STAT6018 Research Frontiers in Data Science

Topic I: Statistical methods for analyzing complex survival data

Yu Gu, PhD
Assistant Professor

Department of Statistics & Actuarial Science
The University of Hong Kong

Table of Contents

Course Logistics

Course website: https://yugu-stat.github.io/teaching/stat6018

Lectures:

- Weeks 1–3
- Mainly discuss papers by Lin–Zeng's group
- Attendance is required

Final presentation:

- Week 4
- Presentation (15 min) + Q&A (5 min)
- Any statistical paper related to survival analysis
- Please send me the paper you want to present via email (yugu@hku.hk) for approval by Week 3.

Censored Data

Univariate survival data: time to the occurrence of a given event/failure

- Time to death
- Time to the occurrence of a disease

Multivariate survival data: times to several events/failures

- Recurrent events: repetitions of a phenomenon (e.g., illness)
  - Tumor recurrences
  - Infection episodes
- Multiple types of events: combination of multiple types of phenomena
  - Ordered events, such as HIV-infection $\rightarrow$ AIDS $\rightarrow$ death
  - Unordered events, such as diseases in several organ systems (cardiovascular disease, cancer, Alzheimer's disease, etc.)

Table of Contents

Reference

Zeng, D., & Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. Biometrika, 93(3), 627-640.

Counting processes

- Counting process is a continuous-time stochastic process $\{N(t) : t \geq 0\}$ with $N(0) = 0$, whose sample paths are step functions with jumps of size 1 only.
- In survival analysis without censoring, $N(t)$ records the number of events that have occurred by time t.
- For univariate survival data, $N(t)$ takes a single jump at the survival time.
- For recurrent events data, $N(t)$ takes jumps at all recurrent event times.

Intensity function

Notation:

- $N^*(t)$: counting process recording the number of events by time t
- $X(t)$: potentially time-dependent covariates
- $\mathcal{F}_t = \{N^*(s), X(s) : 0 \leq s \leq t\}$: history up to time t
- $dN^*(t)$: increment of N^* (i.e., number of events) over $[t, t + dt)$

Intensity function:

$$\lambda(t|X) = \lim_{dt \downarrow 0} \frac{1}{dt} E\{dN^*(t) \mid \mathcal{F}_{t-}\}$$

Cumulative intensity function:

$$\Lambda(t|X) = \int_0^t \lambda(s|X) ds$$

Proportional intensity model

Proportional intensity (PI) model:

$$\Lambda(t|X) = \int_0^t Y^*(s) \exp\left\{\beta^\mathsf{T} X(s)\right\} d\Lambda(s)$$

- $Y^*(t)$: indicator process
  - $Y^*(t) = I(T \geq t)$ for univariate survival data
  - $Y^*(t) \equiv 1$ for recurrent events data
- $\Lambda(t)$: unknown cumulative baseline intensity function
- β: unknown regression parameters

A large-sample theory for this model based on maximum partial likelihood estimation has been established via the counting-process martingale theory[1].

[1] *Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The Annals of Statistics, 1100-1120.*

Discussion about PI model

- For univariate survival data, the PI model reduces to the Cox proportional hazards (PH) model.

- The proportional hazards assumption may be violated in certain applications, especially in long-term studies.

- For example, the initial effect of a treatment may disappear with time, such that the hazard ratio converges to 1 as $t \to \infty$.

- A useful alternative is the proportional odds (PO) model[2]:

$$\frac{\Pr(T \le t|X)}{\Pr(T > t|X)} = g(t) \exp\left\{\beta^{\mathsf{T}} X(t)\right\},$$

  which constrains the hazard ratio to converge to 1 as $t \to \infty$.

[2] *Bennett, S. (1983). Analysis of survival data by the proportional odds model. Statistics in medicine, 2(2), 273-277.*

Semiparametric transformation models

The PH/PI and PO models belong to the broad class of semiparametric transformation models for general counting processes:

$$\Lambda(t|X) = G\left[\int_0^t Y^*(s)\exp\left\{\beta^{\mathsf{T}} X(s)\right\} d\Lambda(s)\right] \tag{1}$$

- $G(\cdot)$: strictly increasing transformation function
  - $G(x) = x \Rightarrow$ PH/PI model
  - $G(x) = \log(1+x) \Rightarrow$ PO model
- $\Lambda(t)$: arbitrary increasing function

Common choices of transformations

Box–Cox transformations:

$$G(x) = \rho^{-1}\left\{(1+x)^{\rho} - 1\right\} \quad (\rho \geq 0)$$

Logarithmic transformations:

$$G(x) = r^{-1}\log(1+rx) \quad (r \geq 0)$$

- $\rho = 1$ or $r = 0 \Rightarrow$ PH/PI model
- $\rho = 0$ or $r = 1 \Rightarrow$ PO model

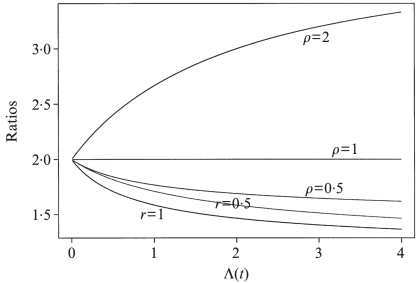

Figure 1: Plots of $\Lambda(t|X=x)/\Lambda(t|X=0)$ against $\Lambda(t)$ with $e^{\beta^{\mathsf{T}}x} = 2$

Censored counting processes

Notation:

- C: censoring time
- $N(t) = N^*(t \wedge C)$: counting process recording the number of events observed by time t
- $Y(t) = Y^*(t)I(C \geq t)$: at-risk indicator process
- τ: study end time

Independent censoring assumption: $N^*(t) \perp\!\!\!\perp C$ conditional on $X(t)$

Observed data from n random samples:

$$\left\{ N_i(t), Y_i(t), X_i(t) : t \in [0, \tau] \right\} \quad \text{for } i = 1, \ldots, n$$

