \\
&\text{KRA5}: \quad H(t|x) = -\log S(t|x) = \int_0^t h(u|x)\,du.
\end{aligned}
\tag{9}
$$

Likelihood A Construction Under Censoring

Given that $g(Y_i;\phi)$ represents the density and $G(Y_i;\phi)$ is the survivor function of the censoring process, and assuming that $T_i \perp C_i$, we can derive the likelihood as follows:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} P(T_i \in [Y_i, Y_i + \Delta t_1), C_i > Y_i)^{\Delta_i} \cdot P(T_i > Y_i, C_i \in [Y_i, Y_i + \Delta t_2))^{1-\Delta_i} \\
&= \prod_{i=1}^{n} [f(Y_i;\theta)\,\Delta t_1\, G(Y_i;\phi)]^{\Delta_i} [S(Y_i;\theta)\, g(Y_i;\phi)\,\Delta t_2]^{1-\Delta_i} \qquad (10) \\
&= \prod_{i=1}^{n} [f(Y_i;\theta)]^{\Delta_i} [S(Y_i;\theta)]^{1-\Delta_i} [\Delta t_1\, G(Y_i;\phi)]^{\Delta_i} [g(Y_i;\phi)\,\Delta t_2]^{1-\Delta_i}.
\end{aligned}
$$

Finally, the likelihood function can be expressed as:

$$
L \propto \prod_{i=1}^{n} [f(Y_i;\theta)]^{\Delta_i} [S(Y_i;\theta)]^{1-\Delta_i}. \qquad (11)
$$

Thus, we obtain the flexible form of the likelihood function:

$$
L := \prod_{i=1}^{n} \left[f(Y_i|X_i)^{\Delta_i} S(Y_i|X_i)^{1-\Delta_i} \right]. \qquad (12)
$$

Likelihood A, Cont.

Using the given relations from Key Relationships (9), one can rewrite equation (12) as

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \left[f(Y_i|X_i)^{\Delta_i} S(Y_i|X_i)^{1-\Delta_i} \right] \\
&\overset{\text{KRA1}}{=} \prod_{i=1}^{n} \left[h(Y_i|X_i)^{\Delta_i} S(Y_i|X_i) \right] \\
&\overset{\text{KRA3}}{=} \prod_{i=1}^{n} \left[h(Y_i|X_i)^{\Delta_i} \exp\left(- \int_0^{Y_i} h(u|X_i)\, du \right) \right].
\end{aligned}
\tag{13}
$$

Likelihood A, Cont.

By substituting $h(t|x) = h(t|x;\theta)$, we obtain:

$$L(\theta) = \prod_{i=1}^{n} \left[h(Y_i|X_i;\theta)^{\Delta_i} \exp\left(- \int_0^{Y_i} h(u|X_i;\theta)\, du \right) \right]. \tag{14}$$

Taking the logarithm of the likelihood function, we get the log-likelihood:

$$\begin{aligned} \log L(\theta) &= \log\left(\prod_{i=1}^{n} \left[h(Y_i|X_i;\theta)^{\Delta_i} \exp\left(- \int_0^{Y_i} h(u|X_i;\theta)\, du \right) \right] \right) \\ &= \sum_{i=1}^{n} \left[\Delta_i \log h(Y_i|X_i;\theta) - \int_0^{Y_i} h(u|X_i;\theta)\, du \right]. \end{aligned} \tag{15}$$

The negative log-likelihood (NLL), which we often use as the loss function, is defined as:

$$\begin{aligned} L_{\text{Hazard - NLL}}(\theta) &:= -\frac{1}{n} \log L(\theta) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \log h(Y_i|X_i;\theta) - \int_0^{Y_i} h(u|X_i;\theta)\, du \right]. \end{aligned} \tag{16}$$

Proportional Hazard Models (PHMs)

Proportional hazards models (PHMs) assume that the hazard function can be factored as follows:

$$h(t|x) = h_0(t;\theta)e^{f(x;\theta)} \quad \text{for } t \geq 0, x \in X, \tag{17}$$

where $h_0(\cdot;\theta) : [0,\infty) \to [0,\infty)$ and $f(\cdot;\theta) : X \to \mathbb{R}$ are functions with parameter vector θ. Specifically:

