# Ghosh Lin Estimation

## 1.1 Setup

Let $D$ denote the time to a terminal event (e.g. death), $C$ an independent censoring time, and $N^*(t)$ the number of recurrent events by time $t$. While censoring is assumed independent of $D$ and $N^*(t)$, no restrictions are placed on the dependence of recurrent or terminal events. Due to censoring, an observation takes the form $\{N(\cdot), U, \delta\}$, where:

$$N(t) = N^*\big(t \wedge C\big), \qquad U = \min(C, D), \qquad \delta = \mathbb{I}\big(D \leq C\big).$$

The observed data $\mathcal{D} = \{N_i(\cdot), U_i, \delta_i\}$ are IID replicates of $\{N(\cdot), U, \delta\}$.

## 1.2 Mean Cumulative Function

**Definition 1.2.1.** The **mean cumulative function** (MCF) is defined as:

$$\mu(t) = \mathbb{E}\big\{N^*(t)\big\}.$$

■

**Proposition 1.2.1.** The MCF is expressible as:

$$\mu(t) = \int_0^t S(u) dR(u),$$

where $S(u) = \mathbb{P}(D > u)$ is the survival function, and $dR(u) = \mathbb{E}\{dN^*(t) | D \geq u\}$. ♦

**Proof.**

$$\begin{aligned}\mu(t) &= \int_0^t \mathbb{E}\{dN^*(u)\} = \int_0^t \mathbb{P}\{dN^*(u) = 1\} = \int_0^t \mathbb{P}\{dN^*(u) = 1, D \geq u\} \\ &= \int_0^t \mathbb{P}(D \geq u)\mathbb{P}\{dN^*(u) = 1 | D \geq u\} = \int_0^t S(u) dR(u).\end{aligned}$$

■

**Example 1.2.1.** Suppose the gap times for the recurrent event process $N^*(t)$ are independent and exponentially distributed with arrival rate $\lambda_A$. In the absence of terminal events, $N^*(t)$ is a Poisson process with MCF $\mu(\tau) = \mathbb{E}\{N^*(\tau)\} = \lambda_A \tau$. Suppose death occurs independently according to an exponential distribution with arrival rate $\lambda_D$, then:

$$\mu(\tau) = \int_0^\tau e^{-\lambda_D t} \lambda_A dt = \frac{\lambda_A}{\lambda_D}\left\{1 - e^{-\lambda_D \tau}\right\}.$$

♠

**Definition 1.2.2.** The **Ghosh Lin** estimator [1] of the MCF is:

$$\hat{\mu}(t) = \int_0^t \hat{S}(u) d\hat{R}(u),$$

where $\hat{S}(u)$ is the Kaplan-Meier estimator for $\mathbb{P}(D > u)$ and $\hat{R}(u)$ is the Nelson-Aalen estimator of cumulative recurrence:

$$\hat{R}(t) = \int_0^t \frac{dN(u)}{Y(u)} = \sum_{i=1}^n \int_0^t \frac{dN_i(u)}{Y(u)}.$$

■

**Discussion 1.2.1.**

- In the recurrent events framework, a subject remains under observation after the occurrence of the event. The subject is only lost to observation due to censoring or a terminal event.
- In the absence of terminal events, $\hat{S}(u) \equiv 1$ and the Ghosh-Lin estimator reduces to the Nelson-Aalen estimator $\hat{R}(t)$.
- Regarding terminal events as censoring results in systematic overestimation of the MCF for $t$ greater than or equal to the first terminal event time.
- If the event of interest can occur at most once per patient, then $\hat{\mu}(t)$ is the standard estimator of the cumulative incidence curve with death acting as a competing risk.

♠

**Example 1.2.2** (**Estimation**). The Ghosh Lin curve may be tabulated as follows. Let $\tau_1 < \cdots < \tau_K$ denote the distinct observed event, censoring, or death times. Let $e_k = dN(\tau_k)$ denote the number of events occurring at $\tau_k$, $d_k = dN^D(\tau_k)$ the number at deaths, and $n_k = Y(\tau_k)$ the number at risk. The instantaneous event rate at time $\tau_k$ is $\hat{r}_k = e_k/n_k$, and the instantaneous hazard is $\hat{h}_k = d_k/n_k$. The Kaplan-Meier estimate of the survival at time $\tau_k$ is the cumulative product:

