Nonparametric Estimation with Recurrent Event Data

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

Talk at Mississippi State Univ, 10/19/01

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 gastroenterology study concerning small bowel motility during fasting state.

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

• Consider unit/subject #3.
• $K = 3$
• Gap Times, $T_j$: 284, 59, 186
• Censored Time, $\tau - S_K$: 4
• Calendar Times, $S_j$: 284, 343, 529
• Limit of Obs. Period: $\tau = 533$

Calendar Scale
Features of Data Set

- Random observation period per subject (administrative constraints).
- Length of period: $\tau$
- 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: $S_1, S_2, \ldots, S_K$.
- Right-censoring variable: $\tau - S_K$. 
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 right-censoring mechanism, to estimate the inter-event distribution, F.
General Form of Data Accrual

<table>
<thead>
<tr>
<th>Unit #</th>
<th>Successive Inter-Event Times or Gaptimes</th>
<th>Length of Study Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$T_{11}, T_{12}, \ldots, T_{1j}, \ldots$ IID $F$</td>
<td>$\tau_1$</td>
</tr>
<tr>
<td>2</td>
<td>$T_{21}, T_{22}, \ldots, T_{2j}, \ldots$ IID $F$</td>
<td>$\tau_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>$T_{n1}, T_{n2}, \ldots, T_{nj}, \ldots$ IID $F$</td>
<td>$\tau_n$</td>
</tr>
</tbody>
</table>

Calendar Times of Event Occurrences

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

Number of Events in Observation Period

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

Upper limit of observation periods, $\tau$’s, could be fixed, or assumed to be IID with unknown distribution $G$. 
### Observables

<table>
<thead>
<tr>
<th>Unit #</th>
<th>Vector of Observables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$D_1 = (K_1, T_{11}, T_{12}, \ldots, T_{1K(1)}, \tau_{1-S_{1K(1)}})$</td>
</tr>
<tr>
<td>2</td>
<td>$D_2 = (K_2, T_{21}, T_{22}, \ldots, T_{2K(2)}, \tau_{2-S_{2K(2)}})$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>$D_n = (K_n, T_{n1}, T_{n2}, \ldots, T_{nK(n)}, \tau_{n-S_{nK(n)}})$</td>
</tr>
</tbody>
</table>

### Theoretical Problem

To obtain an estimator of the gap-time or inter-event time distribution, $F$; and to determine its properties.
Relevance and Applicability

• Recurrent phenomena occur in a variety of settings.
  – 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 ("biased sampling").
  - Estimator actually inconsistent.
Review: Prior Results

Single-Event Complete Data

• \( T_1, T_2, \ldots, T_n \) IID \( F(t) = \Pr(T \leq t) \)

• Empirical Survivor Function (EDF)

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

• Asymptotics of EDF

\[
\sqrt{n} \left( \hat{F} - F \right) \Rightarrow W_1
\]

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

\[
\nu_1(t) = \overline{F}(t)F(t).
\]
In Hazards View

• Hazard rate function

\[
\lambda(t) = \lim_{h \downarrow 0} \frac{1}{h} P\{t \leq T < t + h \mid T \geq t\} = \frac{f(t)}{F(t)}
\]

• Cumulative hazard function

\[
\Lambda(t) = -\log\{F(t)\} = \int_0^t \lambda(w)dw
\]

• Equivalences

\[
f(t) = \lambda(t)e^{-\Lambda(t)}
\]

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

• Another representation of the variance

\[
u_1(t) = F(t)F(t) = F(t)^2 \int_0^t \frac{d\Lambda(w)}{F(w)}
\]
Single-Event Right-Censored Data

- Failure times: \( T_1, T_2, \ldots, T_n \) IID \( F \)
- Censoring times: \( C_1, C_2, \ldots, C_n \) IID \( G \)

- Right-censored data
  \((Z_1, \delta_1), (Z_2, \delta_2), \ldots, (Z_n, \delta_n)\)

  with
  \[ Z_i = \min(T_i, C_i) \]
  \[ \delta_i = I\{T_i \leq C_i\} \]

- Product-limit or Kaplan-Meier Estimator

\[
\hat{F}(t) = \prod_{\{i:Z_{(i)} \leq t\}} \left[ 1 - \frac{1}{n_i} \right]^{\delta_{(i)}}
\]

\( Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)} \)

\( n_{(i)} = \# \text{ at risk at } Z_{(i)} \)
PLE Properties

• Asymptotics of PLE

\[ \sqrt{n} \left( \hat{F} - F \right) \Rightarrow W_2 \]

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

\[ \nu_2(t) = \overline{F(t)}^2 \int_0^t \frac{d\Lambda(w)}{\overline{F(w)G(w)}} \]

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

\[ \nu_1(t) = \nu_2(t) \]
Relevant Stochastic Processes for Recurrent Event Setting

• Calendar-Time Processes for ith unit

\[ N_i^\dagger(s) = \sum_{j=1}^{\infty} I\{S_{ij} \leq s; S_{ij} \leq \tau_i\} \]

\[ Y_i^\dagger(s) = I\{\tau_i \geq s\} \]

\[ \mathcal{F}_s^\dagger = \text{event history up to calendar-time } s \]

\[ A_i^\dagger(s) = \int_0^s Y_i^\dagger(v) \lambda \left( v - S_{iN_i^\dagger(v^-)} \right) dv \]

\[ M_i^\dagger(s) = N_i^\dagger(s) - A_i^\dagger(s) \]

Then,

\[ M^\dagger(s) = (M_1^\dagger(s), \ldots, M_n^\dagger(s)) \]

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 $A_i^+(t)$ in form

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

$v - S_{iN_i^+(v-)}$ is the length since last event at calendar time $v$

• **Needed:** Calendar-GapTime Space

For Unit 3 in MMC Data
• Processes in Calendar-Gaptime Space

\[ Z_i(s, t) = I\{s - S_{iN_i^+(s-)} \leq t\} \]

\[ N_i(s, t) = \int_0^s Z_i(v, t) N_i^+(dv) \]

\[ A_i(s, t) = \int_0^s Z_i(v, t) A_i^+(dv) \]

\[ M_i(s, t) = \int_0^s Z_i(v, t) M_i^+(dv) = N_i(s, t) - A_i(s, t) \]

\[ Y_i(s, t) = \sum_{j=1}^{N_i^+(s-)} I\{T_{ij} \geq t\} + I\{(s \wedge \tau_i) - S_{iN_i^+(s-)} \geq t\} \]

• \( N_i(s,t) = \# \) 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
MMC Unit #3

\[ K_3(s=400) = 2 \]

\[ N_3(s=400,t=100) = 1 \]

\[ Y_3(s=400,t=100) = 1 \]
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 \( s \to \infty \),

\[ N_i(s, t) \xrightarrow{a.s.} N_i(\tau_i, t) = N_i(t) = \sum_{i=1}^{K_i} I\{T_{ij} \leq t\}; \]

\[ Y_i(s, t) \xrightarrow{a.s.} Y_i(t) = \sum_{j=1}^{K_i} I\{T_{ij} \geq t\} + I\{\tau_i - S_i \tau_i I \geq t\}. \]

“Change-of-Variable” Formulas

\[ A(s, t) = \sum_{i=1}^{n} \int_{0}^{s} Z_i(v, t)A_i^\dagger(dv) = \int_{0}^{t} Y(s, w)\lambda(w)dw \]

\[ \int_{0}^{s} H_i(s, v - S_i N_i^\dagger(v_\cdot ))M_i(dv, t) = \int_{0}^{t} H_i(s, w)M_i(s, dw) \]
Estimators of $\Lambda$ and $F$ for the Recurrent Event Setting

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

By “change-of-variable” formula,

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

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

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

Estimator of $\Lambda(t)$

$$\hat{\Lambda}(s, t) = \int_0^t \frac{J(s, w)}{Y(s, w)} N(s, dw) = \int_0^t \frac{N(s, dw)}{Y(s, w)}$$
Estimator of F

• Since

\[ \hat{F}(t) = \prod_{w \leq t} [1 - \Lambda(dw)] \]

by substitution principle,

\[
\hat{F}(s, t) = \prod_{w \leq t} [1 - \hat{\Lambda}(s, dw)] = \prod_{w \leq t} \left[ 1 - \frac{N(s, \Delta w)}{Y(s, w)} \right]
\]

a generalized product-limit estimator (GPLE).