1. Exponential PHM: $h(t|x;\theta) := e^{\beta^\top x+\psi}$ for $t \geq 0$, $x \in X$. In this case, we have
$$h_0(t;\theta) = e^{\psi} \quad \text{and} \quad f(x;\theta) = \beta^\top x, \tag{18}$$
where $\theta = (\beta,\psi) \in \mathbb{R}^d \times \mathbb{R}$.
2. Weibull PHM: $h(t|x;\theta) := te^{\phi} - e^{(\beta^\top x)e^{\phi}+\psi+\phi}$ for $t \geq 0$, $x \in X$. In this case, we have
$$h_0(t;\theta) = t^{e^{\phi}-1}e^{\psi+\phi} \quad \text{and} \quad f(x;\theta) = (\beta^\top x)e^{\phi}, \tag{19}$$
where $\theta = (\beta,\psi,\phi) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$.

Exponential PHM

Substituting $h(t|x;\theta) = e^{\beta^\top x+\psi}$ into equation (16) results in:

$$\begin{aligned} L_{\text{Hazard - NLL}}(\beta,\psi) &= -\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i(\beta^\top X_i+\psi) - \int_0^{Y_i} e^{\beta^\top X_i+\psi}\,du\right] \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i(\beta^\top X_i+\psi) - Y_i e^{\beta^\top X_i+\psi}\right]. \end{aligned} \tag{20}$$

Substituting $h(t|x;\theta) := te^{\phi} - e^{(\beta^\top x)e^{\phi}+\psi+\phi}$ into equation (16) yields the following expression:

$$
\begin{aligned}
&L_{\text{Hazard - NLL}}(\beta,\psi,\varphi) \\
&\quad= -\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i \log\left(Y_i e^{\varphi-1} e^{(\beta^\top X_i)e^{\varphi}+\psi+\varphi}\right) - Y_i e^{\varphi} e^{(\beta^\top X_i)e^{\varphi}+\psi}\right\} \\
&\quad= -\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i\left[(e^{\varphi}-1)\log Y_i + (\beta^\top X_i)e^{\varphi} + \psi + \varphi\right] - Y_i e^{\varphi} e^{(\beta^\top X_i)e^{\varphi}+\psi}\right\}.
\end{aligned}
\tag{21}
$$

Cox PHM

In his discussion of Cox's (1972) paper on proportional hazards regression[2], Breslow (1972) provided the maximum likelihood estimator for the cumulative baseline hazard function[3]. The article by Lin (2007)[4] outlines the historical context of the Cox model.

The hazard function of the Cox PHM is defined as:

$$h(t|x;\theta) := h_0(t;\theta)e^{f(x;\theta)}, \tag{22}$$

where $h_0(t;\theta)$ is a piecewise constant function, given by:

$$h_0(t;\theta) := \begin{cases} \lambda_l \text{ if } \tau_{(l-1)} < t \leqslant \tau_{(l)}, & l \in [L], \\ 0 \qquad\quad \text{if } t > \tau_{(L)}, \end{cases} \tag{23}$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_L) \in [0, \infty)$, and the times $\tau_{(1)}, \tau_{(2)}, \ldots, \tau_{(L)}$ are the unique times of death. Additionally, $\tau_{(0)} := 0$.

[2] D. Cox. "Regression Models and Life-Tables(with Discussion)". In: *Journal of the Royal Statistical Society. Series B* 34(2) (1972), pp. 187–220.

[3] N. Breslow. "Discussion of the Paper by D. R. Cox". In: *Journal of the Royal Statistical Society (B)* 34 (1972), pp. 216–217.

[4] D. Lin. "On the Breslow Estimator". In: *Lifetime Data Analysis* 13 (2007), pp. 471–480.

Cox PHM, Cont.

By substituting equation (22) into equation (16), the following expression is obtained (for more details, refer to the Breslow estimator derivation video[5]):

$$\begin{aligned}\log L(\theta,\lambda) &= \sum_{i=1}^{n}\left[\Delta_i \log h(Y_i|X_i;\theta) - \int_0^{Y_i} h(u|X_i;\theta)\,du\right] \\ &= \sum_{i=1}^{n}\left[\Delta_i \log\left(h_0(Y_i;\theta)e^{f(X_i;\theta)}\right) - \int_0^{Y_i} h_0(u;\theta)e^{f(X_i;\theta)}\,du\right] \\ &= \sum_{m=1}^{L} D[m]\log\lambda_{(m)} + \sum_{i=1}^{n}\Delta_i f(X_i;\theta) - \sum_{m=1}^{L}(\tau_{(m)}-\tau_{(m-1)})\lambda_m \sum_{j=1}^{n}\mathbb{I}\left\{Y_j \geq m\right\} e^{f(X_j;\theta)}.\end{aligned} \tag{24}$$

