Introduction to Empirical Processes and Semiparametric Inference

Yu Gu

June 2, 2021

Outline

1. Introduction: empirical processes
2. Introduction: semiparametric models
3. Examples of theoretical justification
   - Cox model with right-censored data
   - Transformation model with interval-censored data

Outline

1. Introduction: empirical processes
2. Introduction: semiparametric models
3. Examples of theoretical justification
   - Cox model with right-censored data
   - Transformation model with interval-censored data

What is an empirical process?

- A *stochastic process* is a collection of random variables $\{X(t), t \in T\}$ on the same probability space, indexed by an arbitrary index set T.
- In general, an *empirical process* is a stochastic process based on a random sample, usually of n i.i.d. random variables $X_1, \ldots, X_n$.

Example: empirical distribution function

Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with cumulative distribution function (c.d.f.) F. Then the *empirical distribution function* (e.d.f.) is defined as

$$\mathbb{F}_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq t), \quad t \in \mathbb{R}.$$

$\mathbb{F}_n(t)$ is one of the simplest examples of an empirical process.

Example: Kaplan-Meier estimator

Let $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ be a sample of right-censored failure time observations. Then the *Kaplan-Meier estimator* of the survival function is given by

$$\widehat{S}(t) = \prod_{k: T_k^0 \leq t} \left\{ 1 - \frac{\sum_{i=1}^n \delta_i \mathbf{1}(X_i = T_k^0)}{\sum_{i=1}^n \mathbf{1}(X_i \geq T_k^0)} \right\},$$

where $T_1^0 < T_2^0 < \cdots < T_K^0$ are unique observed failure times.

General features of an empirical process

- The i.i.d. sample $X_1, \ldots, X_n$ is drawn from a probability measure P on an arbitrary sample space $\mathcal{X}$.
- Define the *empirical measure* to be $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$, where δ_x denotes the Dirac measure at x.
- For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, define

$$\mathbb{P}_n f := \int f d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

- For any class $\mathcal{F}$ of such real-valued functions on $\mathcal{X}$, $\{\mathbb{P}_n f : f \in \mathcal{F}\}$ is the empirical process indexed by $\mathcal{F}$.

Start with the classical e.d.f. $\mathbb{F}_n$

- Setting $\mathcal{X} = \mathbb{R}$, $\mathbb{F}_n$ can be re-expressed as the empirical process $\{\mathbb{P}_n f : f \in \mathcal{F}\}$, where $\mathcal{F} = \{\mathbf{1}(x \le t), t \in \mathbb{R}\}$.
- By the law of large numbers, $\mathbb{F}_n(t) \stackrel{a.s.}{\to} F(t)$ for each $t \in \mathbb{R}$.
- By the central limit theorem, for each $t \in \mathbb{R}$,

$$\mathbb{G}_n(t) := \sqrt{n}(\mathbb{F}_n(t) - F(t)) \stackrel{d}{\to} N\Big(0, F(t)(1 - F(t))\Big).$$

- From the functional perspective, uniform results over $t \in \mathbb{R}$ would be more appealing.
  - Need theory of empirical processes

Strengthened results on $\mathbb{F}_n$ and $\mathbb{G}_n$

- Glivenko (1933) and Cantelli (1933) demonstrated that the previous result could be strengthened to

$$\|\mathbb{F}_n - F\|_\infty = \sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \stackrel{a.s.}{\to} 0.$$

- Donsker (1952) showed that

$$\mathbb{G}_n \stackrel{d}{\to} \mathbb{B}(F) \quad \text{in } \ell^\infty(\mathbb{R}),$$

where $\mathbb{B}$ is the *standard Brownian bridge process* on [0, 1]; for any index set T, $\ell^\infty(T)$ denotes the space of all bounded functions $f : T \mapsto \mathbb{R}$.

Extend to general empirical processes

- Properties of the approximation of Pf by $\mathbb{P}_n f$, uniformly in $\mathcal{F}$
  - the random quantity $\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|$
  - the empirical process $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P)$
- Two special classes
  - **Glivenko-Cantelli:** $\mathcal{F}$ is P-Glivenko-Cantelli if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| \stackrel{a.s.}{\to} 0.$$

  - **Donsker:** $\mathcal{F}$ is P-Donsker if

$$\mathbb{G}_n \stackrel{d}{\to} \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{F}),$$

    where $\mathbb{G}$ is a mean zero Gaussian process indexed by $\mathcal{F}$.