Likelihood

Define $\lambda(t) = \Lambda'(t)$. Under model (1), the intensity function for $N_i(t)$ is

$$\lambda(t|X_i) = Y_i(t)e^{\beta^\top X_i(t)}\lambda(t)G'\left\{\int_0^t Y_i(s)e^{\beta^\top X_i(s)}d\Lambda(s)\right\}.$$

Thus, the likelihood function is

$$\begin{aligned} L_n(\beta,\Lambda) &= \prod_{i=1}^n \prod_{t\in[0,\tau]} \lambda(t|X_i)^{dN_i(t)} \exp\{-\Lambda(\tau|X_i)\} \\ &= \prod_{i=1}^n \prod_{t\in[0,\tau]} \left[e^{\beta^\top X_i(t)}\lambda(t)G'\left\{\int_0^t Y_i(s)e^{\beta^\top X_i(s)}d\Lambda(s)\right\}\right]^{dN_i(t)} \\ &\qquad\qquad \times \exp\left[-G\left\{\int_0^\tau Y_i(s)e^{\beta^\top X_i(s)}d\Lambda(s)\right\}\right]. \end{aligned}$$

Likelihood (cont.)

And the log-likelihood function is

$$
\begin{aligned}
\ell_n(\beta,\Lambda) = \sum_{i=1}^{n}\Bigg(& \int_0^\tau \left\{\beta^\mathsf{T} X_i(t) + \log\lambda(t)\right\} dN_i(t) \\
& + \int_0^\tau \log G'\left\{\int_0^t Y_i(s)e^{\beta^\mathsf{T} X_i(s)}d\Lambda(s)\right\} dN_i(t) \\
& - G\left\{\int_0^\tau Y_i(s)e^{\beta^\mathsf{T} X_i(s)}d\Lambda(s)\right\}\Bigg).
\end{aligned}
$$

We maximize the log-likelihood over β and Λ.

NPMLE

- We adopt the nonparametric maximum likelihood estimation (NPMLE) approach, where Λ is restricted to be a step function with non-negative jumps at all the observed event times, denoted by $t_1 < t_2 < \cdots < t_m$.
- The log-likelihood function under NPMLE becomes

$$\ell_n(\beta, \Lambda) = \sum_{i=1}^{n} \Bigg(\int_0^{\tau} \{\beta^{\mathsf{T}} X_i(t) + \log \Lambda\{t\}\}\, dN_i(t) + \int_0^{\tau} \log G' \Bigg\{ \sum_{k:t_k \le t} e^{\beta^{\mathsf{T}} X_i(t_k)} \Lambda\{t_k\} \Bigg\} dN_i(t) - G \Bigg\{ \sum_{k:t_k \le C_i} e^{\beta^{\mathsf{T}} X_i(t_k)} \Lambda\{t_k\} \Bigg\} \Bigg),$$

  where $\Lambda\{t\}$ denotes the jump size of Λ at time t.
- The estimators of β and $\Lambda\{t_k\}$ $(k = 1, \ldots, m)$ are obtained via the quasi-Newton method.

Variance estimation

To estimate the limiting covariance function of $\sqrt{n}(\widehat{\beta} - \beta_0, \widehat{\Lambda} - \Lambda_0)$, it suffices to obtain a variance estimator for the linear functional

$$\sqrt{n} \int_0^\tau w(t) d\{\widehat{\Lambda}(t) - \Lambda_0(t)\} + \sqrt{n} b^\mathsf{T}(\widehat{\beta} - \beta_0),$$

where $w(\cdot) \in \mathrm{BV}([0, \tau])$ and $b \in \mathbb{R}^p$.

We can treat β and $\Lambda\{t_k\}$'s as the parameters and estimate their limiting covariance matrix by the inverse of the observed information matrix $n\mathcal{I}_n$.

Since $\sqrt{n} \int_0^\tau w(t) d\{\widehat{\Lambda}(t) - \Lambda_0(t)\} + \sqrt{n} b^\mathsf{T}(\widehat{\beta} - \beta_0)$ is linear with all parameter estimates, its limiting variance V can be estimated by

$$\widehat{V} = \begin{pmatrix} W^\mathsf{T} & b^\mathsf{T} \end{pmatrix} \mathcal{I}_n^{-1} \begin{pmatrix} W \\ b \end{pmatrix},$$

where W is the vector of $w(\cdot)$ evaluated at all observed event times.

Asymptotic properties

Let $(\widehat{\beta}, \widehat{\Lambda})$ and (β_0, Λ_0) denote the nonparametric maximum likelihood estimates and the true values of (β, Λ), respectively. We have:

Consistency: $\|\widehat{\beta} - \beta_0\| + \sup_{t \in [0,\tau]} |\widehat{\Lambda} - \Lambda_0| \overset{a.s.}{\to} 0$.

Asymptotic normality: $\sqrt{n}(\widehat{\beta} - \beta_0, \widehat{\Lambda} - \Lambda_0)$ converges weakly to a mean-zero Gaussian process.

Semiparametric efficiency: The limiting covariance matrix of $\widehat{\beta}$ attains the semiparametric efficiency bound.

Consistency of variance estimators: $\widehat{V} \overset{a.s.}{\to} V$.

Table of Contents

Reference

Zeng, D., & Lin, D. Y. (2007). Semiparametric transformation models with random effects for recurrent events. Journal of the American Statistical Association, 102(477), 167-180.

Motivation

- Recall the proportional intensity model for recurrent events

$$\lambda(t|X) = \lambda(t) \exp\left\{\beta^{\mathsf{T}} X(t)\right\}$$

- Under the above model, the occurrence of an event is independent of any earlier events of the same subjects, which may not hold true in practice.

- For example, people who had a previous COVID-19 infection tend to have a lower risk of reinfection, while people who develop tumors more quickly than others tend to experience tumor recurrence more quickly.

- We could let $X(t)$ include the past event history, but this is not ideal since modeling the within-subject correlation through time-dependent covariates is very difficult.

