Nonparametric Estimation with Recurrent Event Data

Edsel A. Pena Department of Statistics University of South Carolina E-mail: pena@stat.sc.edu

Talk at UNCC, 10/12/01

Based on joint works with R. Strawderman (Cornell) and M. Hollander (Florida State)

Research supported by NIH and NSF Grants

A Real Recurrent Event Data (Source: Aalen and Husebye ('91), Statistics in Medicine)

Variable: Migrating motor complex (MMC) periods, in minutes, for 19 individuals in a study concerning small bowel motility during fasting state.

Unit #	#Complete	Complete Observed Successive	Censored
i	(K _i =K(i))	Periods (T _{ij})	$(\tau_i - S_{iK(i)})$
1	8	112 145 39 52 21 34 33 51	54
2	2	206 147	30
3	3	284 59 186	4
4	3	94 98 84	87
5	1	67	131
6	9	124 34 87 75 43 38 58 142 75	23
7	5	116 71 83 68 125	111
8	4	111 59 47 95	110
9	4	98 161 154 55	44
10	2	166 56	122
11	5	63 90 63 103 51	85
12	4	47 86 68 144	72
13	3	120 106 176	6
14	4	112 25 57 166	85
15	3	132 267 89	86
16	5	120 47 165 64 113	12
17	4	162 141 107 69	39
18	6	106 56 158 41 41 168	13
19	5	147 134 78 66 100	4

Pictorial Representation of Data for a Unit or Subject

- Consider unit/subject #3.
- K = 3
- Gap Times, T_i: 284, 343, 529
- Censored Time, τ -S_K: 4
- Calendar Times, S_i: 284, 343, 529
- Limit of Obs. Period: $\tau = 533$



Features of Data Set

- Random observation period per subject (administrative constraints).
- Length of period: τ
- Event of interest is recurrent. A subject may have more than one event during observation period.
- # of events (K) informative about MMC period distribution (F).
- Last MMC period right-censored by a variable informative about F.
- Calendar times: $S_1, S_2, ..., S_K$.
- Right-censoring variable: τ -S_K.

Assumptions and Problem

- Aalen and Husebye: "Consecutive MMC periods for each individual appear (to be) approximate renewal processes."
- *Translation*: The inter-event times T_{ij} 's are assumed stochastically independent.
- *Problem*: Under this IID assumption, and taking into account the informativeness of K and the rightcensoring mechanism, to estimate the inter-event distribution, F.

General Form of Data Accrual

Unit	Successive Inter-Event	Length of
#	Times or Gaptimes	Study Period
1	$T_{11}, T_{12}, \dots, T_{1j}, \dots$ IID F	$ au_1$
2	$T_{21}, T_{22}, \dots, T_{2j}, \dots$ IID F	$ au_2$
•••	• • •	•••
n	$T_{n1}, T_{n2},, T_{nj}, IID F$	$ au_{ m n}$

Calendar Times of Event Occurrences

$$S_{i0}=0$$
 and $S_{ij}=T_{i1}+T_{i2}+...+T_{ij}$

Number of Events in Observation Period

$$K_i = \max\{j: S_{ij} \le \tau_i\}$$

Upper limit of observation periods, τ 's, could be fixed, or assumed to be IID with unknown distribution G.

Observables

Unit	Vector of Observables
#	
1	$\mathbf{D}_1 = (K_1, T_{11}, T_{12}, \dots, T_{1K(1)}, \tau_1 - S_{1K(1)})$
2	$\mathbf{D}_2 = (K_2, T_{21}, T_{22}, \dots, T_{2K(2)}, \tau_2 - S_{2K(2)})$
• • •	•••
n	$\mathbf{D}_{n} = (K_{n}, T_{n1}, T_{n2}, \dots, T_{nK(n)}, \tau_{n} - S_{nK(n)})$

Main Theoretical Problem

Based on this data, to obtain an estimator of the unknown gaptime or inter-event time distribution, F; and determine its properties.

Relevance and Applicability

- Recurrent phenomena occur in a variety of settings.
 - Nuclear power plant stoppages.
 - Outbreak of a disease.
 - Terrorist attacks.
 - Labor strikes.
 - Hospitalization of a patient.
 - Tumor occurrence.
 - Epileptic seizures.
 - Non-life insurance claims.
 - When stock index (e.g., Dow Jones) decreases by at least 6% in one day.

