# 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

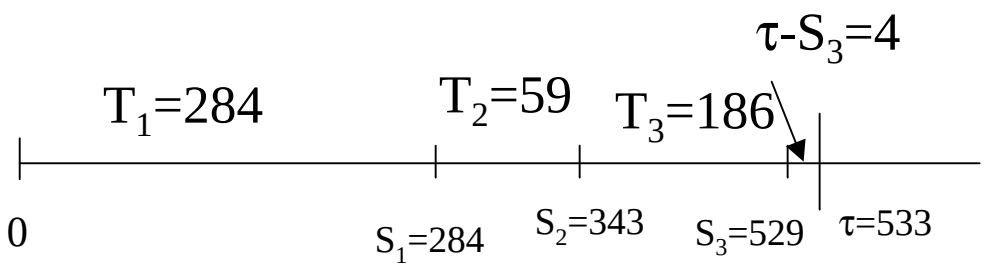
(Source: Aalen and Husebye ('91), *Statistics in Medicine*)

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

| Unit # i | #Complete ($K_i$=K(i)) | Complete Observed Successive Periods ($T_{ij}$) | Censored ($\tau_i - S_{iK(i)}$) |
|---|---|---|---|
| 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, $T_j$: 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

| Unit # | Successive Inter-Event Times or Gaptimes | Length of Study Period |
|---|---|---|
| 1 | $T_{11}, T_{12}, \ldots, T_{1j}, \ldots$ IID F | $\tau_1$ |
| 2 | $T_{21}, T_{22}, \ldots, T_{2j}, \ldots$ IID F | $\tau_2$ |
| … | … | … |
| n | $T_{n1}, T_{n2}, \ldots, T_{nj}, \ldots$ IID F | $\tau_n$ |

Calendar Times of Event Occurrences

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

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

| Unit # | Vector of Observables |
|---|---|
| 1 | $\mathbf{D}_1 = (K_1, T_{11}, T_{12},\ldots,T_{1K(1)}, \tau_1\text{-}S_{1K(1)})$ |
| 2 | $\mathbf{D}_2 = (K_2, T_{21}, T_{22},\ldots,T_{2K(2)}, \tau_2\text{-}S_{2K(2)})$ |
| … | … |
| n | $\mathbf{D}_n = (K_n, T_{n1}, T_{n2},\ldots,T_{nK(n)}, \tau_n\text{-}S_{nK(n)})$ |

## 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 $F(t) = P(T \leq t)$
- Empirical Survivor Function (EDF)

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

- Asymptotics of EDF

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

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

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

# 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

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

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

- Another representation of the variance

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

# 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

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

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

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

# PLE Properties

- Asymptotics of PLE

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

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

$$\upsilon_2(t) = \overline{F}(t)^2 \int_0^t \frac{d\Lambda(w)}{\overline{F}(w)\overline{G}(w)}$$

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

$$\upsilon_1(t) = \upsilon_2(t)$$

# Relevant Stochastic Processes for Recurrent Event Setting

- Calendar-Time Processes for ith unit

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

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

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

$$A_i^\dagger(s) = \int_0^s Y_i^\dagger(v)\lambda\left(v - S_{iN_i^\dagger(v-)}\right) \mathrm{d}v$$