PI model with frailty

- A useful approach to accommodating the dependence of the recurrent event times within the same subject is to incorporate a random effect (or frailty) into the model:

$$\lambda(t|X) = \xi\lambda(t) \exp\left\{\beta^{\mathsf{T}} X(t)\right\}$$

- The frailty ξ may capture the within-subject correlation and is usually assumed to follow the Gamma distribution.

- However, gamma frailty induces a very restrictive form of dependence.

Transformation models with random effects

We specify that the cumulative intensity function of $N^*(t)$ takes the form

$$\Lambda(t|X, Z, b) = G\left[\int_0^t \exp\left\{\beta^\mathsf{T} X(s) + b^\mathsf{T} Z(s)\right\} d\Lambda(s)\right]$$

- b: subject-specific random effects with mean 0 and density function $\phi(b; \gamma)$, used to capture the within-subject correlation
- $X(t)$ and $Z(t)$: potentially time-dependent covariates, may include covariates derived from the event history before time t
- b is usually assumed to follow a mean-zero multivariate normal distribution.

Recurrent events data

Observed data from n random samples:

$$\left\{N_i(t), Y_i(t), X_i(t), Z_i(t) : t \in [0, \tau]\right\} \quad \text{for } i = 1, \ldots, n$$

- $N_i(t) = N_i^*(t \wedge C_i)$
- $Y_i(t) = I(C_i \geq t)$

Independent censoring assumption: The conditional density of C at t given $\{N^*(s), X(s), Z(s) : s \in [0, \tau]\}$ and b depends only on $\{X(s), Z(s) : s \leq t\}$ and is noninformative about (β, γ, Λ).

Noninformative covariate processes assumption: The conditional distribution of $\{X(t), Z(t)\}$ given $\{N(s), Y(s), X(s), Z(s) : s < t\}$ is noninformative about (β, γ, Λ).

Likelihood and NPMLE

Let $\theta = (\beta^{\mathsf{T}}, \gamma^{\mathsf{T}})^{\mathsf{T}}$. The likelihood function under the preceding two assumptions is

$$L_n(\theta, \Lambda) = \prod_{i=1}^{n} \int_{b_i} \prod_{t \in [0,\tau]} \left[\lambda(t) e^{\beta^{\mathsf{T}} X_i(t) + b_i^{\mathsf{T}} Z_i(t)} G' \left\{ \int_0^t Y_i(s) e^{\beta^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} d\Lambda(s) \right\} \right]^{dN_i(t)}$$
$$\times \exp\left[-G \left\{ \int_0^\tau Y_i(s) e^{\beta^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} d\Lambda(s) \right\} \right] \phi(b_i; \gamma) db_i$$

NPMLE: Λ is treated as a step function with non-negative jumps at all the observed event times.

EM algorithm

The estimators can be computed via an EM algorithm, treating the random effects b_i as missing data.

The complete-data log-likelihood function is

$$\ell_c(\theta, \Lambda) = \sum_{i=1}^{n} \left(\int_0^\tau \left\{ \beta^\mathsf{T} X_i(t) + b_i^\mathsf{T} Z_i(t) + \log \Lambda\{t\} \right\} dN_i(t) \right.$$
$$+ \int_0^\tau \log G' \left\{ \int_0^t Y_i(s) e^{\beta^\mathsf{T} X_i(s) + b_i^\mathsf{T} Z_i(s)} d\Lambda(s) \right\} dN_i(t)$$
$$\left. - G \left\{ \int_0^\tau Y_i(s) e^{\beta^\mathsf{T} X_i(s) + b_i^\mathsf{T} Z_i(s)} d\Lambda(s) \right\} + \log \phi(b_i; \gamma) \right).$$

E-step

Let $\widehat{E}(\cdot)$ denote the conditional expectation given the observed data.

In the E-step, we compute $\widehat{E}\{H(b_i)\}$ for some function $H(\cdot)$ based on the posterior density of b_i, which is proportional to

$$\prod_{i=1}^{n} \prod_{t \in [0,\tau]} \left[\lambda(t) e^{\beta^{\mathsf{T}} X_i(t) + b_i^{\mathsf{T}} Z_i(t)} G' \left\{ \int_0^t Y_i(s) e^{\beta^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} d\Lambda(s) \right\} \right]^{dN_i(t)}$$
$$\times \exp\left[-G \left\{ \int_0^{\tau} Y_i(s) e^{\beta^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} d\Lambda(s) \right\} \right] \phi(b_i; \gamma)$$

The integral over b_i in $\widehat{E}\{H(b_i)\}$ can be approximated by Gauss–Hermite quadrature.

M-step

In the M-step, we maximize the objective function

$$M(\theta, \Lambda) = \sum_{i=1}^{n} \Bigg(\int_0^{\tau} \left\{ \beta^{\mathsf{T}} X_i(t) + \log \Lambda\{t\} \right\} dN_i(t) \\ + \int_0^{\tau} \widehat{E} \left[b_i^{\mathsf{T}} Z_i(t) + \log G' \left\{ \int_0^{t} Y_i(s) e^{\beta^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} d\Lambda(s) \right\} \right] dN_i(t) \\ - \widehat{E} \left[G \left\{ \int_0^{\tau} Y_i(s) e^{\beta^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} d\Lambda(s) \right\} \right] + \widehat{E} \left\{ \log \phi(b_i; \gamma) \right\} \Bigg).$$

We update γ by maximizing $\sum_{i=1}^{n} \widehat{E} \left\{ \log \phi(b_i; \gamma) \right\}$.

M-step (cont.)

To update β and Λ, define $F(t) = \Lambda(t)/\Lambda(\tau)$. We expand β to $[\log \Lambda(\tau), \beta]$ and expand $X_i(t)$ to $[1, X_i(t)]$. For simplicity, we still denote the expanded terms by β and $X_i(t)$.