Remarks

- Glivenko-Cantelli: uniform almost surely convergence
- Donsker: uniform central limit theorem
- Donsker $\Rightarrow$ Glivenko-Cantelli (GC)
- GC or Donsker properties depend crucially on the complexity (or entropy) of $\mathcal{F}$.

Complexity of $(\mathcal{F}, \|\cdot\|)$

- Covering number
  - denoted by $N(\epsilon, \mathcal{F}, \|\cdot\|)$
  - minimum number of balls $B(f;\epsilon) := \{g : \|g - f\| < \epsilon\}$ needed to cover $\mathcal{F}$
  - entropy: $\log N(\epsilon, \mathcal{F}, \|\cdot\|)$
- Bracketing number
  - denoted by $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$
  - minimum number of brackets $[\ell, u]$[1] with $\|\ell - u\| < \epsilon$ needed to cover $\mathcal{F}$
  - entropy with bracketing: $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$

[1] *Given two functions $\ell(\cdot)$ and $u(\cdot)$, the bracket $[\ell, u]$ is the set of all functions $f \in \mathcal{F}$ with $\ell(x) \leq f(x) \leq u(x)$, for all $x \in \mathcal{X}$.*

GC theorems

Theorem (GC by bracketing)

Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is a GC class.

Theorem (GC by entropy)

Let $\mathcal{F}$ be a class of measurable functions with envelope[a] F such that $PF < \infty$. Let $\mathcal{F}_M$ be the class of functions $f\mathbf{1}\{F \leq M\}$ where f ranges over $\mathcal{F}$. Then $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ both almost surely and in mean, if and only if

$$\frac{1}{n}\log N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) \xrightarrow{p} 0,$$

for every $\epsilon > 0$ and $M > 0$.

[a] *An envelop function is any function that can bound every function in $\mathcal{F}$ everywhere. That is, for each $f \in \mathcal{F}$, $|f(x)| \leq F(x)$ for any $x \in \mathcal{X}$.*

Donsker theorems

Define the *bracketing entropy integral* as

$$J_{[]}(\delta, \mathcal{F}, L_r(P)) := \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_r(P))} d\epsilon.$$

Theorem (Donsker by bracketing entropy integral)

Suppose that $\mathcal{F}$ is a class of measurable functions with square-integrable (measurable) envelope F and such that $J_{[]}(\infty, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is P-Donsker.

Donsker theorems (cont.)

Define the *uniform entropy integral* as

$$J(\delta, \mathcal{F}, L_r) = \int_0^\delta \sup_Q \sqrt{\log N\left(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right)}\, d\epsilon.$$

Theorem (Donsker by uniform entropy integral)

Let $\mathcal{F}$ be a pointwise-measurable class of measurable functions with (measurable) envelope F such that $PF^2 < \infty$. If $J(\infty, \mathcal{F}, L_2) < \infty$ then $\mathcal{F}$ is P-Donsker.

Some useful results

Suppose $\mathcal{F}$ is Donsker.

1. Any subset of $\mathcal{F}$ is Donsker.
2. $\overline{\mathcal{F}}$ is Donsker, where $\overline{\mathcal{F}}$ denotes the set of all f for which there exists a sequence f_n in $\mathcal{F}$ with $f_n \to f$ both pointwise and in $L_2(P)$.
3. The symmetric convex hull of $\mathcal{F}$ is Donsker, where $\text{sconv}\mathcal{F} = \left\{\sum_i \lambda_i f_i : \ f_i \in \mathcal{F}, \ \sum_i |\lambda_i| \le 1\right\}$.
4. Any Lipschitz-continuous transformation of $\mathcal{F}$ is Donsker.

M-estimators

- Definition:
  - Metric space: (Θ, d)
  - $m_\theta : \mathcal{X} \to \mathbb{R}$, for each $\theta \in \Theta$
  - "Empirical gain": $M_n(\theta) = \mathbb{P}_n m_\theta$
  - *M*-estimator: $\hat{\theta}_n = \arg\max_{\theta \in \Theta} M_n(\theta)$
- Examples:
  - Maximum (penalized) likelihood estimator
  - Least squares estimator
  - Nonparametric maximum likelihood estimator, e.g., Grenander estimator, where Θ is the set of all non-increasing densities on $[0, \infty)$ and $m_\theta(x) = \log \theta(x)$