$$M_i^\dagger(s) = N_i^\dagger(s) - A_i^\dagger(s)$$

Then,

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

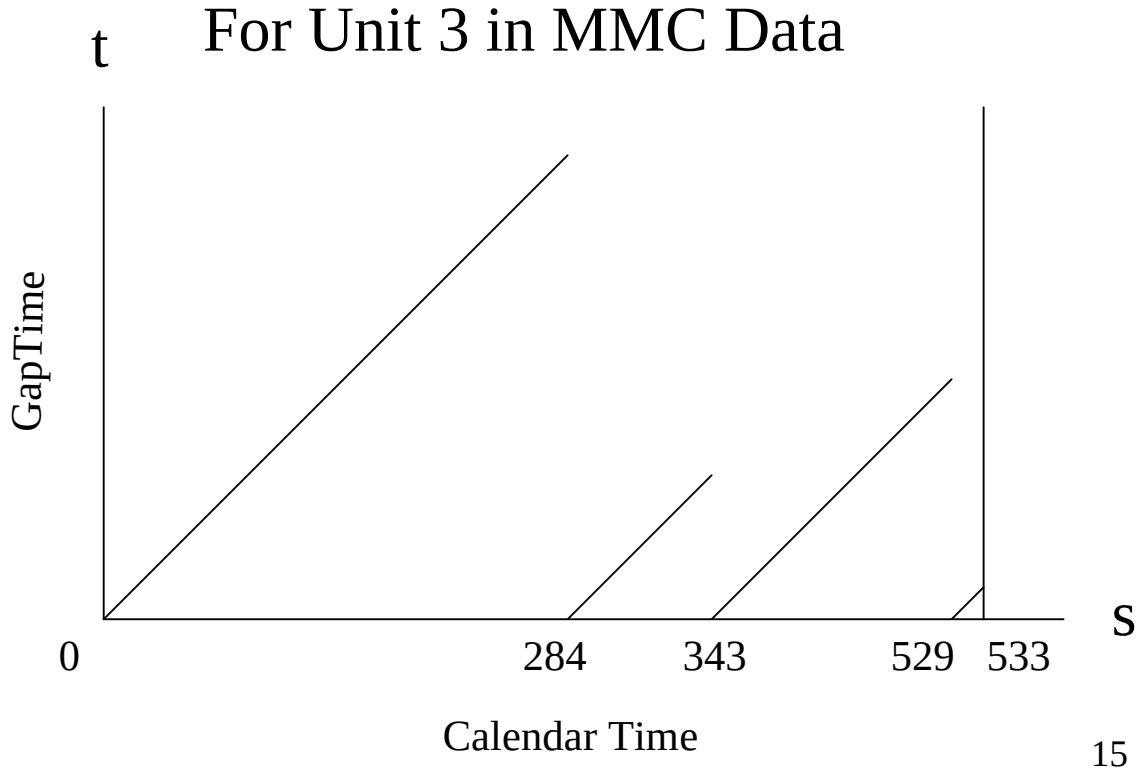
is a vector of square-integrable zero-mean martingales.

• Difficulty: arises because interest is on λ(.) or Λ(.), 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

$$Z_i(s,t) = I\{s - S_{iN_i^\dagger(s-)} \leq t\}$$

$$N_i(s,t) = \int_0^s Z_i(v,t) N_i^\dagger(\mathrm{d}v)$$

$$A_i(s,t) = \int_0^s Z_i(v,t) A_i^\dagger(\mathrm{d}v)$$

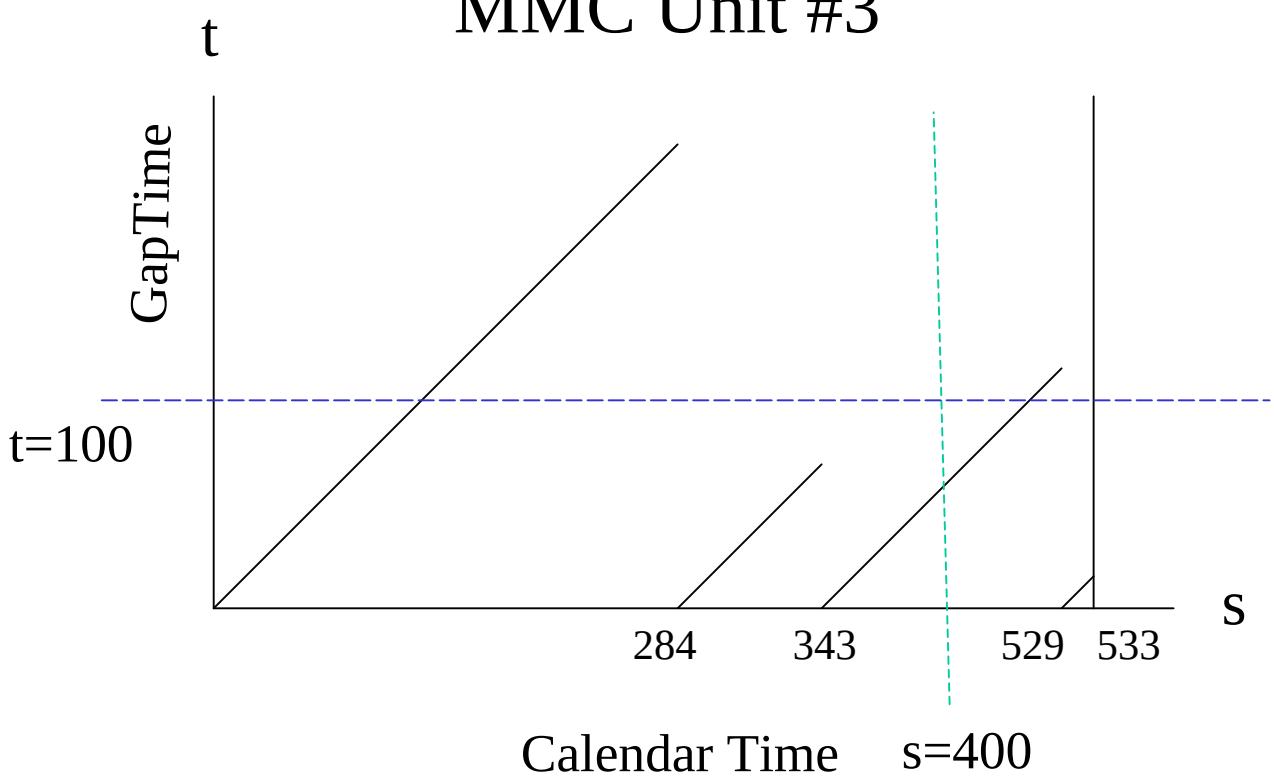
$$M_i(s,t) = \int_0^s Z_i(v,t) M_i^\dagger(\mathrm{d}v) = N_i(s,t) - A_i(s,t)$$