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

\[ F^{*j} = j\text{th convolution of } F, j = 1, 2, \ldots \]

\[ R(t) = \sum_{j=1}^{\infty} F^{*j}(t) = \text{renewal function of } F; \]

\[ G_s(w) = \begin{cases} G(w) & \text{if } w < s \\ 1 & \text{if } w \geq s \end{cases} \]

\[ E\{Y_1(s, t)\} = y(s, t) = \bar{F}(t) \left\{ \bar{G}_s(t-) + \int_{[t, \infty)} R(w-t) dG_s(w) \right\} \]

\[ d(s, t) = \int_0^t \frac{\Lambda(dw)}{y(s, w)} \]

**Special Case:** If \( F = \text{EXP}(\theta) \) and \( G = \text{EXP}(\eta) \)

\[ d(s, t) = I\{t \leq s\} \times \theta \int_0^t \frac{\exp\{(\theta + \eta)w\}}{1 + \frac{\theta}{\eta} \left[ 1 - \exp\{-\eta(s-w)\} - \eta(s-w) \exp\{-\eta(s-w)\} \right]} dw \]

\[ d(\infty, t) = \frac{\theta \eta}{(\theta + \eta)^2} \left\{ \exp\{(\theta + \eta)t\} - 1 \right\} \]
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

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

converges weakly to a zero-mean Gaussian process

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

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

Comparison of Limiting Variance Functions

- **EDF**: \( v_1(t) = \bar{F}(t)F(t) = \bar{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\bar{F}(w)} \)

- **PLE**: \( v_2(t) = \bar{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\bar{F}(w)\bar{G}(w)} \)

- **GPLE** (recurrent event): For large \( s \),

\[
\nu_3(t) = \bar{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\bar{F}(w)\bar{G}(w)} \left\{1 + \frac{1}{G(w)} \int_w^\infty R(u - w)dG(u)\right\}
\]

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

\[
\nu_3(t) = \bar{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\bar{F}(w)\bar{G}(w)} \left\{1 + (\mu_F)^{-1}\mu_G(w)\right\}
\]

with \( \mu_G(w) \) being the **mean residual life** of \( \tau \) given \( \tau \geq w \).
Wang-Chang Estimator
(JASA, ‘99)

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

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

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

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

- **Beware!** Wang and Chang developed this estimator to be able to handle correlated inter-event times, so comparison with GPLE is not completely fair to their estimator!
Frailty-Induced Correlated Model

• Correlation induced according to a frailty model:

• $U_1, U_2, \ldots, U_n$ are IID unobserved $\text{Gamma}(\alpha, \alpha)$ random variables, called frailties.

• Given $U_i = u$, $(T_{i1}, T_{i2}, T_{i3}, \ldots)$ are independent inter-event times with

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

• Marginal survivor function of $T_{ij}$:

$$
\bar{F}(t) = E\left\{[\bar{F}_0(t)]^U\right\} = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^\alpha
$$
Frailty-Model Estimator

• Frailty parameter, $\alpha$, determines dependence among inter-event times. Small (Large) $\alpha$: Strong (Weak) dependence.


• GPLE needed in EM algorithm.

• GPLE is not consistent when frailty parameter is finite, that is, when IID model does not hold.
Monte Carlo Studies

- Under gamma frailty model.
- \( F = \text{EXP}(\theta): \theta = 6 \)
- \( G = \text{EXP}(\eta): \eta = 1 \)
- \( n = 50 \)
- # of Replications = 1000
- Frailty parameter \( \alpha \) took values in \{Infty (IID), 6, 2\}
- Computer programs: S-Plus and Fortran routines.
Simulated Comparison of the Three Estimators for Varying Frailty Parameter
Black=GPLE; Blue=WCPLE; Red=FRMLE

<table>
<thead>
<tr>
<th>Frailty Parameter $\alpha$</th>
<th>Simulated Bias Function</th>
<th>Simulated RMSE Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\infty$</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>IID</td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>6</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
</tr>
</tbody>
</table>
Effect of the Frailty Parameter ($\alpha$) for Each of the Three Estimators (Black=Infty; Blue=6; Red=2)
The Three Estimates of Inter-Event Survivor Function for the MMC Data Set

IID assumption seems acceptable. Estimate of $\alpha$ is 10.2.