Application: consistency of M-estimators

- Two assumptions:
  1. $\mathcal{F} := \{m_\theta(\cdot) : \theta \in \Theta\}$ is P-GC
  2. θ_0 is a well-separated maximizer of $M(\theta) = Pm_\theta$, i.e., for every $\delta > 0$, $M(\theta_0) > \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta)$.
- For fixed $\delta > 0$, let $\psi(\delta) = M(\theta_0) - \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta)$

$$
\begin{aligned}
\left\{ d(\hat{\theta}_n, \theta_0) \geq \delta \right\} &\Rightarrow M(\hat{\theta}_n) \leq \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta) \\
&\Leftrightarrow M(\hat{\theta}_n) - M(\theta_0) \leq -\psi(\delta) \\
&\Rightarrow M(\hat{\theta}_n) - M(\theta_0) + \left(M_n(\theta_0) - M_n(\hat{\theta}_n) \right) \leq -\psi(\delta) \\
&\Rightarrow 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \psi(\delta)
\end{aligned}
$$

$$
\Rightarrow \mathbb{P}\left(d(\hat{\theta}_n, \theta_0) \geq \delta \right) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \psi(\delta)/2 \right) \to 0.
$$

Outline

1. Introduction: empirical processes
2. Introduction: semiparametric models
3. Examples of theoretical justification
   - Cox model with right-censored data
   - Transformation model with interval-censored data

General statistical models

- Collection of probability measures $\{P \in \mathcal{P}\}$ that specify the distribution of a random observation X.
- Parametric models: $\mathcal{P} = \left\{P_\theta : \theta \in R^d\right\}$
- Nonparametric models: $\mathcal{P} = \{P : P \text{ is any distribution}\}$
- Semiparametric models: $\mathcal{P} = \left\{P_{\theta,\eta} : \theta \in R^d, \eta \in \mathcal{M}\right\}$
  - $\mathcal{M}$ is an infinite-dimensional space
  - θ: parameter of interest
  - η: nuisance parameter

Why semiparametric models?

- Only interested in some specific variable relationships: treatment effect, risk effect, etc.
- Not necessary to specify delicately those parameters that contribute to the probability distribution but are less interesting.
- Models are flexible and robust and parameters of interest are easy to be interpreted.

Primary goals of semiparametric inference

- Select an appropriate model for inference on X.
- Estimate (θ, η) (sometimes θ alone is the main focus).
- Conduct inference (e.g., confidence intervals or bands) for the parameters of interest.
  - Usually for θ only
  - Sometimes the convergence rate for η is not $O_p(n^{-1/2})$

Asymptotic properties of an estimator

- Consistency: $\hat{\theta}_n \stackrel{p}{\to} \theta_0$
- Asymptotic normality: $\sqrt{n}(\hat{\theta}_n - \theta_0) \stackrel{d}{\to} G$
- Semiparametric efficiency

On semiparametric efficiency

- *Efficient:* achieve the smallest asymptotic variance among all *regular* estimators[2].
- *Information:* inverse of the asymptotic variance.
- The information for estimation under $\mathcal{P}$ is **worse** than the information under any parametric submodel $\mathcal{P}_0$.
- *Semiparametric efficient:* attain minimum information over all efficient estimators for all $\mathcal{P}_0$.
- $\mathcal{P}_0$ with minimum information is called a *least favorable* submodel.
- Usually only need to consider one-dimensional parametric submodels $\{P_t : t \in [0, \epsilon)\}$.

[2] *A regular sequence of estimators is one whose asymptotic distribution remains the same in shrinking neighborhoods of the true parameter value.*

Semiparametric regression models in survival analysis

- Right-censored survival data
  - Proportional hazards model: $\lambda(t \mid X) = \lambda(t)\exp\{\beta^T X\}$
  - Proportional odds model: $\text{logit}\, S(t \mid X) = h(t) + \beta^T X$
  - Accelerated failure time model: $\log T = \beta^T X + \epsilon$
  - Linear transformation model: $\log \Lambda(T) = \beta^T X + \epsilon$
  - Additive risk model: $\lambda(t \mid X) = \lambda(t) + \beta^T X$
- Interval-censored survival data
  - Proportional hazards model
  - Proportional odds model
  - AFT model
  - Transformation models

Mathematical tools

- Martingale theory for counting process
- Empirical process theory
- Semiparametric efficiency theory

Outline

Outline

Introduction

- Data: $\left(Y_i = T_i \wedge C_i,\ R_i = \mathbf{1}\left(T_i \leq C_i\right),\ X_i\right), \quad i = 1, \ldots, n$
- Assumptions:
  1. T and C are independent given X
  2. At least a proportion of subjects survive up to the study end time τ, i.e., $Pr(T > \tau) > \delta > 0$
- Model: Cox PH model

$$h_{T|X}(t \mid x) = \lambda(t)e^{x'\beta}$$

- Parameters of interest: β and $\Lambda(t) = \int_0^t \lambda(s)ds$

Introduction (cont.)