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

- $N_i(s,t)$ = # 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 $s \to \infty$,

$$N_i(s,t) \xrightarrow{\text{a.s.}} N_i(\tau_i,t) = N_i(t) = \sum_{i=1}^{K_i} I\{T_{ij} \le t\};$$

$$Y_i(s,t) \xrightarrow{\text{a.s.}} Y_i(t) = \sum_{j=1}^{K_i} I\{T_{ij} \ge t\} + I\{\tau_i - S_{iK_i} \ge t\}.$$

## "Change-of-Variable" Formulas

$$A(s,t) = \sum_{i=1}^{n} \int_0^s Z_i(v,t) A_i^{\dagger}(\mathrm{d}v) = \int_0^t Y(s,w)\lambda(w)\mathrm{d}w$$

$$\int_0^s H_i(s, v - S_{iN_i^{\dagger}(v-)}) M_i(\mathrm{d}v, t) = \int_0^t H_i(s,w) M_i(s, \mathrm{d}w)$$

# 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 Λ(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

$$\bar{F}(t) = \prod_{w \le t} [1 - \Lambda(\mathrm{d}w)]$$

by substitution principle,

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

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

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

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

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

$$E\{Y_1(s,t)\} = y(s,t) \equiv \bar{F}(t)\left\{\bar{G}_s(t-) + \int_{[t,\infty)} R(w-t)\mathrm{d}G_s(w)\right\}$$

$$d(s,t) = \int_0^t \frac{\Lambda(\mathrm{d}w)}{y(s,w)}$$

Special Case: If F = EXP(θ) and G = EXP(η)

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

$$d(\infty,t) = \frac{\theta\eta}{(\theta+\eta)^2}\{\exp\{(\theta+\eta)t\} - 1\}$$

# 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

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

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

Proof relied on weak convergence theorem for recurrent and renewal settings in Pena, Strawderman and Hollander (2000), which utilized ideas in Sellke (1988) and Gill (1980).

# 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,

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

- 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 $\tau$ given $\tau \geq w$.

# Wang-Chang Estimator (JASA, '99)

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

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

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

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

- 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 Gamma($\alpha, \alpha$) random variables, called frailties.
- Given $U_i = u$, $(T_{i1}, T_{i2}, T_{i3}, \ldots)$ are independent inter-event times with