$$\hat{S}_k = \hat{S}(\tau_k) = \prod_{j \leq k} \big(1 - \hat{h}_j\big).$$

Finally, the Ghosh-Lin estimate MCF at time $\tau_k$ is:

$$\hat{\mu}_k = \hat{\mu}(\tau_k) = \sum_{j \leq k} \hat{S}_j \cdot \hat{r}_j.$$

Consider the following table in which an event occurs at $\tau_1$, followed by a death at $\tau_2$, then a second event at $\tau_3$, a censoring at $\tau_4$, and a final event at $\tau_5$. Note that the number at risk decreases following a terminal event (or censoring), but not after an event. For clarity, the number of censorings $c_k$ is also tracked.

| $\tau_k$ | $c_k$ | $e_k$ | $d_k$ | $n_k$ | $\hat{r}_k$ | $\hat{h}_k$ | $\hat{S}_k$ | $\hat{\mu}_k$ |
|---|---|---|---|---|---|---|---|---|
| $\tau_1$ | 0 | 1 | 0 | 10 | 1/10 | 0 | 1 | 1/10 |
| $\tau_2$ | 0 | 0 | 1 | 10 | 0 | 1/10 | 9/10 | 1/10 |
| $\tau_3$ | 0 | 1 | 0 | 9 | 1/9 | 0 | 9/10 | 2/10 |
| $\tau_4$ | 1 | 0 | 0 | 9 | 0 | 0 | 9/10 | 2/10 |
| $\tau_5$ | 0 | 1 | 0 | 8 | 1/8 | 0 | 9/10 | 5/16 |
| $\vdots$ | | | | | | | | $\vdots$ |

♠

## 1.3 1-Sample Asymptotics

**Proposition 1.3.2** (**Influence Function Expansion**).

$$\sqrt{n}\big\{\hat{\mu}(t) - \mu(t)\big\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(t) + o_p(1), \tag{1.3.1}$$

where:

$$\psi_i(t) = \int_0^t \frac{\mu(u)}{y(u)}dM_i^D(u) - \mu(t)\int_0^t \frac{1}{y(u)}dM_i^D(u) + \int_0^t \frac{S(u)}{y(u)}dM_i(u), \tag{1.3.2}$$

is the influence function, where $y(u) = \mathbb{P}(U \geq u)$ is the probability limit of $n^{-1}Y(u)$,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)dR(u),$$

is the recurrent events martingale and:

$$M_i^D(t) = N_i^D(t) - \int_0^t Y_i(u)dA^D(u)$$

is the terminal event martingale.

♦

**Proof.** The normalized difference is expressible as:

$$\sqrt{n}\big\{\hat{\mu}(t) - \mu(t)\big\} = \sqrt{n}\int_0^t \hat{S}(u)d\hat{R}(u) - \sqrt{n}\int_0^t S(u)dR(u).$$

Adding and subtracting $\sqrt{n}\int_0^t \hat{S}(u)dR(u)$ gives:

$$\sqrt{n}\big\{\hat{\mu}(t) - \mu(t)\big\} = I_1(t) + I_2(t),$$

where:

$$I_1(t) = \sqrt{n}\int_0^t \big\{\hat{S}(u) - S(u)\big\}dR(u)$$
$$I_2(t) = \sqrt{n}\int_0^t \hat{S}(u)d\big\{\hat{R}(u) - R(u)\big\}.$$

From:

$$\sqrt{n}\frac{\big\{\hat{S}(u) - S(u)\big\}}{-S(u)} = \sqrt{n}\big\{\hat{A}^D(u) - A^D(u)\big\} + o_p(1),$$

the first term is asymptotically equivalent to:

$$\begin{aligned} I_1(t) &= -\sqrt{n}\int_0^t \frac{\big\{\hat{S}(u) - S(u)\big\}}{-S(u)} \cdot S(u)dR(u) \\ &= -\int_0^t \sqrt{n}\big\{\hat{A}^D(u) - A^D(u)\big\} \cdot d\mu(u) + o_p(1), \end{aligned}$$

where $A^D(u)$ is the cumulative hazard for terminal events, and $\hat{A}^D(u)$ is the corresponding Nelson-Aalen estimator.