- Observed likelihhod function:

$$\prod_{i=1}^{n}\left\{\left[\lambda\left(Y_i\right)e^{X_i'\beta}\right]^{R_i}e^{-\Lambda(Y_i)e^{X_i'\beta}}h_{C|X}\left(Y_i\mid X_i\right)^{1-R_i}e^{-H_{C|X}(Y_i|X_i)}f_X\left(X_i\right)\right\}$$

- Parameter space:

$$\left\{(\beta,\Lambda):\ \beta\in\mathbb{R}^p,\ \Lambda(t)\text{ is an increasing function with }\Lambda(0)=0\right\}$$

- Nonparametric maximum likelihood approach:

$$\ell_n(\beta,\Lambda)=\sum_{i=1}^{n}\left\{R_i\left[X_i'\beta+\log\Delta\Lambda\left(Y_i\right)\right]-\Lambda\left(Y_i\right)e^{X_i'\beta}\right\}$$

- Facts:
  1. $\hat{\Lambda}_n$ is a step function with non-negative jumps only at Y_i.
  2. Under Assumption 2, $\hat{\Lambda}_n(\tau)<\infty$.

NPMLEs

Differentiating ℓ_n with respect to $\{\beta, \Delta\Lambda(Y_1), \ldots, \Delta\Lambda(Y_n)\}$ and solving the resulting equations, we obtain

$$\sum_{i=1}^{n} R_i \left[X_i - \frac{\sum_{Y_j \geq Y_i} X_j e^{X_j' \hat{\beta}_n}}{\sum_{Y_j \geq Y_i} e^{X_j' \hat{\beta}_n}} \right] = 0$$

and

$$\hat{\Lambda}_n(t) = \sum_{Y_i \leq t} \frac{R_i}{\sum_{Y_j \geq Y_i} e^{X_j' \hat{\beta}_n}}.$$

Here, $\hat{\beta}_n$ is exactly the maximizer of the *partial likelihood function* proposed in Cox (1972), and $\hat{\Lambda}_n(t)$ is exactly the *Breslow estimator*.

Consistency

Theorem (Consistency)

Assume that X is bounded and has a continuous density, λ_0 is continuous and positive on $[0, \tau]$. Then

$$\|\hat{\beta}_n - \beta_0\| + \sup_{t \in [0,\tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \xrightarrow{p} 0.$$

Lemma ($\theta \in \mathbb{R}^k$)

Suppose $M_n(\theta)$ and $M_0(\theta)$ are strictly concave function and for any compact set $K \subset \Theta$,

$$\sup_{\theta \in K} |M_n(\theta) - M_0(\theta)| \xrightarrow{p} 0.$$

Moreover, $\hat{\theta}_n$ and θ_0 are unique maximizer of $M_n(\theta)$ and $M_0(\theta)$ respectively. Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Notations

$$S_{0,n}(t,\beta) = \mathbb{P}_n \mathbf{1}(Y \geq t) e^{X'\beta}, \quad S_0(t,\beta) = \mathbb{P}\mathbf{1}(Y \geq t) e^{X'\beta},$$

$$S_{1,n}(t,\beta) = \mathbb{P}_n \mathbf{1}(Y \geq t) X e^{X'\beta}, \quad S_1(t,\beta) = \mathbb{P}\mathbf{1}(Y \geq t) X e^{X'\beta},$$

$$S_{2,n}(t,\beta) = \mathbb{P}_n \mathbf{1}(Y \geq t) XX' e^{X'\beta}, \quad S_2(t,\beta) = \mathbb{P}\mathbf{1}(Y \geq t) XX' e^{X'\beta},$$

$$M_n(\beta) = \mathbb{P}_n R \log \frac{e^{X'\beta}}{S_{0,n}(Y,\beta)}, \quad M_0(\beta) = \mathbb{P} R \log \frac{e^{X'\beta}}{S_0(Y,\beta)},$$

$$\hat{\Lambda}_n(t) = \mathbb{P}_n \frac{R\mathbf{1}(Y \leq t)}{S_{0,n}(Y,\hat{\beta}_n)}, \quad \Lambda_0(t) = \mathbb{P} \frac{R\mathbf{1}(Y \leq t)}{S_0(Y,\beta_0)}.$$