Next, setting $\left.\frac{d\log L(\theta)}{d\lambda(l)}\right|_{\lambda(l)=\widetilde{\lambda}(l)} = 0$, results in the following expression:

$$\widetilde{\lambda}_{(l)} = \frac{D[l]}{(\tau_l - \tau_{l-1})\sum_{j=1}^{n}\mathbb{I}\left\{Y_j \geq \ell\right\} e^{f(X_j;\theta)}}. \tag{25}$$

[5] D. Refaeli. *Survival Analysis – Cox PH – Breslow Estimator.*

Cox PHM, Cont.

We substitute equation (25) into (24), resulting in:

$$\begin{aligned}\log L(\theta) &= \sum_{m=1}^{L} D[m] \log \frac{D[l]}{(\tau_{(m)} - \tau_{(m-1)}) \sum_{j=1}^{n} \mathbb{I}\left\{Y_j \geqslant m\right\} e^{f(X_j;\theta)}} + \sum_{i=1}^{n} \Delta_i f(X_i;\theta) - \sum_{m=1}^{L} D[m] \\ &= \sum_{i=1}^{n} \Delta_i \left[f(X_i;\theta) - \sum_{i=1}^{n} \log \left(\sum_{j=1}^{n} \mathbb{I}\left\{Y_j \geqslant Y_i\right\} e^{f(X_j;\theta)} \right) \right] + \text{constant}.\end{aligned} \tag{26}$$

Finally, we obtain:

$$\mathrm{L}_{\text{Hazard - NLL}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \Delta_i \left[f(X_i;\theta) - \sum_{i=1}^{n} \log \left(\sum_{j=1}^{n} \mathbb{I}\left\{Y_j \geqslant Y_i\right\} e^{f(X_j;\theta)} \right) \right]. \tag{27}$$

Survival Analysis in Discrete Time

Key Assumptions B

Suppose time is discretized into a user-defined grid of L time points $\tau_{(1)}, \tau_{(2)}, \ldots, \tau_{(L)} \in [0, \infty)$, such that $\tau_{(1)} < \tau_{(2)} < \cdots < \tau_{(L)}$. Assume that all training values Y_i have been discretized to take values from $\tau_{(1)}, \tau_{(2)}, \ldots, \tau_{(L)}$

The CDF and PMF of the distribution $P(T \mid X = x)$ are defined as follows:

- The PMF: $f[l \mid x] := P(T = \tau_{(l)} \mid X = x)$, for $l \in [L]$,
- The CDF: $F[l \mid x] := P(T \leq \tau_{(l)} \mid X = x) = \sum_{m=1}^{l} f[m \mid x]$.

The PMF $f[\cdot \mid x]$ satisfies:

- $f[l \mid x] \geq 0$ for all $l \in [L]$,
- $\sum_{l=1}^{L} f[l \mid x] = 1$.

The relationship between the CDF and PMF is: $f[l \mid x] = F[l \mid x] - F[l-1 \mid x]$, with $F[0 \mid x] := 0$.

Survival Function B

For $x \in X$, the discrete-time survival function at time index $l \in [L]$ is defined as:

$$S[l \mid x] := P(T > \tau_{(\ell)} \mid X = x) = 1 - F[l \mid x] = 1 - \sum_{m=1}^{l} f[m \mid x]. \tag{28}$$

Additionally, the PMF $f[l \mid x]$ can be expressed via equation (28) as:

$$\begin{aligned} f[l \mid x] &= F[l \mid x] - F[l-1 \mid x] \\ &= (1 - S[l \mid x]) - (1 - S[l-1 \mid x]) \\ &= S[l-1 \mid x] - S[l \mid x] \quad \text{for } l \in [L], \end{aligned} \tag{29}$$

where $S[0 \mid x] := 1$.