Then the objective function to be maximized is equivalent to

$$\begin{aligned}\widetilde{M}(\beta, F) = \sum_{i=1}^{n} \Bigg(& \int_0^\tau \left\{\beta^\mathsf{T} X_i(t) + \log F\{t\}\right\} dN_i(t) \\ & + \int_0^\tau \widehat{E}\left[b_i^\mathsf{T} Z_i(t) + \log G'\left\{\int_0^t Y_i(s) e^{\beta^\mathsf{T} X_i(s) + b_i^\mathsf{T} Z_i(s)} dF(s)\right\}\right] dN_i(t) \\ & - \widehat{E}\left[G\left\{\int_0^\tau Y_i(s) e^{\beta^\mathsf{T} X_i(s) + b_i^\mathsf{T} Z_i(s)} dF(s)\right\}\right]\Bigg),\end{aligned}$$

with the constraint that $\sum_{i=1}^n \int_0^\tau F\{t\} dN_i(t) = 1$ (by NPMLE).

M-step (cont.)

Notation:

- T_{ij}: jth event time of the ith subject ($i = 1, \ldots, n$ and $j = 1, \ldots, n_i$)
- $t_1 < t_2 < \cdots < t_m$: sorted sequence of all distinct values of T_{ij}
- $f_k = F\{t_k\}$, for $k = 1, \ldots, m$
- μ: Lagrange multiplier

The objective function can be written as

$$\begin{aligned}\widetilde{M}(\beta, F) = \sum_{k=1}^{m} \log(f_k) + \sum_{i=1}^{n} \Bigg(& \sum_{j=1}^{n_i} \beta^{\mathsf{T}} X_i(T_{ij}) \\ & + \sum_{j=1}^{n_i} \widehat{E} \left[b_i^{\mathsf{T}} Z_i(T_{ij}) + \log G' \left\{ \sum_{k: t_k \le T_{ij}} e^{\beta^{\mathsf{T}} X_i(t_k) + b_i^{\mathsf{T}} Z_i(t_k)} f_k \right\} \right] \\ & - \widehat{E} \left[G \left\{ \sum_{k: t_k \le C_i} e^{\beta^{\mathsf{T}} X_i(t_k) + b_i^{\mathsf{T}} Z_i(t_k)} f_k \right\} \right] \Bigg) - \mu \left(\sum_{k=1}^{m} f_k - 1 \right).\end{aligned}$$

M-step (cont.)

We then solve the score equations for β and $(f_1, \dots, f_m)$:

$$0 = \sum_{i=1}^{n} \Bigg(\sum_{j=1}^{n_i} X_i(T_{ij}) + \sum_{j=1}^{n_i} \widehat{E} \left[\frac{G''\{\sum_{k:t_k \le T_{ij}} e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} f_k\}}{G'\{\sum_{k:t_k \le T_{ij}} e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} f_k\}} \times \sum_{k:t_k \le T_{ij}} e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} X_i(t_k) f_k \right] - \widehat{E} \left[G' \left\{ \sum_{k:t_k \le C_i} e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} f_k \right\} \times \sum_{k:t_k \le C_i} e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} X_i(t_k) f_k \right] \Bigg).$$

and

$$\mu = \frac{1}{f_k} + \sum_{i=1}^{n} \Bigg(\sum_{j=1}^{n_i} \widehat{E} \left[\frac{G''\{\sum_{l:t_l \le T_{ij}} e^{\beta^\mathsf{T} X_i(t_l) + b_i^\mathsf{T} Z_i(t_l)} f_l\}}{G'\{\sum_{l:t_l \le T_{ij}} e^{\beta^\mathsf{T} X_i(t_l) + b_i^\mathsf{T} Z_i(t_l)} f_l\}} \times I(t_k \le T_{ij}) e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} \right] - \widehat{E} \left[G' \left\{ \sum_{l:t_l \le C_i} e^{\beta^\mathsf{T} X_i(t_l) + b_i^\mathsf{T} Z_i(t_l)} f_l \right\} \times I(t_k \le C_i) e^{\beta^\mathsf{T} X_i(t_k) + b_i^\mathsf{T} Z_i(t_k)} \right] \Bigg)$$

Recursive formula

When $X(t)$ and $Z(t)$ are both time-independent, it is easy to observe that the second equation provides a recursive formula for calculating $(f_1, \ldots, f_m)$:

$$\frac{1}{f_{k+1}} = \frac{1}{f_k} + \sum_{i=1}^{n} \left(\sum_{j=1}^{n_i} \widehat{E} \left[\frac{G''\{e^{\beta^{\mathsf{T}} X_i + b_i^{\mathsf{T}} Z_i} F(t_k)\}}{G'\{e^{\beta^{\mathsf{T}} X_i + b_i^{\mathsf{T}} Z_i} F(t_k)\}} \times I(T_{ij} = t_k) e^{\beta^{\mathsf{T}} X_i + b_i^{\mathsf{T}} Z_i} \right] \right.$$
$$\left. - \widehat{E} \left[G' \left\{ e^{\beta^{\mathsf{T}} X_i + b_i^{\mathsf{T}} Z_i} F(t_k) \right\} \times I(t_k \leq C_i < t_{k+1}) e^{\beta^{\mathsf{T}} X_i + b_i^{\mathsf{T}} Z_i} \right] \right)$$

Write f_k as $f_k(f_1, \beta)$. We can solve (f_1, β) via the Newton-Raphson method, where the derivatives of f_k w.r.t. f_1 and β are calculated based on the above recursive formula, with initial values $\partial f_1 / \partial f_1 = 1$ and $\partial f_1 / \partial \beta = 0$.

This addresses the issue of high-dimensional parameters in NPMLE.

Variance estimation

As in the previous paper, the limiting variances of $(\widehat{\beta}, \widehat{\Lambda})$ can be consistently estimated by the inverse of the observed information matrix $n\mathcal{I}_n$.