Proof of consistency of $\hat{\beta}_n$

1. Concavity of $M_n(\beta)$ and $M_0(\beta)$.

$$\nabla^2_{\beta\beta} M_n(\beta) = -\mathbb{P}_n[R \frac{S_{2,n}(Y,\beta)\, S_{0,n}(Y,\beta) - S_{1,n}(Y,\beta)^{\otimes 2}}{S_{0,n}(Y,\beta)^2}],$$

$$\nabla^2_{\beta\beta} M_0(\beta) = -\mathbb{P}[R \frac{S_2(Y,\beta)\, S_0(Y,\beta) - S_1(Y,\beta)^{\otimes 2}}{S_0(Y,\beta)^2}].$$

$$\left\{ \mathbf{1}(Y \geq t) e^{X'\beta}, \mathbf{1}(Y \geq t) X e^{X'\beta}, \mathbf{1}(Y \geq t) X X' e^{X'\beta} :\ \beta \in K, t \in [0,\tau] \right\}$$

is a GC class, for any compact set $K \subset \mathbb{R}$.

$\Rightarrow S_{q,n}(t,\beta) \to S_q(t,\beta)$ uniformly in $K \times [0,\tau]$.

$\Rightarrow \nabla^2_{\beta\beta} M_n(\beta) \to \nabla^2_{\beta\beta} M_0(\beta) < 0$ uniformly.

2. $\sup_{\beta \in K} |M_n(\beta) - M_0(\beta)| \xrightarrow{p} 0.$ $\checkmark$
3. $\hat{\beta}_n$ is the unique maximizer of $M_n(\beta)$. $\checkmark$ ($M_n(\beta)$ is essentially the PLL)
4. $\hat{\beta}_0$ is the unique maximizer of $M_0(\beta)$. It suffices to show $\nabla_\beta M_0(\beta_0) = 0$.

Proof of consistency of $\hat{\Lambda}_n$

$$
\begin{aligned}
&\hat{\Lambda}_n(t) - \Lambda_0(t) \\
&= \mathbb{P}_n[\mathbf{1}(Y \le t)R/S_{0,n}(Y, \hat{\beta}_n)] - \mathbb{P}[\mathbf{1}(Y \le t)R/S_0(Y, \beta_0)] \\
&= (\mathbb{P}_n - \mathbb{P})[\mathbf{1}(Y \le t)R/S_{0,n}(Y, \hat{\beta}_n)] \\
&\quad + \mathbb{P}[\mathbf{1}(Y \le t)R/S_{0,n}(Y, \hat{\beta}_n) - \mathbf{1}(Y \le t)R/S_0(Y, \hat{\beta}_n)] \\
&\quad + \mathbb{P}[\mathbf{1}(Y \le t)R/S_0(Y, \hat{\beta}_n) - \mathbf{1}(Y \le t)R/S_0(Y, \beta_0)] \\
&=: \text{(i)} + \text{(ii)} + \text{(iii)} \\
&\to 0 \text{ uniformly over } t \in [0, \tau].
\end{aligned}
$$

(i): GC theorem.
(ii): $S_{0,n}(t, \beta) \to S_0(t, \beta)$ uniformly in $K \times [0, \tau]$ & $\hat{\beta}_n \xrightarrow{p} \beta_0$.
(iii): $S_0(t, \beta)$ is differentiable with respect to β.

Asymptotic normality

Theorem (Asymptotic normality)

Under regularity conditions,

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0, \hat{\Lambda}_n - \Lambda_0\right) \xrightarrow{d} G_1 \times G_2 \quad \textit{in } \mathbb{R}^p \times \ell^\infty[0, \tau],$$

where G_1 is a normal distribution with mean zero and variance Σ_β, and G_2 is a Brownian bridge with covariance $\Sigma_\Lambda(t, s)$.