Hazard Function B

The discrete-time hazard function $h[l \mid x]$ is defined as:

$$\begin{aligned} h[l \mid x] :=& P(T = \tau_{(l)} \mid X = x, T \geqslant \tau_{(l)}) \\ =& \frac{P(T = \tau_{(l)}, T \geqslant \tau_{(l)} \mid X = x)}{P(T > \tau_{(l-1)} \mid X = x)} \\ =& \frac{P(T = \tau_{(l)} \mid X = x)}{P(T \geqslant \tau_{(l)} \mid X = x)} = \frac{P(T = \tau_{(l)} \mid X = x)}{P(T > \tau_{(l-1)} \mid X = x)} \\ =& \frac{f[l \mid x]}{S[l-1 \mid x]} = \frac{S[l \mid x] - S[l-1 \mid x]}{S[l-1 \mid x]} \end{aligned} \tag{30}$$

It is important to note that while the continuous-time version $h(t \mid x)$ can be nonnegative and may exceed 1, in discrete time, $h[l \mid x]$ is a probability and thus cannot exceed 1.

Hazard Function B, Cont.

It is evident from equation (30) that:

$$h[l \mid x] = \frac{S[l \mid x] - S[l-1 \mid x]}{S[l-1 \mid x]} \Leftrightarrow S[l \mid x] = S[l-1 \mid x]\,(1 - h[l \mid x])\,. \tag{31}$$

In general, the survival function $S[l \mid x]$ can be expressed as:

$$S[l \mid x] = \prod_{m=1}^{l} (1 - h[m \mid x])\,, \quad l \in [L], \tag{32}$$

Equation (32) illustrates how the hazard function $h[\cdot \mid x]$ can be used to estimate the survival function $S[\cdot \mid x]$.

Cumulative Hazard Function B

The discrete-time cumulative hazard function is defined as:

$$H[l \mid x] := \sum_{m=1}^{l} h[m \mid x], \tag{33}$$

The following relation holds based on equation (20):

$$h[l \mid x] = H[l \mid x] - H[l-1 \mid x], \tag{34}$$

where $H[0 \mid x] := 0$.

It is important to note that in continuous time, the relationship $-\log S[l \mid x] = H[l \mid x]$ holds. However, in discrete time, the corresponding expression is:

$$-\log S[l \mid x] = H[l \mid x] + \sum_{m=1}^{l} \sum_{p=2}^{\infty} \frac{(h[m \mid x])^p}{p}, \quad l \in [L]. \tag{35}$$

Key Relationships B

The following equations show how the PMF $f(\cdot|x)$, the CDF $F(\cdot|x)$, the survival function $S(\cdot|x)$, the hazard function $h(\cdot|x)$, and the cumulative hazard function $H(\cdot|x)$ are related:

$$
\begin{aligned}
&\text{KRB1}: \quad f[l \mid x] = F[l \mid x] - F[l-1 \mid x] = S[l-1 \mid x] - S[l \mid x] = h[l \mid x] S[l-1 \mid x] \\
&\text{KRB2}: \quad F[l \mid x] = \sum_{m=1}^{l} f[m \mid x] = 1 - S[l \mid x] \\
&\text{KRB3}: \quad S[l \mid x] = 1 - F[l \mid x] = \sum_{m=l+1}^{L} f[m \mid x] = \prod_{m=1}^{l} (1 - h[m \mid x]) \\
&\text{KRB4}: \quad h[l \mid x] = H[l \mid x] - H[l-1 \mid x] = \frac{S[l-1 \mid x] - S[\ell \mid x]}{S[l-1 \mid x]} = \frac{f[l \mid x]}{S[l-1 \mid x]} \\
&\text{KRB5}: \quad H[l \mid x] = \sum_{m=1}^{\ell} \frac{S[m-1 \mid x] - S[m \mid x]}{S[m-1 \mid x]} = \sum_{m=1}^{l} h[m \mid x]
\end{aligned}
\tag{36}
$$

where $F[0 \mid x] = 0, \quad S[0 \mid x] = 1, \quad H[0 \mid x] = 0.$

Likelihood B

The likelihood function is expressed as:

$$L := \prod_{i=1}^{n} [f[\kappa(Y_i) \mid X_i]]^{\Delta_i} [S[\kappa(Y_i) \mid X_i]]^{1-\Delta_i}, \tag{37}$$

where $\kappa(Y_i)$ represents the specific time index corresponding to the observed time Y_i, with Y_i discretized to one of the values in $\tau_{(1)}, \tau_{(2)}, \ldots, \tau_{(L)}$. The function $f[\kappa(Y_i) \mid X_i]$ denotes the probability mass of the event occurring at the time $\kappa(Y_i)$, conditional on the covariates X_i, while $S[\kappa(Y_i) \mid X_i]$ represents the survival function at the corresponding time index, conditional on X_i. The indicator Δ_i takes the value 1 if the event is occurred ($\Delta_i = 1$) or 0 if the event is censored ($\Delta_i = 0$) for subject i.