Limitations of Existing Estimation Methods

- Consider only the first, possibly right-censored, observation per unit and use the product-limit estimator (PLE).
 - Loss of information
 - Inefficient
- Ignore the right-censored last observation, and use empirical distribution function (EDF).
 - Leads to bias.
 - Estimator actually inconsistent.

Review: Prior Results

Complete Data Setting

- $T_1, T_2, ..., T_n \text{ IID } F(t) = P(T \le t)$
- Empirical Survivor Function (EDF)

$$\hat{\overline{F}}(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i > t)$$

Asymptotics of EDF

$$\sqrt{n} \left(\hat{\overline{F}} - \overline{F} \right) \Rightarrow W_1$$

where W_1 is a zero-mean Gaussian process with covariance function

$$v_1(t) = \overline{F}(t)F(t).$$

In Hazards View

• Hazard rate function

$$\lambda(t) = \lim_{h \downarrow 0} \frac{1}{h} P\{t \le T < t + h \mid T \ge t\} = \frac{f(t)}{\overline{F}(t)}$$

Cumulative hazard function

$$\Lambda(t) = -\log\{\overline{F}(t)\} = \int_{0}^{t} \lambda(w)dw$$

• Equivalences

$$f(t) = \lambda(t)e^{-\Lambda(t)}$$
$$\overline{F}(t) = e^{-\Lambda(t)} = \prod_{s=0}^{t} \left[1 - d\Lambda(s)\right]$$

• Another representation of the variance

$$\upsilon_1(t) = \overline{F}(t)F(t) = \overline{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\overline{F}(w)}$$

Right-Censored Data

- Failure times: $T_1, T_2, ..., T_n \text{ IID F}$
- Censoring times: C₁, C₂, ..., C_n IID G
- Right-censored data

$$(Z_1, \delta_1), (Z_2, \delta_2), ..., (Z_n, \delta_n)$$

with

$$Z_i = \min(T_i, C_i)$$

$$\delta_i = I\{T_i \le C_i\}$$

• Product-limit or Kaplan-Meier Estimator

$$\hat{\overline{F}}(t) = \prod_{\{i:Z_{(i)} \le t\}} \left[1 - \frac{1}{n_i} \right]^{\delta_{(i)}}$$
$$Z_{(1)} \le Z_{(2)} \le \dots \le Z_{(n)}$$
$$n_{(i)} = \# \text{ at risk at } Z_{(i)}$$

PLE Properties

Asymptotics of PLE

$$\sqrt{n} \left(\hat{\overline{F}} - \overline{F} \right) \Rightarrow W_2$$

where W_2 is a zero-mean Gaussian process with covariance function

$$v_2(t) = \overline{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\overline{F}(w)\overline{G}(w)}$$

• If G(w) = 0 for all w, so no censoring,

$$\mathcal{U}_1(t) = \mathcal{U}_2(t)$$

Relevant Stochastic Processes for Recurrent Event Setting

• Calendar-Time Processes for ith unit

$$N_i^{\dagger}(s) = \sum_{j=1}^{\infty} I\{S_{ij} \le s; S_{ij} \le \tau_i\}$$
$$Y_i^{\dagger}(s) = I\{\tau_i \ge s\}$$

$$\begin{split} \mathcal{F}_{s}^{\dagger} &= \text{event history up to calendar-time } s \\ A_{i}^{\dagger}(s) &= \imath_{0}^{s} Y_{i}^{\dagger}(v) \lambda \left(v - S_{iN_{i}^{\dagger}(v-)}^{\dagger} \right) \mathrm{d} v \\ M_{i}^{\dagger}(s) &= N_{i}^{\dagger}(s) - A_{i}^{\dagger}(s) \end{split}$$

Then,

$$M^{\dagger}(s) = (M_1^{\dagger}(s), \dots, M_n^{\dagger}(s))$$

is a vector of square-integrable zero-mean martingales.