Proof of asymptotic normality of $\hat{\beta}_n$

Since $\nabla_\beta M_n(\hat{\beta}_n) = \nabla_\beta M_0(\beta_0) = 0$, we can use telescopic expansion:

$$
\begin{aligned}
0 &= \mathbb{P}_n[R(X - \frac{S_{1,n}(Y,\hat{\beta}_n)}{S_{0,n}(Y,\hat{\beta}_n)})] - \mathbb{P}[R(X - \frac{S_{1,n}(Y,\hat{\beta}_n)}{S_{0,n}(Y,\hat{\beta}_n)})] \\
&\quad + \mathbb{P}[R(X - \frac{S_{1,n}(Y,\hat{\beta}_n)}{S_{0,n}(Y,\hat{\beta}_n)})] - \mathbb{P}[R(X - \frac{S_1(Y,\hat{\beta}_n)}{S_0(Y,\hat{\beta}_n)})] \\
&\quad + \mathbb{P}[R(X - \frac{S_1(Y,\hat{\beta}_n)}{S_0(Y,\hat{\beta}_n)})] - \mathbb{P}[R(X - \frac{S_1(Y,\beta_0)}{S_0(Y,\beta_0)})] \\
&= (\mathbb{P}_n - \mathbb{P})[R(X - \frac{S_{1,n}(Y,\hat{\beta}_n)}{S_{0,n}(Y,\hat{\beta}_n)})] \quad \cdots\cdots \text{(i)} \\
&\quad - \mathbb{P}[\frac{R(\mathbb{P}_n - \mathbb{P})[Xe^{X'\hat{\beta}_n}\mathbf{1}(Y \geq y)]|_{y=Y}}{S_{0,n}(Y,\hat{\beta}_n)}] \quad \cdots\cdots \text{(ii)} \\
&\quad + \mathbb{P}[\frac{RS_1(Y,\hat{\beta}_n)(\mathbb{P}_n - \mathbb{P})[\mathbf{1}(Y \geq y)e^{X'\hat{\beta}_n}]|_{y=Y}}{S_{0,n}(Y,\hat{\beta}_n)S_0(Y,\hat{\beta}_n)}] \quad \cdots\cdots \text{(iii)} \\
&\quad - \mathbb{P}[R\frac{S_2(Y,\beta^*)S_0(Y,\beta^*) - S_1(Y,\beta^*)^{\otimes 2}}{S_0(Y,\beta^*)^2}](\hat{\beta}_n - \beta_0)
\end{aligned}
$$

Proof of asymptotic normality of $\hat{\beta}_n$ (cont.)

(i)+(ii)+(iii) is an empirical process

$$(\mathbb{P}_n - \mathbb{P})[R(X - \frac{S_{1,n}(Y,\hat{\beta}_n)}{S_{0,n}(Y,\hat{\beta}_n)}) - Xe^{X'\hat{\beta}_n}\widetilde{\mathbb{P}}\frac{\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_{0,n}(\widetilde{Y},\hat{\beta}_n)} + e^{X'\hat{\beta}_n}\widetilde{\mathbb{P}}\frac{S_1(\widetilde{Y},\hat{\beta}_n)\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_{0,n}(\widetilde{Y},\hat{\beta}_n)S_0(\widetilde{Y},\hat{\beta}_n)}]. \tag{1}$$

By applying the functional central limit theorem[3],

$$(1) = (\mathbb{P}_n - \mathbb{P})[R(X - \frac{S_1(Y,\beta_0)}{S_0(Y,\beta_0)}) - Xe^{X'\beta_0}\widetilde{\mathbb{P}}\frac{\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_0(\widetilde{Y},\beta_0)} + e^{X'\beta_0}\widetilde{\mathbb{P}}\frac{S_1(\widetilde{Y},\beta_0)\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_0(\widetilde{Y},\beta_0)^2}] + o_p(1/\sqrt{n}).$$

Thus,

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_n - \beta_0) \\
&= \left\{\mathbb{P}[R\frac{S_2(Y,\beta_0)S_0(Y,\beta_0) - S_1(Y,\beta_0)^{\otimes 2}}{S_0(Y,\beta_0)^2}]\right\}^{-1} \times \\
&\quad \mathbb{G}_n[R(X - \frac{S_1(Y,\beta_0)}{S_0(Y,\beta_0)}) - Xe^{X'\beta_0}\widetilde{\mathbb{P}}\frac{\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_0(\widetilde{Y},\beta_0)} + e^{X'\beta_0}\widetilde{\mathbb{P}}\frac{S_1(\widetilde{Y},\beta_0)\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_0(\widetilde{Y},\beta_0)^2}] + o_p(1).
\end{aligned}$$