By Louis' formula[3], $n\mathcal{I}_n$ can be calculated within the EM algorithm by

$$-\sum_{i=1}^{n} \widehat{E}\left\{\nabla^2 \ell_i(b_i;\theta,\Lambda)\right\} - \sum_{i=1}^{n}\left[\widehat{E}\left\{\nabla \ell_i(b_i;\theta,\Lambda)^{\otimes 2}\right\} - \widehat{E}\left\{\nabla \ell_i(b_i;\theta,\Lambda)\right\}^{\otimes 2}\right],$$

where ℓ_i is the ith subject's contribution to the complete-data log-likelihood function, and $\nabla \ell_i$ denotes the gradient of ℓ_i w.r.t. β and $\Lambda\{t_k\}$'s.

[3] *Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B: Statistical Methodology, 44(2), 226-233.*

Asymptotic properties under known G

Let $(\widehat{\theta}, \widehat{\Lambda})$ and (θ_0, Λ_0) denote the nonparametric maximum likelihood estimates and the true values of (θ, Λ), respectively.

When the transformation $G(\cdot)$ is completely specified, we have:

Consistency: $\|\widehat{\theta} - \theta_0\| + \sup_{t \in [0,\tau]} |\widehat{\Lambda} - \Lambda_0| \xrightarrow{a.s.} 0$.

Asymptotic normality: $\sqrt{n}(\widehat{\theta} - \theta_0, \widehat{\Lambda} - \Lambda_0)$ converges weakly to a mean-zero Gaussian process.

Semiparametric efficiency: The limiting covariance matrix of $\widehat{\theta}$ attains the semiparametric efficiency bound.

Asymptotic properties under unknown G

- When the transformation $G(\cdot)$ belongs to a one-parameter family $\{G_\eta : \eta \in (a_0, b_0)\}$, η is another unknown parameter.
- Write $\theta = (\beta^\mathsf{T}, \gamma^\mathsf{T}, \eta)^\mathsf{T}$. With some additional conditions, all the asymptotic properties on the previous slide still hold.
  - Linear independence of covariates at time 0
  - Smoothness conditions for G_η w.r.t. η
- The Box–Cox and logarithmic transformations introduced before satisfy those additional conditions, so their parameters (ρ or r) can also be estimated from the data.

Table of Contents

Reference

Zeng, D., & Lin, D. Y. (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. Biometrics, 65(3), 746-752.

Motivation

- In practice, recurrent event times are subject to censoring. Most of the existing methods require independent censoring.
- This is OK if censoring is caused by the end of the study or random loss to follow-up.
- In many medical studies, however, recurrent events may be terminated by the subject's withdrawal from the study due to deteriorating health or the subject's death.
- In those cases, the censoring time is likely correlated with the recurrent event times, and existing methods may yield misleading results.
- To address the dependent censoring issue, we consider joint analysis of recurrent and terminal evnets through shared random effects models.

Joint transformation models

Submodel for recurrent event process $N^*(t)$:

$$\Lambda_R(t|X,Z,b) = H\left[\int_0^t \exp\left\{\alpha^{\mathsf{T}} X(s) + b^{\mathsf{T}} Z(s)\right\} dA(s)\right]$$

Submodel for terminal event time T:

$$\Lambda_T(t|X,Z,b) = G\left[\int_0^t \exp\left\{\beta^{\mathsf{T}} X(s) + b^{\mathsf{T}}\left(\gamma \circ Z(s)\right)\right\} d\Lambda(s)\right]$$

- $H(\cdot)$ and $G(\cdot)$: transformation functions
- α, β, and γ: unknown regression parameters
- $X(t)$ and $Z(t)$: potentially time-dependent covariates, $Z(t)$ contains 1
- $\gamma \circ Z(s)$: component-wise product of γ and $Z(s)$
- b: shared random effects, with mean 0 and density function $\phi(b;\eta)$

Joint transformation models (cont.)

Submodel for recurrent event process $N^*(t)$:

$$\Lambda_R(t|X,Z,b) = H\left[\int_0^t \exp\left\{\alpha^\mathsf{T} X(s) + b^\mathsf{T} Z(s)\right\} dA(s)\right]$$

Submodel for terminal event time T:

$$\Lambda_T(t|X,Z,b) = G\left[\int_0^t \exp\left\{\beta^\mathsf{T} X(s) + b^\mathsf{T}\left(\gamma \circ Z(s)\right)\right\} d\Lambda(s)\right]$$

- The variance of b characterizes the dependence among recurrent event times.
- γ characterizes the dependence between recurrent and terminal events attributed to the unobserved random effects. $\gamma = 0$ implies that the dependence can be fully explained by the covariates.

Data and assumption

Data: $\{Y_i, \Delta_i, N_i^*(t), X_i(t), Z_i(t) : t \leq Y_i\}$ $(i = 1, \ldots, n)$

- $Y_i = \min(T_i, C_i)$
- $\Delta_i = I(T_i \leq C_i)$
- C_i: censoring time

Independent censoring assumption: $C_i \perp\!\!\!\perp (N_i^*, T_i, b_i)$ conditional on the covariates X_i and Z_i

Conditional independence: $N_i^* \perp\!\!\!\perp T_i$ conditional on b_i, X_i, and Z_i

Likelihood

Let $a(t) = A'(t)$, $\lambda(t) = \Lambda'(t)$, and $R_i(t) = I(Y_i \geq t)$. The observed-data likelihood function concerning $(\alpha, \beta, \gamma, \eta, A, \Lambda)$ is

$$
\begin{aligned}
\prod_{i=1}^{n} \int_{b_i} \Bigg[\prod_{t} \Bigg\{ a(t) e^{\alpha^{\mathsf{T}} X_i(t) + b_i^{\mathsf{T}} Z_i(t)} H' \left(\int_0^t e^{\alpha^{\mathsf{T}} X_i(s) + b_i^{\mathsf{T}} Z_i(s)} dA(s) \right) \Bigg\}^{R_i(t) dN_i^*(t)} \\
\times \exp \left\{ -H \left(\int_0^{Y_i} e^{\alpha^{\mathsf{T}} X_i(t) + b_i^{\mathsf{T}} Z_i(t)} dA(t) \right) \right\} \Bigg] \\
\times \Bigg[\Bigg\{ \lambda\left(Y_i\right) e^{\beta^{\mathsf{T}} X_i(Y_i) + b_i^{\mathsf{T}} (\gamma \circ Z_i(Y_i))} G' \left(\int_0^{Y_i} e^{\beta^{\mathsf{T}} X_i(t) + b_i^{\mathsf{T}} (\gamma \circ Z_i(t))} d\Lambda(t) \right) \Bigg\}^{\Delta_i} \\
\times \exp \left\{ -G \left(\int_0^{Y_i} e^{\beta^{\mathsf{T}} X_i(t) + b_i^{\mathsf{T}} (\gamma \circ Z_i(t))} d\Lambda(t) \right) \right\} \Bigg] \phi(b_i; \eta) db_i
\end{aligned}
$$

NPMLE

We consider A as a step function with jumps only at the observed recurrent event times, and consider Λ as a step function with jumps only at the observed terminal event times.