Integrating by parts:

$$\begin{aligned} &-\int_0^t \sqrt{n}\big\{\hat{A}^D(u) - A^D(u)\big\} \cdot d\mu(u) \\ &\quad = -\sqrt{n}\Big[\{\hat{A}^D(u) - A^D(u)\}\mu(u)\Big]_{u=0}^{u=t} + \sqrt{n}\int_0^t \mu(u) \cdot d\big\{\hat{A}^D(u) - A^D(u)\big\} \\ &\quad = -\sqrt{n}\{\hat{A}^D(t) - A^D(t)\}\mu(t) + \sqrt{n}\int_0^t \mu(u) \cdot d\big\{\hat{A}^D(u) - A^D(u)\big\}. \end{aligned}$$

Using the martingale expansion of the Nelson-Aalen estimator:

$$\sqrt{n}\{\hat{A}^D(t) - A^D(t)\} = \frac{1}{\sqrt{n}}\int_0^t \frac{dM^D(u)}{y(u)} + o_p(1),$$

$I_1(t)$ becomes:

$$I_1(t) = -\frac{\mu(t)}{\sqrt{n}}\int_0^t \frac{dM^D(u)}{y(u)} + \frac{1}{\sqrt{n}}\int_0^t \frac{\mu(u)}{y(u)}dM^D(u) + o_p(1).$$

Now consider $I_2(t)$. By uniform consistency of the Kaplan-Meier estimator:

$$I_2(t) = \sqrt{n}\int_0^t S(u)d\big\{\hat{R}(u) - R(u)\big\} + o_p(1).$$

From Ghosh Lin (A.8):

$$\sqrt{n}\big\{\hat{R}(t) - R(t)\big\} = \frac{1}{\sqrt{n}}\int_0^t \frac{dM(u)}{y(u)} + o_p(1),$$

where:

$$M(t) = N(t) - \int_0^t Y(u)dR(u).$$

Consequently:

$$\sqrt{n}\int_0^t S(u)d\big\{\hat{R}(u) - R(u)\big\} = \frac{1}{\sqrt{n}}\int_0^t \frac{S(u)}{y(u)}dM(u) + o_p(1),$$

and overall:

$$\sqrt{n}\big\{\hat{\mu}(t) - \mu(t)\big\} = \int_0^t \frac{\mu(u)}{y(u)}dM^D(u) - \mu(t)\int_0^t \frac{1}{y(u)}dM^D(u) + \int_0^t \frac{S(u)}{y(u)}dM(u) + o_p(1).$$

■

**Corollary 1.3.1.** $\sqrt{n}\big\{\hat{\mu}(t) - \mu(t)\big\}$ converges weakly to a mean-zero Guassian process with covariance function:

$$\gamma(s,t) = \mathbb{E}\big\{\psi(s)\cdot\psi(t)\big\},$$

which is estimable by:

$$\hat{\gamma}(s,t) = \frac{1}{n}\sum_{i=1}^n \psi_i(s)\cdot\psi_i(t).$$

In particular:

$$\sqrt{n}\big\{\hat{\mu}(t) - \mu(t)\big\} \rightsquigarrow W\big\{\sigma^2_{\text{MCF}}(t)\big\},$$

with:

$$\sigma^2_{\text{MCF}}(t) = \mathbb{E}\big\{\psi^2(t)\big\},$$

♣

**Discussion 1.3.1.** Since $\mu(t)$ is non-negative, point-wise confidence intervals may be constructed via:

$$\sqrt{n}\big\{\ln\hat{\mu}(t) - \ln\mu(t)\big\} = \sqrt{n}\frac{\big\{\hat{\mu}(t) - \mu(t)\big\}}{\mu(t)} + o_p(1).$$

In particular, an asymptotic $(1-\alpha)$ level confidence interval is given by:

$$\hat{\mu}(t)\exp\left\{\pm\frac{z_{1-\alpha/2}\hat{\sigma}^2_{\text{MCF}}(t)}{\sqrt{n}\hat{\mu}(t)}\right\}.$$