[3] *Theorem 2 of Section 4.3.4 in Zeng's lecture notes*

Proof of asymptotic normality of $\hat{\Lambda}_n$

From the consistency proof, by applying the functional central limit theorem, we have

$$\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$$

$$= \mathbb{G}_n[\mathbf{1}(Y \leq t)R/S_0(Y, \beta_0)] - e^{X'\beta_0}\widetilde{\mathbb{P}}[\frac{\mathbf{1}(\widetilde{Y} \leq t)\mathbf{1}(Y \geq \widetilde{Y})\widetilde{R}}{S_0(\widetilde{Y}, \beta_0)^2}]]$$

$$- \mathbb{P}[\frac{\mathbf{1}(Y \leq t)RS_1(Y, \beta_0)}{S_0(Y, \beta_0)^2}]\sqrt{n}(\hat{\beta}_n - \beta_0)$$

$$+ o_p(1).$$

Outline

Introduction

- **Interval censoring:** event occurs within an interval
- Data: $(L_i, R_i, X_i), \quad i = 1, \ldots, n$
- **Transformation models:**

$$\Lambda(t; X) = G\left\{\int_0^t e^{\beta^T X} d\Lambda(s)\right\}$$

  - $G(\cdot)$: specific transformation function, strictly increasing
  - $\Lambda(\cdot)$: unknown increasing function
  - $G(x) = x \Rightarrow$ proportional hazards
  - $G(x) = \log(1 + x) \Rightarrow$ proportional odds

Introduction (cont.)

- Observed likelihood function (under PH model):

$$L_n(\beta,\Lambda) = \prod_{i=1}^{n} \left[\exp\left\{ -\int_0^{L_i} e^{\beta^{\mathrm{T}} X_i(s)} \mathrm{d}\Lambda(s) \right\} - \exp\left\{ -\int_0^{R_i} e^{\beta^{\mathrm{T}} X_i(s)} \mathrm{d}\Lambda(s) \right\} \right]$$

- Nonparametric maximum likelihood approach:

$$\begin{aligned} &\prod_{i=1}^{n} \left[\exp\left\{ -\sum_{t_k \le L_i} \lambda_k e^{\beta^{\mathrm{T}} X_i(t_k)} \right\} - \exp\left\{ -\sum_{t_k \le R_i} \lambda_k e^{\beta^{\mathrm{T}} X_i(t_k)} \right\} \right] \\ =& \prod_{i=1}^{n} \exp\left(-\sum_{t_k \le L_i} \lambda_k e^{\beta^{\mathrm{T}} X_{ik}} \right) \left\{ 1 - \exp\left(-\sum_{t_k \le R_i} \lambda_k e^{\beta^{\mathrm{T}} X_{ik}} \right) \right\}^{\mathbf{1}(R_i<\infty)} \end{aligned} \tag{2}$$

  - $t_1 < \cdots < t_m$: unique values of $L_i > 0$ and $R_i < \infty$
  - λ_k: jump size of Λ at t_k

Introduction (cont.)

- Direct maximization of (2) is difficult.
- Introduce latent independent Poisson random variables:

$$W_{ik} \sim \text{Poisson}(\lambda_k e^{\beta^T X_{ik}})$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, m$.

- (2) is equivalent to observing

$$\sum_{t_k \le L_i} W_{ik} = 0 \quad \text{and} \quad \mathbf{1}(R_i < \infty) \sum_{L_i < t_k \le R_i} W_{ik} > 0.$$

- EM algorithm treating W_{ik} as missing data.
  $\Rightarrow$ NPMLEs $(\hat{\beta}_n, \hat{\Lambda}_n)$

Asymptotic properties

Theorem (Consistency)

Under regularity conditions,

$$\|\hat{\beta}_n - \beta_0\| + \sup_{t \in [0,\tau]} |\hat{\Lambda}_n(t) - \hat{\Lambda}_0(t)| \xrightarrow{a.s.} 0.$$

Theorem (Asymptotic normality)

Under regularity conditions,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Sigma),$$

where Σ attains the semiparametric efficiency bound.