Thus, we maximize the following modified log-likelihood function over $(\alpha, \beta, \gamma, \eta)$ and the jump sizes of A and Λ:

$$\sum_{i=1}^{n} \log \int_{b_i} \left[\prod_t \left\{ A\{t\} e^{\alpha^\top X_i(t) + b_i^\top Z_i(t)} H' \left(\int_0^t e^{\alpha^\top X_i(s) + b_i^\top Z_i(s)} dA(s) \right) \right\}^{R_i(t) dN_i^*(t)} \right.$$
$$\left. \times \exp\left\{ -H\left(\int_0^{Y_i} e^{\alpha^\top X_i(t) + b_i^\top Z_i(t)} dA(t) \right) \right\} \right]$$
$$\times \left[\left\{ \Lambda\left\{Y_i\right\} e^{\beta^\top X_i(Y_i) + b_i^\top (\gamma \circ Z_i(Y_i))} G' \left(\int_0^{Y_i} e^{\beta^\top X_i(t) + b_i^\top (\gamma \circ Z_i(t))} d\Lambda(t) \right) \right\}^{\Delta_i} \right.$$
$$\left. \times \exp\left\{ -G\left(\int_0^{Y_i} e^{\beta^\top X_i(t) + b_i^\top (\gamma \circ Z_i(t))} d\Lambda(t) \right) \right\} \right] \phi(b_i; \eta) db_i$$

Computing algorithm

- We may use quasi-Newton or other optimization algorithms to obtain the NPMLEs.
- Alternatively, we can use an EM algorithm for computation, with the subject-specific random effects b_i treated as missing data.
- In the M-step, the maximization is taken over only a small set of parameters, thanks to some recursive formulae among the jump sizes of A and Λ.

Asymptotic properties

Let $\theta = (\alpha^\mathsf{T}, \beta^\mathsf{T}, \gamma^\mathsf{T}, \eta^\mathsf{T})^\mathsf{T}$ denote the set of all finite-dimensional parameters. We have:

Consistency: $\|\widehat{\theta} - \theta_0\| + \sup_{t\in[0,\tau]} |\widehat{A} - A_0| + \sup_{t\in[0,\tau]} |\widehat{\Lambda} - \Lambda_0| \overset{a.s.}{\to} 0.$

Asymptotic normality: $\sqrt{n}(\widehat{\theta} - \theta_0, \widehat{A} - A_0, \widehat{\Lambda} - \Lambda_0)$ converges weakly to a mean-zero Gaussian process.

Semiparametric efficiency: The limiting covariance matrix of $\widehat{\theta}$ attains the semiparametric efficiency bound.

The limiting variances and covariances can be consistently estimated by inverting the observed information matrix for all parameters, including θ and the jump sizes of A and Λ. The observed information matrix can be calculated by Louis' formula.

Table of Contents

Reference

Zeng, D., Chen, Q., & Ibrahim, J. G. (2009). Gamma frailty transformation models for multivariate survival times. Biometrika, 96(2), 277-291.

Multivariate failure time data

- Multivariate failure time data arise when each study subject can experience several events.
- It is interesting to determine risk factors that are predictive for some or all of the failures.
- For example, in COVID-19 vaccine trials, investigators want to access the efficacy of a vaccine against infection, hospitalization, and death.
- Like recurrent events data, multivariate failure times from the same subject are potentially correlated. Ignoring such correlation may lead to biased inference.

Gamma frailty transformation models

Let T_k denote the failure time of the kth event type $(k = 1, \ldots, K)$. We specify the following gamma frailty transformation model:

$$\Lambda_k(t|X, \xi) = \xi G_k \left\{ \Lambda_k(t) e^{\beta_k^\top X} \right\} \tag{2}$$

- $\xi \sim \text{Gamma}(\gamma^{-1}, \gamma)$: captures the within-subject correlation
- $G_k(\cdot)$: type-specific transformation function
- $\Lambda_k(t)$: unspecified type-specific increasing function
- β_k: type-specific regression parameters

Gamma frailty transformation models (cont.)

- Under model (2), the marginal cumulative hazard function for T_k is

$$\Lambda_{T_k}(t) = \gamma^{-1} \log \left[1 + \gamma G_k \left\{\Lambda_k(t) e^{\beta_k^{\mathsf{T}} X}\right\}\right]$$

- The above marginal distribution is equivalent to another linear transformation model:

$$\log \Lambda_k(T_k) = -\beta_k^{\mathsf{T}} X + \epsilon_k,$$

  with ϵ_k following the distribution $\log G_k^{-1}[\gamma^{-1}\{\mathrm{Unif}(0,1)^{-\gamma} - 1\}]$.

- The dependence among failure times can be evaluated through γ. We allow $\gamma = 0$, which corresponds to the scenario with independent failure times.

Reparameterization

Let τ denote the study end time. We define $F_k(t) = \Lambda_k(t)/\Lambda_k(\tau)$ and $\alpha_k = \log \Lambda_k(\tau)$. Model (2) can be rewritten as

$$\Lambda_k(t|X,\xi) = \xi G_k \left\{ F_k(t) e^{\alpha_k + \beta_k^{\mathsf{T}} X} \right\} \tag{3}$$

Clearly, $F_k(\cdot)$ is a distribution function in $[0,\tau]$, with $F_k(0) = 0$ and $F_k(\tau) = 1$.