• Difficulty: arises because interest is on $\lambda(.)$ or $\Lambda(.)$, but these appear in the compensator process $A_i^{\dagger}(s)$ in form

$$\lambda \left(v - S_{iN_i^\dagger(v-)} \right)$$

$$v-S_{iN_i^\dagger(v-)}$$

is the length since last event at calendar time v

• Needed: Calendar-Gaptime Space



• Processes in Calendar-Gaptime Space

$$\begin{split} Z_i(s,t) &= I\{s - S_{iN_i^{\dagger}(s-)} \leq t\} \\ N_i(s,t) &= \int_0^s Z_i(v,t) N_i^{\dagger}(\mathrm{d}v) \\ A_i(s,t) &= \int_0^s Z_i(v,t) A_i^{\dagger}(\mathrm{d}v) \\ M_i(s,t) &= \int_0^s Z_i(v,t) M_i^{\dagger}(\mathrm{d}v) = N_i(s,t) - A_i(s,t) \\ Y_i(s,t) &= \sum_{j=1}^{N_i^{\dagger}(s-)} I\{T_{ij} \geq t\} + I\{(s \wedge \tau_i) - S_{iN_i^{\dagger}(s-)} \geq t\} \end{split}$$

• N_i(s,t) = # of events in calendar time [0,s] for ith unit whose gaptimes are at most t

• Y_i(s,t) = number of events in [0,s] for ith unit whose gaptimes are at least t: "at-risk" process Aggregated processes:

$$N(s,t) = \sum_{i=1}^{n} N_i(s,t);$$
$$A(s,t) = \sum_{i=1}^{n} A_i(s,t);$$
$$M(s,t) = \sum_{i=1}^{n} M_i(s,t).$$

• As $s \to \infty$,

$$\begin{split} N_i(s,t) &\xrightarrow{\text{a.s.}} N_i(\tau_i,t) = N_i(t) = \sum_{i=1}^{K_i} I\{T_{ij} \leq t\}; \\ Y_i(s,t) &\xrightarrow{\text{a.s.}} Y_i(t) = \sum_{j=1}^{K_i} I\{T_{ij} \geq t\} + I\{\tau_i - S_{iK_i} \geq t\}. \end{split}$$

"Change-of-Variable" Formulas

$$\begin{split} A(s,t) \, &= \, \sum_{i=1}^n \int_0^s Z_i(v,t) A_i^\dagger(\mathrm{d} v) = \int_0^t Y(s,w) \lambda(w) \mathrm{d} w \\ \int_0^s H_i(s,v-S_{iN_i^\dagger(v-)}) M_i(\mathrm{d} v,t) = \int_0^t H_i(s,w) M_i(s,\mathrm{d} w) \end{split}$$

Estimators of Λ and F for the Recurrent Event Setting

 $J(v,w)=I\{Y(v,w)>0\}$

By "change-of-variable" formula,

$$\int_0^t \frac{J(s,w)}{Y(s,w)} M(s,\mathrm{d}w) = \sum_{i=1}^n \int_0^s \frac{J(s,v-S_{iN_i^\dagger(v-)})}{Y(s,v-S_{iN_i^\dagger(v-)})} M_i(\mathrm{d}v,t)$$

RHS is a sq-int. zero-mean martingale, so

$$E\left\{\int_0^t \frac{J(s,w)}{Y(s,w)} N(s,\mathrm{d}w)\right\} = E\left\{\int_0^t J(s,w)\mathrm{d}\Lambda(w)\right\}$$

Estimator of $\Lambda(t)$

$$\hat{\Lambda}(s,t) = \int_0^t \frac{J(s,w)}{Y(s,w)} N(s,\mathrm{d} w) = \int_0^t \frac{N(s,\mathrm{d} w)}{Y(s,w)}$$

Estimator of F

• Since

$$\bar{F}(t) = \mathop{\Pi}_{w \leq t} \left[1 - \Lambda(\mathrm{d}w) \right]$$

by substitution principle,

$$\hat{\bar{F}}(s,t) = \mathop{\Pi}_{w \leq t} \left[1 - \hat{\Lambda}(s,\mathrm{d}w) \right] = \mathop{\Pi}_{w \leq t} \left[1 - \frac{N(s,\Delta w)}{Y(s,w)} \right]$$

a generalized product-limit estimator (GPLE).

• GPLE extends the EDF for complete data, and the PLE or KME for single-event right-censored data.

Computational Forms

• If T_{ij}'s are distinct,

$$\begin{split} \hat{\bar{F}}(s,t) &= \prod_{i=1}^{n} \prod_{\substack{j: \\ j: \\ S_{ij} \leq (s \wedge \tau_i)}} \prod_{\substack{T_{ij} \leq t \\ S_{ij} \leq (s \wedge \tau_i)}} \left[1 - \frac{1}{Y(s,T_{ij})} \right] \\ Y(s,T_{ij}) &= \sum_{k=1}^{n} \left\{ \sum_{l=1}^{N_k^{\dagger}(s-)} I\{T_{kl} \geq T_{ij}\} + I\{(s \wedge \tau_k) - S_{kN_k^{\dagger}(s-)} \geq T_{ij}\} \right\} \end{split}$$

• If s is large, combine data for all n units taking into account right-censoring of each unit's last observation. GPLE could be computed like the product-limit estimator. Statistical packages such as SAS, Splus, etc., could be utilized.