♠

## 1.4 2-Sample Asymptotics

**Definition 1.4.1.** The two sample log-rank statistic for equality of two MCFs, $H_0$ : $\mu_1(t) = \mu_0(t)$ for all $t \in [0, \tau]$ is:

$$T_{\text{LR}}(\tau) = \int_0^\tau \omega(t) d\big\{\hat{\mu}_1(t) - \hat{\mu}_0(t)\big\},$$

where:

$$\omega(t) = \frac{n}{n_1 n_0} \frac{Y_1(t) Y_0(t)}{Y_1(t) + Y_0(t)}.$$

■

**Proposition 1.4.3.** Under $H_0 : \mu_1(t) = \mu_0(t)$:

$$\Big(\frac{n_1 n_0}{n}\Big)^{1/2} T_{\text{LR}} \xrightarrow{d} N\big(0, \sigma^2_{\text{LR}}\big),$$

where:

$$\hat{\sigma}^2_{\text{LR}} = \frac{n_0}{n n_1} \sum_{i=1}^{n_1} \left\{ \int_0^\tau \omega(t) d\psi_{1i}(t) \right\}^2 + \frac{n_1}{n n_0} \sum_{i=1}^{n_0} \left\{ \int_0^\tau \omega(t) d\psi_{0i}(t) \right\}^2$$

and $\psi_{ij}(t)$ is the influence function for subject $i$ in arm $j$ from (1.3.2). ♦

**Proof.** Consider the log-rank statistic:

$$\begin{aligned}\Big(\frac{n_1 n_0}{n}\Big)^{1/2} \hat{T}_{\text{LR}} = & \Big(\frac{n_1 n_0}{n}\Big)^{1/2} \int_0^\tau \omega(t) d\big\{\hat{\mu}_1(t) - \mu_1(t)\big\} \\ & - \Big(\frac{n_1 n_0}{n}\Big)^{1/2} \int_0^\tau \omega(t) d\big\{\hat{\mu}_0(t) - \mu_0(t)\big\}\end{aligned}$$

From (1.3.1):

$$\sqrt{n_j}\big\{\hat{\mu}_j(t) - \mu_j(t)\big\} = \frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} \psi_{ij}(t) + o_p(1).$$

Applied to the log-rank statistic:

$$\begin{aligned}\Big(\frac{n_1 n_0}{n}\Big)^{1/2} \hat{T}_{\text{LR}} = & \Big(\frac{n_0}{n}\Big)^{1/2} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \int_0^\tau \omega(t) d\psi_{i1}(t) \\ & - \Big(\frac{n_1}{n}\Big)^{1/2} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \int_0^\tau \omega(t) d\psi_{i0}(t) + o_p(1).\end{aligned}$$

Suppose $n_1, n_0 \to \infty$ such that $n_0/n \to \pi_0$ and $n_1/n \to \pi_1$, the normalized log-rank statistic converges in distribution to a zero-mean random variable with variance:

$$\sigma^2_{\text{LR}} = \pi_0 \cdot \mathbb{E}\left\{ \int_0^\tau \omega(t) d\psi_1(t) \right\}^2 + \pi_1 \cdot \mathbb{E}\left\{ \int_0^\tau \omega(t) d\psi_0(t) \right\}^2$$

■

# Area Under the Mean Cumulative Function

## 2.1 Asymptotics

**Definition 2.1.1.** Define the area under the MCF as:

$$U(\tau) = \int_0^\tau \mu(t)dt = \int_0^\tau \left\{ \int_0^t S(u)dR(u) \right\} dt.$$

By exchanging the order of integration:

$$U(\tau) = \int_0^\tau \int_0^\tau \mathbb{I}(u \leq t) S(u)dR(u)dt = \int_0^\tau (\tau - u)S(u)dR(u).$$

■

**Proposition 2.1.1** (**Influence Function Expansion**).

$$\sqrt{n}\big\{\hat{U}(\tau) - U(\tau)\big\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i(\tau) + o_p(1),$$

where:

$$\psi_i(\tau) = \int_0^\tau \frac{(\tau - u)S(u)}{y(u)} dM_i(u) + \int_0^\tau \frac{(\tau - u)\mu(u)}{y(u)} dM_i^D(u),$$

is the influence function, $y(u) = \mathbb{P}(U \geq u)$ is the probability of being at risk,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)dR(u),$$

is the recurrent events martingale and:

$$M_i^D(t) = N_i^D(t) - \int_0^t Y_i(u)dA^D(u)$$

is the terminal event martingale.