Proof of consistency

1. Define $m(\beta, \Lambda) = \log[\{L(\beta, \Lambda) + L(\beta_0, \Lambda_0)\}/2]$. Show that
$$\mathcal{M} := \left\{ m(\beta, \Lambda) : \ \beta \in \mathcal{B}, \ \Lambda \in BV[0, \tau] \right\}$$
is a Donsker class.
2. Show $\limsup_n \hat{\Lambda}_n(\tau) < \infty$, so that $m(\hat{\beta}_n, \hat{\Lambda}_n) \in \mathcal{M}$.
3. By Helly's selection lemma, for any subsequence of $(\hat{\beta}_n, \hat{\Lambda}_n)$, there exists a further subsequence such that $\hat{\beta}_n \to \beta^*$ and $\hat{\Lambda}_n \to \Lambda^*$ pointwise on $[0, \tau]$.
4. Construct a step function $\widetilde{\Lambda}$ that converges uniformly to Λ_0. Use the fact that $\mathbb{P}_n m(\hat{\beta}_n, \hat{\Lambda}_n) \geq \mathbb{P}_n m(\beta_0, \widetilde{\Lambda})$, together with the Donsker property of $\mathcal{M}$ to show $\mathbb{P} m(\beta^*, \Lambda^*) \geq \mathbb{P} m(\beta_0, \Lambda_0)$. Thus, by the properties of the Kullback-Leibler information, $L(\beta^*, \Lambda^*) = L(\beta_0, \Lambda_0)$.
5. Verify identifiability of the model. Then $\beta^* = \beta_0$ and $\Lambda^*(t) = \Lambda_0(t)$ for $t \in [0, \tau]$.
6. Pointwise convergence of $\hat{\Lambda}_n$ to Λ_0 can be strengthened to uniform convergence since Λ_0 is continuous.

Convergence rate

Lemma (Convergence rate)

Under regularity conditions,

$$E\left(\sum_{k=1}^{K}\left[\int_0^{U_k} e^{\hat{\beta}^{\mathrm{T}}X(s)}\mathrm{d}\hat{\Lambda}_n(s) - \int_0^{U_k} e^{\beta_0^{\mathrm{T}}X(s)}\mathrm{d}\Lambda_0(s)\right]^2\right)^{1/2} = O_p\left(n^{-1/3}\right),$$

where K is a random number of monitoring times, and $(U_1, \ldots, U_K)$ is a random sequence of monitoring times.

To prove the lemma, calculate the bracketing number for $\mathcal{M}$. Ideally,

$$\varphi(\delta) = \int_0^{\delta} \sqrt{1 + \log N_{[]}\left(\epsilon, \mathcal{M}, L_2(\mathbb{P})\right)}\,\mathrm{d}\epsilon \leq O(\delta^{1/2}).$$

Then check each condition in Theorem 3.4.1 of van der Vaart & Wellner and obtain the order of the Hellinger distance between $(\hat{\beta}_n, \hat{\Lambda}_n)$ and (β_0, Λ_0).

Proof of asymptotic normality of $\hat{\beta}_n$

1. To obtain the score operator for Λ, consider a one-dimensional submodel $\Lambda_{\epsilon,h}$ defined by $\mathrm{d}\Lambda_{\epsilon,h} = (1+\epsilon h)\,\mathrm{d}\Lambda$.
2. Consider the least favorable direction h^* such that the corresponding parametric submodel achieves the semiparametric efficient information.
3. Apply Taylor expansion at (β_0, Λ_0). By the previous lemma on convergence rate, we obtain
$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left(E[\{\ell_\beta - \ell_\Lambda(h^*)\}^{\otimes 2}]\right)^{-1} \mathbb{G}_n\{\ell_\beta(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_\Lambda(\hat{\beta}_n, \hat{\Lambda}_n)(h^*)\} + o_{\mathrm{p}}(1).$$
4. Show the existence of h^*. Need some functional theories.
5. Show that $\ell_\beta(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_\Lambda(\hat{\beta}_n, \hat{\Lambda}_n)(h^*)$ belongs to a Donsker class and converges in $L_2(\mathbb{P})$-norm to $\ell_\beta - \ell_\Lambda(h^*)$. This follows from the continuous differentiability of $h^*(t)$ over $[0, \tau]$.
6. Show the nonsingularity of the matrix $E[\{\ell_\beta - \ell_\Lambda(h^*)\}^{\otimes 2}]$. Prove by contradiction.
7. Finally, we have $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$ and
$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left(E[\{\ell_\beta - \ell_\Lambda(h^*)\}^{\otimes 2}]\right)^{-1} \mathbb{G}_n\{\ell_\beta - \ell_\Lambda(h^*)\} + o_{\mathrm{p}}(1).$$

Extension: multivariate interval-censored data

- Multiple types of events / clustering of study subjects
- Need to account for potential dependence
- Semiparametric transformation models with random effects (i—cluster, j—subject, k—event):

$$\Lambda_{ijk}(t) = G_k\left[\int_0^t \exp\left\{\beta^{\mathrm{T}} X_{ijk}(s) + b_i^{\mathrm{T}} Z_{ijk}(s)\right\} \mathrm{d}\Lambda_k(s)\right]$$

- Latent Poisson random variables + EM algorithm (treat random effects as missing data)
- Be careful with the random effects in the proofs. Everything else is similar to the univariate setting!