GPLE Finite-Sample Properties

• GPLE is positively biased, but with bias disappearing at an exponential rate as sample size increases.

$$0 \leq \operatorname{Bias}\{\hat{\bar{F}}(s,t)\} \leq F(t)[\pi_U(t)]^n$$

 $\pi_U(t) = \int_0^t \bar{F}(w) \mathrm{d}G(w) + \int_0^\infty \int_z^{z+t} \bar{F}(w-z) \mathrm{d}G(w) \mathrm{d}R(z)$

R(t) = renewal function of F

Special Case

 $F = \text{EXP}(\theta)$ and $G = \text{EXP}(\eta)$: $\pi_U(t) = 1 - e^{-(\theta + \eta)t}$

• Finite-sample variance function estimate:

$$\operatorname{Var}(\widehat{\bar{F}}(s,t)) = \widehat{\bar{F}}(s,t))^2 \int_0^t \frac{N(s,\mathrm{d}w)}{Y(s,w)[Y(s,w) - N(s,\Delta w)]}$$

Asymptotic Properties of GPLE

$$\begin{split} F^{\star j} &= j \text{th convolution of } F, j = 1, 2, \dots \\ R(t) &= \sum_{j=1}^{\infty} F^{\star j}(t) = \text{renewal function of } F; \\ G_s(w) &= \begin{cases} G(w) & \text{if } w < s \\ 1 & \text{if } w \ge s \end{cases} \\ E\{Y_1(s,t)\} &= y(s,t) \equiv \bar{F}(t) \left\{ \bar{G}_s(t-) + \int_{[t,\infty)} R(w-t) \mathrm{d} G_s(w) \right\} \\ d(s,t) &= \int_0^t \frac{\Lambda(\mathrm{d} w)}{y(s,w)} \end{split}$$

<u>Special Case</u>: If $F = EXP(\theta)$ and $G = EXP(\eta)$

$$\begin{split} d(s,t) &= I\{t \leq s\} \times \\ \theta \int_0^t \frac{\exp\{(\theta + \eta)w\}}{1 + \frac{\theta}{\eta} [1 - \exp\{-\eta(s - w)\} - \eta(s - w)\exp\{-\eta(s - w)\}]} \mathrm{d}w \\ d(\infty,t) &= \frac{\theta \eta}{(\theta + \eta)^2} \left\{ \exp\{(\theta + \eta)t\} - 1 \right\} \end{split}$$

Weak Convergence

Theorem: If $s \in (0, \infty)$ and $t^* \in (0, \infty)$ such that $y(s, t^*) > 0$ and if $\Lambda(t^*) < \infty$, then

$$\{W(s,t) = \sqrt{n}[\hat{\bar{F}}(s,t) - \bar{F}(t)] : t \in [0,t^*]\}$$

converges weakly to a zero-mean Gaussian process

 $\{W^{\infty}(s,t):t\in [0,t^{*}]\}$;

 $Cov[W^{\infty}(s,t_1), W^{\infty}(s,t_2)] = \bar{F}(t_1)\bar{F}(t_2)d[s,\min(t_1,t_2)].$

The proof of this result relied on a weak convergence theorem for recurrent and renewal settings developed in Pena, Strawderman and Hollander (2000), which utilized some ideas in Sellke (1988) and Gill (1980).

Comparison of Limiting Variance Functions

• EDF:
$$v_1(t) = \overline{F}(t)F(t) = \overline{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\overline{F}(w)}$$

• PLE:
$$v_2(t) = \overline{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\overline{F}(w)\overline{G}(w)}$$

• GPLE (recurrent event): For large s,

$$v_3(t) = \bar{F}(t)^2 \int_0^t \frac{\mathrm{d}\Lambda(w)}{\bar{F}(w)\bar{G}(w)\left\{1 + \frac{1}{\bar{G}(w)}\int_w^\infty R(u-w)\mathrm{d}G(u)\right\}}$$

• For large t or if in stationary state, $R(t) = t/\mu_F$, so approximately,

$$v_3(t) = \bar{F}(t)^2 \int_0^t \frac{\mathrm{d}\Lambda(w)}{\bar{F}(w)\bar{G}(w) \left\{1 + (\mu_F)^{-1}\mu_G(w)\right\}}$$

with $\mu_G(w)$ being the mean residual life of τ given $\tau \ge w$.