Under some mild conditions on the true parameter values, the transformation functions, and the censoring distributions, all the parameters, including (α_k, β_k, F_k) $(k = 1, \ldots, K)$ and γ, are identifiable.

Data and likelihood

Data: $\{Y_{ik}, \Delta_{ik}, X_i : i = 1, \ldots, n \text{ and } k = 1, \ldots, K\}$

- $Y_{ik} = \min(T_{ik}, C_{ik})$
- $\Delta_{ik} = I(T_{ik} \leq C_{ik})$
- C_{ik}: censoring time for the kth event type of the ith subject

Independent censoring assumption: $C_{ik} \perp\!\!\!\perp (T_{ik}, \xi_i)$ given X_i

Likelihood function:

$$L_n(\alpha, \beta, \gamma, F) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[G_k'\{F_k(Y_{ik}) e^{\alpha_k + \beta_k^{\mathsf{T}} X_i}\} F_k'(Y_{ik}) e^{\alpha_k + \beta_k^{\mathsf{T}} X_i} \right]^{\Delta_{ik}}$$

$$\times \int_{\xi_i} \xi_i^{\sum_{k=1}^{K} \Delta_{ik}} \exp\left[-\xi_i \sum_{k=1}^{K} G_k\{F_k(Y_{ik}) e^{\alpha_k + \beta_k^{\mathsf{T}} X_i}\} \right] g(\xi_i; \gamma) d\xi_i,$$

where $g(\xi; \gamma)$ is the density of Gamma(γ^{-1}, γ).

NPMLE

We treat F_k as a discrete distribution function with positive jumps at all Y_{ik} with $\Delta_{ik} = 1$.

Then the log-likelihood function is

$$\ell_n(\alpha, \beta, \gamma, F) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left[\log G_k' \left\{ F_k \left(Y_{ik} \right) e^{\alpha_k + \beta_k^{\mathsf{T}} X_i} \right\} + \log F_k \left\{ Y_{ik} \right\} + \alpha_k + \beta_k^{\mathsf{T}} X_i \right]$$
$$+ \sum_{i=1}^{n} \log \int_{\xi_i} \xi_i^{\sum_{k=1}^{K} \Delta_{ik}} \exp \left(-\xi_i \left[\sum_{k=1}^{K} G_k \left\{ F_k \left(Y_{ik} \right) e^{\alpha_k + \beta_k^{\mathsf{T}} X_i} \right\} \right] \right) g(\xi_i; \gamma) d\xi_i$$

We maximize the log-likelihood over α_k, β_k, γ, and the jump sizes of F_k, under the constraint that the sum of all jumps of F_k equals 1.

EM algorithm

The maximization can be solved via an EM algorithm, with gamma frailties ξ_i treated as missing data.

The complete-data log-likelihood function is

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\left[\log G_k'\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\}+\log F_k\left\{Y_{ik}\right\}+\alpha_k+\beta_k^{\mathsf{T}}X_i+\log\xi_i\right]$$
$$-\sum_{i=1}^{n}\xi_i\sum_{k=1}^{K}G_k\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\}+\sum_{i=1}^{n}\log g(\xi_i;\gamma)$$

E-step

In the E-step, we evaluate the conditional expectation of some function $H(\xi_i)$ given the observed data.

The conditional density of ξ_i given the observed data is proportional to

$$
\begin{aligned}
&\xi_i^{\sum_{k=1}^{K} \Delta_{ik}} \exp\left[-\xi_i \sum_{k=1}^{K} G_k\{F_k(Y_{ik}) e^{\alpha_k + \beta_k^\top X_i}\}\right] g(\xi_i; \gamma) \\
&\sim \text{Gamma}\left(\gamma^{-1} + \sum_{k=1}^{K} \Delta_{ik}, \left[\gamma^{-1} + \sum_{k=1}^{K} G_k\{F_k(Y_{ik}) e^{\alpha_k + \beta_k^\top X_i}\}\right]^{-1}\right)
\end{aligned}
$$

The integral over ξ_i can be calculated analytically or by a Laplace approximation.

M-step

Notation:

- $t_{1k} < t_{2k} < \cdots < t_{m_k,k}$: sorted sequence of all Y_{ik} with $\Delta_{ik} = 1$
- $f_{lk} = F_k\{t_{lk}\}$, for $k = 1, \ldots, K$ and $l = 1, \ldots, m_k$

In the M-step, we maximize the following objective function:

$$\begin{aligned} M(\alpha, \beta, \gamma, F) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \Bigg[\log G_k' \Bigg\{ \sum_{l:t_{lk} \le Y_{ik}} f_{lk} \left(Y_{ik} \right) e^{\alpha_k + \beta_k^{\mathrm{T}} X_i} \Bigg\} + \log \sum_{l:t_{lk} \le Y_{ik}} f_{lk} \\ + \alpha_k + \beta_k^{\mathrm{T}} X_i + \widehat{E} \left(\log \xi_i \right) \Bigg] - \sum_{i=1}^{n} \widehat{E} \left(\xi_i \right) \sum_{k=1}^{K} G_k \Bigg\{ \sum_{l:t_{lk} \le Y_{ik}} f_{lk} e^{\alpha_k + \beta_k^{\mathrm{T}} X_i} \Bigg\} \\ - n \log \gamma^{1/\gamma} \Gamma(\gamma^{-1}) + (\gamma^{-1} - 1) \sum_{i=1}^{n} \widehat{E} \left(\log \xi_i \right) - \gamma^{-1} \sum_{i=1}^{n} \widehat{E} \left(\xi_i \right) \end{aligned}$$

under the constraint $\sum_{l=1}^{m_k} f_{lk} = 1$, for $k = 1, \ldots, K$.

M-step (cont.)