Likelihood B, Cont.

Using the given relations from Key Relationships (36), one can rewrite equation (37) as:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} [f[\kappa(Y_i) \mid X_i]]^{\Delta_i} [S[\kappa(Y_i) \mid X_i]]^{1-\Delta_i} \\
&\overset{\text{KRB1}}{=} \prod_{i=1}^{n} \left[(h[\kappa(Y_i) \mid X_i] S[\kappa(Y_i) - 1 \mid X_i])^{\Delta_i} S[\kappa(Y_i) \mid X_i]^{1-\Delta_i} \right] \\
&\overset{\text{KRB3}}{=} \prod_{i=1}^{n} \left[h[\kappa(Y_i) \mid X_i] \left(\prod_{m=1}^{\kappa(Y_i)-1} (1 - h[m \mid X_i]) \right)^{\Delta_i} \left(\prod_{m=1}^{\kappa(Y_i)} (1 - h[m \mid X_i]) \right)^{1-\Delta_i} \right] \\
&= \prod_{i=1}^{n} \left[h[\kappa(Y_i) \mid X_i]^{\Delta_i} (1 - h[\kappa(Y_i) \mid X_i])^{1-\Delta_i} \left(\prod_{m=1}^{\kappa(Y_i)} (1 - h[m \mid X_i]) \right) \right]
\end{aligned}
\tag{38}
$$

Likelihood B, Cont.

Taking the logarithm of both sides of equation (38), the log-likelihood function is given by:

$$\log L = \sum_{i=1}^{n} [\Delta_i \log(h[\kappa(Y_i) \mid X_i]) + (1 - \Delta_i) \log(1 - h[\kappa(Y_i) \mid X_i])] + \sum_{m=1}^{\kappa(Y_i)-1} \log(1 - h[m \mid X_i]) \tag{39}$$

To maximize equation (39), which can be equivalently stated as minimizing the negative log-likelihood, the expression becomes:

$$L_{\text{Hazard-NLL}}(\xi) = -\frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \log(h[\kappa(Y_i)|X_i; \xi]) + (1 - \Delta_i) \log(1 - h[\kappa(Y_i)|X_i; \xi]) + \sum_{m=1}^{\kappa(Y_i)-1} \log(1 - h[m|X_i; \xi]) \right]. \tag{40}$$

Reduced Version of DeepHit

Note: Below is a **Reduced Version** of DeepHit (RVD) model[6]. The RVD model defines the PMF $f[\cdot \mid x]$ using a neural network $f(\cdot;\xi) : X \to [0,1]^L$ with parameter ξ, such that:

$$\begin{bmatrix} f[1 \mid x] \\ f[2 \mid x] \\ \vdots \\ f[L \mid x] \end{bmatrix} = \begin{bmatrix} f_1(x;\xi) \\ f_2(x;\xi) \\ \vdots \\ f_L(x;\xi) \end{bmatrix} =: f(x;\xi) \tag{41}$$

We use the transformed form of (37), that is,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [f[\kappa(Y_i) \mid X_i]]^{\Delta_i} S[\kappa(Y_i) \mid X_i]^{1-\Delta_i} \\ &\overset{\text{KR23}}{=} \prod_{i=1}^{n} [f[\kappa(Y_i) \mid X_i]]^{\Delta_i} \sum_{m=\kappa(Y_i)+1}^{L} f[m \mid X_i]^{1-\Delta_i} \\ &= \prod_{i=1}^{n} \left[f_{\kappa(Y_i)}(X_i;\xi)\right]^{\Delta_i} \sum_{m=\kappa(Y_i)+1}^{L} f_m(X_i;\xi)^{1-\Delta_i} \end{aligned} \tag{42}$$

[6] C. Lee et. al. "DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks". In: *Thirty-Second AAAI Conference on Artificial Intelligence* 32(1) (2018), pp. 2314–2321.