Wang-Chang Estimator
(JASA, '99)

$$K_i^* = \begin{cases} 1 & \text{if } K_i = 0 \\ K_i & \text{if } K_i > 0 \end{cases}$$

$$d^*(t) = \sum_{i=1}^n \left\{ \frac{I\{K_i > 0\}}{K_i^*} \sum_{j=1}^{K_i} I\{T_{ij} = t\} \right\}$$

$$= \sum_{i=1}^n \frac{1}{2} \left[\sum_{i=1}^{K_i} I\{T_{ii} > t\} + I\{\tau_i - S_i \kappa_i > t\} I\{K_i = 0\} \right]$$

$$R^*(t) = \sum_{i=1}^n \frac{1}{K_i^*} \left[\sum_{j=1}^{K_i} I\{T_{ij} \ge t\} + I\{\tau_i - S_{iK_i} \ge t\} I\{K_i = 0\} \right]$$

$$\hat{S}(t) = \prod_{i=1}^{n} \prod_{\{j: \ T_{ij} \le t\}} \left[1 - \frac{d^{*}(T_{ij})}{R^{*}(T_{ij})} \right]$$

• Beware! Wang and Chang developed this estimator to be able to handle correlated interevent times, so comparison with GPLE is not completely fair to their estimator!

Frailty-Induced Correlated Model

- Correlation induced according to a frailty model:
- $U_1, U_2, ..., U_n$ are IID unobserved Gamma(α, α) random variables, called frailties.
- Given $U_i = u$, $(T_{i1}, T_{i2}, T_{i3}, ...)$ are independent inter-event times with

$$\overline{F}(t | U_i = u) = [\overline{F}_0(t)]^u = \exp\left\{-u \int_0^t \lambda_0(w) dw\right\}.$$

• Marginal survivor function of T_{ij:}

$$\overline{F}(t) = E\left\{\left[\overline{F}_0(t)\right]^U\right\} = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^{\alpha}$$

Frailty-Model Estimator

- Frailty parameter, α, determines dependence among inter-event times.
- Small (Large) α: Strong (Weak) dependence.
- EM algorithm is needed to obtain the estimator, where the unobserved frailties are viewed as missing values.
- EM implementation parallels that of Nielsen, Gill, Andersen, and Sorensen (1992).

EM Algorithm

• E-STEP: Given $\hat{\Lambda}_0(.), \hat{\alpha}$, estimate Z's via

$$\hat{Z}_{i} = \frac{\hat{\alpha} + N_{i}^{\dagger}(s)}{\hat{\alpha} + \sum_{l=1}^{Q} \hat{\lambda}_{(l)}(s) Y_{i}(s, t_{(l)})}, \quad i = 1, 2, \dots, n.$$

• M-STEP: Given \hat{Z}_i 's

$$\hat{\lambda}_{(l)}(s) = \frac{\sum_{i=1}^{n} d_{(l)i}(s)}{\sum_{i=1}^{n} \hat{Z}_i Y_i(s, t_{(l)})}, \quad l = 1, 2, \dots, Q.$$

 α is estimated from profile likelihood.

• Resulting Estimator: FRMLE

• Able to prove that the GPLE is inconsistent in the presence of dependence: that is, when the frailty parameter α is finite.

Monte Carlo Studies

- Under gamma frailty model.
- $F = EXP(\theta)$: $\theta = 6$
- $G = EXP(\eta): \eta = 1$
- n = 50
- # of Replications = 1000
- Frailty parameter α took values in {Infty (IID), 6, 2}
- Computer programs: combinations of S-Plus and Fortran routines.
- Black = GPLE; Blue = WCPLE;
 Red = FRMLE

Simulated Comparison of the Three Estimators for Varying Frailty Parameter



Effect of the Frailty Parameter for Each of the Three Estimators



The Three Estimates of Inter-Event Survivor Function for the MMC Data Set



Migrating Moto Complex (MMC) Time, in minutes

IID assumption seems acceptable. Estimate of α is 10.2.