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

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

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

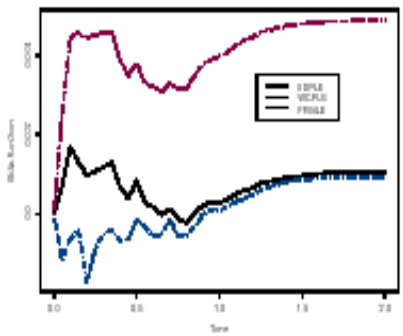
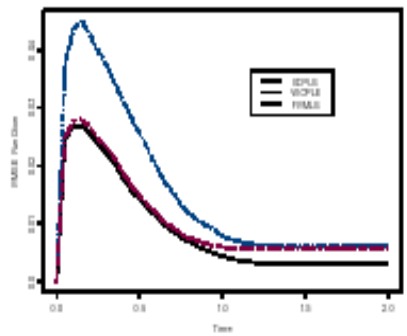
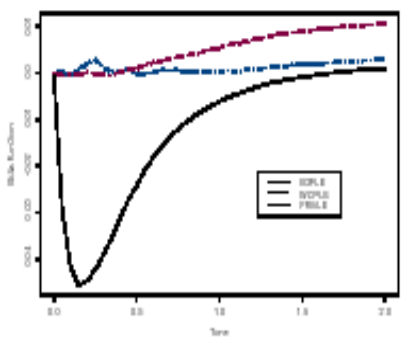
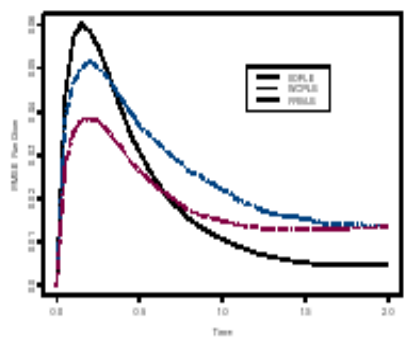
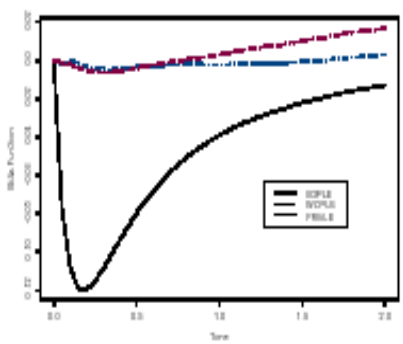
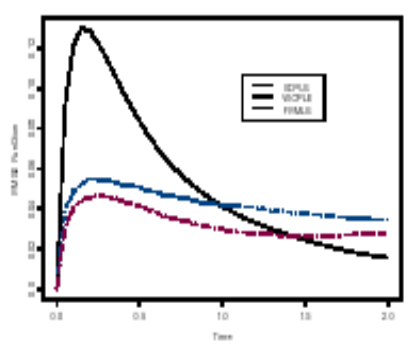
# Frailty-Model Estimator

- Frailty parameter, $\alpha$, determines dependence among inter-event times. Small (Large) $\alpha$: Strong (Weak) dependence.
- EM algorithm needed to obtain estimator (FRMLE). Unobserved frailties viewed as missing. Parallels Nielsen, Gill, Andersen, and Sorensen (1992).
- GPLE needed in EM algorithm.
- GPLE is not consistent when frailty parameter is finite, that is, when IID model does not hold.

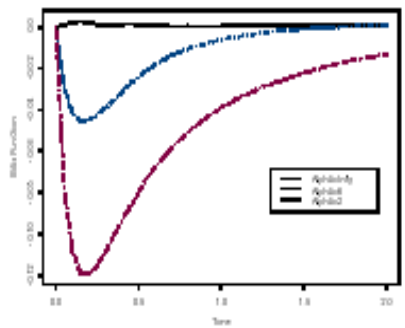
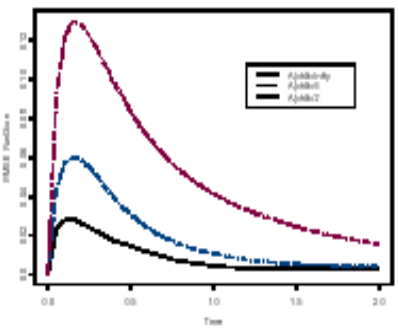
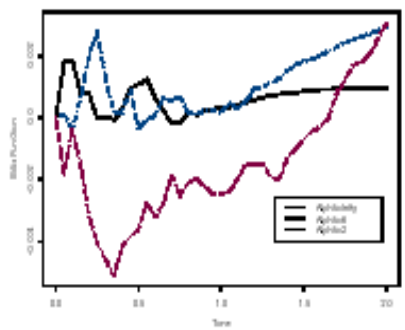
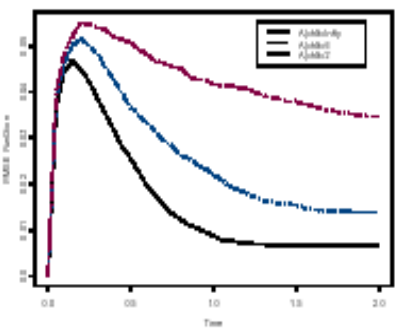
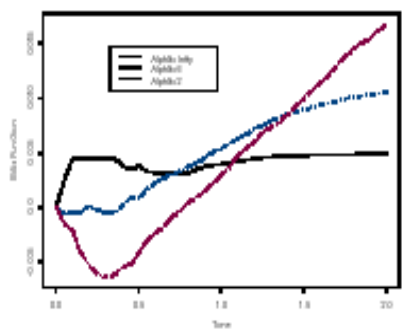
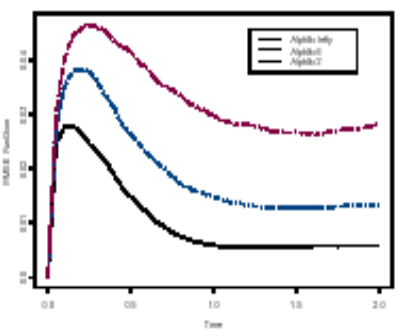
# Monte Carlo Studies