In practice, to maximize $L(\xi)$, a user-specified neural network optimizer can be used to minimize the negative log-likelihood averaged over the training data:

$$\begin{aligned} L_{\text{PMF-NLL}}(\xi) &:= -\frac{1}{n}\log L(\xi) \\ &= -\frac{1}{n}\log \prod_{i=1}^{n}\left[f_{\kappa(Y_i)}(X_i;\xi)\right]^{\Delta_i}\sum_{m=\kappa(Y_i)+1}^{L} f_m(X_i;\xi)^{1-\Delta_i} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i \log\left(f_{\kappa(Y_i)}(X_i;\xi)\right) + (1-\Delta_i)\log\left(\sum_{m=\kappa(Y_i)+1}^{L} f_m(X_i;\xi)\right)\right] \end{aligned} \tag{43}$$

Note: Below is a **Revised Version** of Nnet-survival (RVNS) model[7].

The hazard function $h[\cdot|x]$ in RVNS model is specified by $g(\cdot;\xi) : X \to \mathbb{R}^L$ with parameter variable ξ. In particular, $g(x;\xi)$ is defined as:

$$g(x;\xi) := \begin{bmatrix} g_1(x;\xi) \\ g_2(x;\xi) \\ \vdots \\ g_L(x;\xi) \end{bmatrix}. \tag{44}$$

Then, the RVNS model defines $h[\ell|x]$ as:

$$h[\ell|x;\xi] := \frac{1}{1+\exp(-g_\ell(x;\xi))}, \quad \ell \in [L], x \in X, \tag{45}$$

where equation (45) ensures that $h[\ell|x;\xi] \in [0,1]$ for every $\ell \in [L]$.

[7] M. Gensheimer et. al. "A Scalable Discrete-Time Survival Model for Neural Networks". In: *PeerJ* 7 (2019), e6257.

By substituting equation (45) into equation (40), the result is:

$$
\begin{aligned}
L_{\text{RVNS - NLL}}(\xi) &= -\frac{1}{n}\left[\Delta_i \log\left(\frac{1}{1+\exp(-g_{\kappa(Y_i)}(X_i;\xi))}\right) + (1-\Delta_i)\log\left(\frac{1}{1+\exp(g_{\kappa(Y_i)}(X_i;\xi))}\right) + \sum_{m=1}^{\kappa(Y_i)-1}\log\left(\frac{1}{1+\exp(g_m(X_i;\xi))}\right)\right] \\
&= \frac{1}{n}\left[\Delta_i \log\left(1+\exp(-g_{\kappa(Y_i)}(X_i;\xi))\right) + (1-\Delta_i)\log\left(1+\exp(g_{\kappa(Y_i)}(X_i;\xi))\right) + \sum_{m=1}^{\kappa(Y_i)-1}\log\left(1+\exp(g_m(X_i;\xi))\right)\right].
\end{aligned}
\tag{46}
$$

Summary

In this talk, the following equations are particularly useful:

1. Key Relationships A (9) and B (36)
2. The original likelihood functions (12) and (37)
3. The negative log-likelihoods (16) and (40)

References

The references not listed in the previous slides are as follows:

8. Q. Wang. *Survival Data Analysis (in Chinese)*. China Science Publishing & Media Ltd., 2006.
9. Y. Li and C. He. *Applied Survival Models: Analysis of Incomplete Data (in Chinese)*. South China Unviersity of Technology Press, 2015.
10. H. Kvamme et. al. “Time-to-Event Prediction with Neural Networks and Cox Regression”. In: *Journal of Machine Learning Research* 20 (2019), pp. 1–30.
11. D. Spicker. *Statistical Methods for Life History Analysis*. University of Waterloo Faculty of Mathematics, 2022. URL: https://dylanspicker.com/courses/STAT437/course_index.html/.
12. S. Wiegrebe et. al. “Deep Learning for Survival Analysis: A Review”. In: *Artificial Intelligence Review* 57 (2024), pp. 1–65.
13. G. Chen. “An Introduction to Deep Survival Analysis Models for Predicting Time-to-Event Outcomes”. In: *Foundations and Trends® in Machine Learning* 17(6) (2024), pp. 921–1100.

Discussion

Any comments or questions?

We may not always find an answer, and since we're not very familiar with (deep) survival analysis, we will need to dedicate more time to this topic.