♦

**Proof.** Consider the distribution of:

$$\sqrt{n}\big\{\hat{U}(\tau) - U(\tau)\big\} = \sqrt{n} \int_0^\tau (\tau - u)\hat{S}(u)d\hat{R}(u) - \sqrt{n} \int_0^\tau (\tau - u)S(u)dR(u).$$

Adding and subtracting $\sqrt{n} \int_0^\tau (\tau - u)\hat{S}(u)dR(u)$ gives:

$$\sqrt{n}\big\{\hat{U}(\tau) - U(\tau)\big\} = I_1(\tau) + I_2(\tau),$$

where:

$$I_1(\tau) = \sqrt{n}\int_0^\tau (\tau - u)\hat{S}(u)d\big\{\hat{R}(u) - R(u)\big\},$$
$$I_2(\tau) = \sqrt{n}\int_0^\tau (\tau - u)\big\{\hat{S}(u) - S(u)\big\}dR(u).$$

From Ghosh and Lin (A.8):

$$\sqrt{n}\big\{\hat{R}(t) - R(t)\big\} = \frac{1}{\sqrt{n}}\int_0^t \frac{dM(u)}{y(u)} + o_p(1),$$

the first integral is expressible as:

$$\begin{aligned} I_1(\tau) &= \int_0^\tau (\tau - u)\hat{S}(u)\cdot\sqrt{n}d\big\{\hat{R}(u) - R(u)\big\} \\ &= \int_0^\tau (\tau - u)\hat{S}(u)\cdot\frac{1}{\sqrt{n}}\frac{dM(u)}{y(u)} + o_p(1) \\ &= \frac{1}{\sqrt{n}}\int_0^\tau \frac{(\tau - u)S(u)}{y(u)}dM(u) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\int_0^\tau \frac{(\tau - u)S(u)}{y(u)}dM_i(u)\right\} + o_p(1). \end{aligned}$$

From:

$$\sqrt{n}\frac{\big\{\hat{S}(u) - S(u)\big\}}{-S(u)} = \sqrt{n}\big\{\hat{A}^D(u) - A^D(u)\big\} + o_p(1),$$

the second integral is expressible as:

$$\begin{aligned} I_2(\tau) &= -\sqrt{n}\int_0^\tau (\tau - u)\frac{\big\{\hat{S}(u) - S(u)\big\}}{-S(u)}\cdot S(u)dR(u) \\ &= -\sqrt{n}\int_0^\tau (\tau - u)\big\{\hat{A}^D(u) - A^D(u)\big\}\cdot d\mu(u) + o_p(1). \end{aligned}$$

Integrating by parts:

$$\begin{aligned} &-\sqrt{n}\int_0^\tau (\tau - u)\big\{\hat{A}^D(u) - A^D(u)\big\}d\mu(u) \\ &\quad = -\Big[\sqrt{n}(\tau - u)\big\{\hat{A}^D(u) - A^D(u)\big\}\mu(u)\Big]_{u=0}^{u=\tau} + \sqrt{n}\int_0^\tau (\tau - u)\mu(u)d\big\{\hat{A}^D(u) - A^D(u)\big\}. \end{aligned}$$

The first term vanishes, leaving:

$$I_2(\tau) = \sqrt{n}\int_0^\tau (\tau - u)\mu(u)d\big\{\hat{A}^D(u) - A^D(u)\big\}.$$

Using the martingale expansion for the cumulative hazard:

$$\sqrt{n}\big\{\hat{A}^D(t) - A^D(t)\big\} = \frac{1}{\sqrt{n}}\int_0^t \frac{dM^D(u)}{y(u)} + o_p(1),$$

the second integral takes the form:

$$\begin{aligned} I_2(\tau) &= \sqrt{n}\int_0^\tau (\tau - u)\mu(u) d\big\{\hat{A}^D(u) - A^D(u)\big\} \\ &= \int_0^\tau (\tau - u)\mu(u) \cdot \sqrt{n} d\big\{\hat{A}^D(u) - A^D(u)\big\} \\ &= \int_0^\tau (\tau - u)\mu(u) \cdot \frac{1}{\sqrt{n}} \frac{dM^D(u)}{y(u)} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \int_0^\tau \frac{(\tau - u)\mu(u)}{y(u)} dM_i^D(u) \right\} + o_p(1). \end{aligned}$$

■

# References

[1] Lin DY. "Non-parametric inference for cumulative incidence functions in competing risks studies". In: *Statistics in Medicine* 16.3 (1997), pp. 901–910.