- Under gamma frailty model.
- F = EXP(θ): θ = 6
- G = EXP(η): η = 1
- n = 50
- # of Replications = 1000
- Frailty parameter α took values in {Infty (IID), 6, 2}
- Computer programs: S-Plus and Fortran routines.

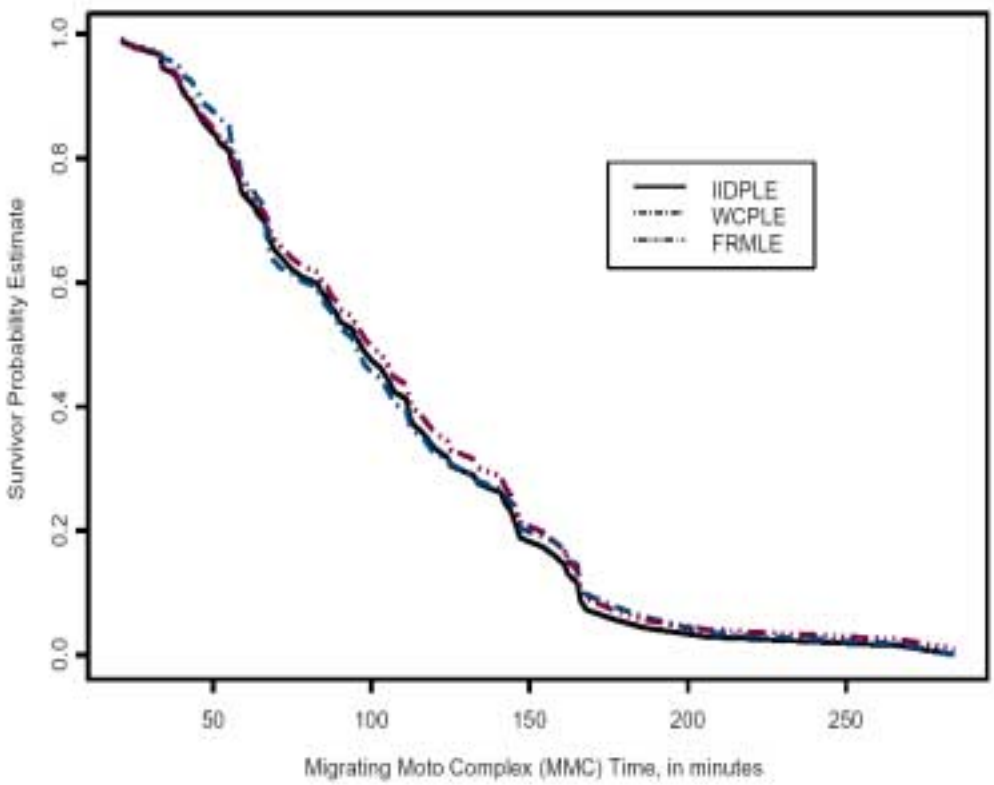
# Simulated Comparison of the Three Estimators for Varying Frailty Parameter Black=GPLE; Blue=WCPLE; Red=FRMLE

| Frailty Parameter $\alpha$ | Simulated Bias Function | Simulated RMSE Function |
|---|---|---|
| $\infty$ IID |  |  |
| 6 |  |  |
| 2 |  |  |

# Effect of the Frailty Parameter ($\alpha$) for Each of the Three Estimators (Black=Infty; Blue=6; Red=2)

| Estimator | Simulated Bias Function | Simulated RMSE Function |
|---|---|---|
| IIDPLE |  |  |
| WCPLE |  |  |
| FRMLE |  |  |

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




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