The score equation for f_{lk} is

$$\frac{1}{f_{lk}} = -\sum_{i=1}^{n} I(Y_{ik} \geq t_{lk})\Delta_{ik} \frac{G_k''\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\}}{G_k'\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\}} e^{\alpha_k+\beta_k^{\mathsf{T}}X_i} + \sum_{i=1}^{n} I(Y_{ik} \geq t_{lk})\widehat{E}\left(\xi_i\right) G_k'\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\} e^{\alpha_k+\beta_k^{\mathsf{T}}X_i} + \mu_k,$$

where μ_k is the Lagrange multiplier.

This yields a recursive formula

$$\frac{1}{f_{l+1,k}} = \frac{1}{f_{lk}} + \sum_{i=1}^{n} I(t_{lk} \leq Y_{ik} < t_{l+1,k})\Delta_{ik} \frac{G_k''\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\}}{G_k'\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\}} e^{\alpha_k+\beta_k^{\mathsf{T}}X_i} - \sum_{i=1}^{n} I(t_{lk} \leq Y_{ik} < t_{l+1,k})\widehat{E}\left(\xi_i\right) G_k'\left\{F_k\left(Y_{ik}\right)e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}\right\} e^{\alpha_k+\beta_k^{\mathsf{T}}X_i}$$

M-step (cont.)

Similar to the previous paper, we can then treat $(\alpha_k, \beta_k, f_{1k})$ $(k = 1, \ldots, K)$ and γ as the parameters to be updated in the M-step, since all other f_{lk} can be expressed as a function of these parameters.

This way, the maximization is carried out over only a small set of parameters, such that the EM algorithm is immune to the high-dimensional parameters in NPMLE.

M-step (cont.)

We can update $(\alpha_k, \beta_k, f_{1k})$ $(k = 1, \ldots, K)$ and γ via the one-step Newton-Raphson method. The equations to be solved are

$$0 = \sum_{i=1}^{n} \Delta_{ik} \left[\frac{G_k''\left\{F_k\left(Y_{ik}\right) e^{\alpha_k + \beta_k^{\mathrm{T}} X_i}\right\}}{G_k'\left\{F_k\left(Y_{ik}\right) e^{\alpha_k + \beta_k^{\mathrm{T}} X_i}\right\}} F_k\left(Y_{ik}\right) e^{\alpha_k + \beta_k^{\mathrm{T}} X_i} + 1 \right] \left(1, X_i^{\mathrm{T}}\right)^{\mathrm{T}}$$

$$- \sum_{i=1}^{n} \widehat{E}\left(\xi_i\right) G_k'\left\{F_k\left(Y_{ik}\right) e^{\alpha_k + \beta_k^{\mathrm{T}} X_i}\right\} F_k\left(Y_{ik}\right) e^{\alpha_k + \beta_k^{\mathrm{T}} X_i} \left(1, X_i^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\sum_{l=1}^{m_k} f_{lk} = 1,$$

for $k = 1, \ldots, K$, and

$$\frac{n}{\gamma^2} \log \gamma - \frac{n}{\gamma^2} + n \frac{\Gamma'(\gamma^{-1})}{\gamma^2 \Gamma(\gamma^{-1})} - \frac{1}{\gamma^2} \sum_{i=1}^{n} \widehat{E}\left(\log \xi_i\right) + \frac{1}{\gamma^2} \sum_{i=1}^{n} \widehat{E}\left(\xi_i\right) = 0.$$

Note that f_{lk} is now a function of $(\alpha_k, \beta_k, f_{1k})$, and the derivatives can be calculated based on the recursive formula.

Boundary issue

- One limitation of this EM algorithm is that the estimate of γ must be positive.
- However, when $\gamma = 0$ (i.e., no correlation among all event types), the MLE of γ can be 0 or even negative. The EM algorithm is not applicable due to an improper density of ξ_i.
- In that case, we estimate the other parameters using the same EM algorithm while fixing $\gamma = 0$ and $\widehat{E}(\xi_i) = 1$.
- We then compare the observed-data likelihoods with and without the constraint $\gamma = 0$. The estimates with a larger observed-data likelihood will be treated as the final estimates.

Asymptotic properties

Consistency:

$$\sum_{k=1}^{K}\left(|\widehat{\alpha}_k - \alpha_{0k}| + |\widehat{\beta}_k - \beta_{0k}|\right) + |\widehat{\gamma} - \gamma_0| + \sum_{k=1}^{K}\sup_{t\in[0,\tau]}|\widehat{F}_k - F_{0k}| \stackrel{a.s.}{\to} 0$$

Asymptotic normality: $\sqrt{n}(\widehat{\beta}_k - \beta_{0k}, \widehat{\gamma} - \gamma_0, \widehat{\Lambda}_k - \Lambda_{0k})_{k=1,\ldots,K}$ converges weakly to a mean-zero Gaussian process.

Semiparametric efficiency: The limiting covariances of $\widehat{\beta}_k$ $(k = 1, \ldots, K)$ and $\widehat{\gamma}$ attains the semiparametric efficiency bound.

The limiting covariance for $(\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{F}_k)$ $(k = 1, \ldots, K)$ and $\widehat{\gamma}$ can be consistently estimated based on the inverse of the observed information matrix (treating the jump sizes of F_k as usual parameters) and the delta method.

Concluding remarks

- All these papers are rediscussed in Zeng & Lin (2007)[4]. Their likelihood functions can be written in a generic form

$$L_n(\theta, \mathcal{A}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{l=1}^{n_{ik}} \prod_{t \leqslant \tau} \lambda_k(t)^{\mathrm{d}N_{ikl}(t)} \Psi\left(\mathcal{O}_i; \theta, \mathcal{A}\right)$$

- A general asymptotic theory has been established in Zeng & Lin (2010)[5].

- To prove the asymptotic properties for each specific problem, we only need to check the regularity conditions of the general theory.

[4] *Zeng, D., & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 69(4), 507-564*

[5] *Zeng, D., & Lin, D. Y. (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. Statistica Sinica, 20(2), 871.*