Nonparametric Bayesian modeling and inference for renewal processes

Athanasios Kottas

Department of Applied Mathematics and Statistics, University of California, Santa Cruz

Joint work with Sai Xiao and Bruno Sansó, *University of California, Santa Cruz*

ISBA 2014 World Meeting, Cancún, Mexico
July 14–18, 2014

Outline

Motivation

- Renewal processes are continuous-time counting processes that extend (homogeneous) Poisson processes.
- Applications of renewal processes in software reliability, queueing systems, and modeling of earthquake occurrences.

Motivation

- Renewal processes are continuous-time counting processes that extend (homogeneous) Poisson processes.
- Applications of renewal processes in software reliability, queueing systems, and modeling of earthquake occurrences.
- **Objective:** develop a Bayesian nonparametric modeling approach for renewal processes that balances
  - model flexibility: general inter-arrival densities; clustering and declustering for temporal point patterns
  - efficient posterior simulation for inference that properly incorporates the normalizing constant of the renewal process likelihood.

Motivation

- Renewal processes are continuous-time counting processes that extend (homogeneous) Poisson processes.
- Applications of renewal processes in software reliability, queueing systems, and modeling of earthquake occurrences.
- **Objective:** develop a Bayesian nonparametric modeling approach for renewal processes that balances
  - model flexibility: general inter-arrival densities; clustering and declustering for temporal point patterns
  - efficient posterior simulation for inference that properly incorporates the normalizing constant of the renewal process likelihood.
- Extension to modeling for inhomogeneous Markov interval processes.

Renewal processes

- A renewal process $\{N(t) : t \geq 0\}$ is defined as $N(t) = \max\{n : T_n \leq t\}$, where

$$T_0 = 0, \quad T_n = X_1 + X_2 + ... + X_n, \ \ n \geq 1$$

  and the X_i are independent and identically distributed r.v.s supported by $\mathbb{R}^+$ (with $0 < \mu = \mathbb{E}(X_1) < \infty$).

Renewal processes

- A renewal process $\{N(t) : t \geq 0\}$ is defined as $N(t) = \max\{n : T_n \leq t\}$, where

$$T_0 = 0, \quad T_n = X_1 + X_2 + ... + X_n, \;\; n \geq 1$$

  and the X_i are independent and identically distributed r.v.s supported by $\mathbb{R}^+$ (with $0 < \mu = \mathbb{E}(X_1) < \infty$).

- Think of T_n as the "n-th arrival time" and X_n as the "n-th inter-arrival time".

Renewal processes

- A renewal process $\{N(t) : t \geq 0\}$ is defined as $N(t) = \max\{n : T_n \leq t\}$, where

$$T_0 = 0, \quad T_n = X_1 + X_2 + ... + X_n, \;\; n \geq 1$$

  and the X_i are independent and identically distributed r.v.s supported by $\mathbb{R}^+$ (with $0 < \mu = \mathbb{E}(X_1) < \infty$).

- Think of T_n as the "n-th arrival time" and X_n as the "n-th inter-arrival time".

- Renewal process distribution characterized from the inter-arrival distribution, i.e., the distribution F of the X_i.
  - H_k d.f. of T_k: defined through $H_1 = F$, and

$$H_{k+1}(t) = \int_0^t H_k(t-u)\mathrm{d}F(u), \;\; \text{for } k \geq 1.$$

  - $\Pr(N(t) = k) = H_k(t) - H_{k+1}(t)$.

Renewal processes

- Renewal function. The expected number of occurrences of the event of interest in time interval $[0, t]$:

$$M(t) = \mathbb{E}(N(t)) = \sum_{k=1}^{\infty} H_k(t).$$

- Renewal equation: $M(t) = F(t) + \int_0^t M(t-u) \mathrm{d}F(u)$.

Renewal processes

- Renewal function. The expected number of occurrences of the event of interest in time interval $[0, t]$:

$$M(t) = \mathbb{E}(N(t)) = \sum_{k=1}^{\infty} H_k(t).$$

- Renewal equation: $M(t) = F(t) + \int_0^t M(t-u)\mathrm{d}F(u)$.

- K function, $K(t) = \mu M(t)$, commonly used to study clustering properties of a process, borrowing ideas from spatial point patterns
  - $K(t) > t \rightarrow$ clustering
  - $K(t) < t \rightarrow$ declustering
  - relative to the (homogeneous) Poisson process: special case of a renewal process with exponential inter-arrival distribution, for which $K(t) = t$.

Renewal process likelihood

- Temporal point pattern $\{0 = t_0 < t_1 < t_2 < \ldots < t_n < T\}$ observed in $[0, T]$.

Renewal process likelihood

- Temporal point pattern $\{0 = t_0 < t_1 < t_2 < \ldots < t_n < T\}$ observed in $[0, T]$.
- For the likelihood:

$$\begin{aligned} & \Pr(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n, T_{n+1} > T) \\ = \; & \Pr(X_1 = t_1, X_2 = t_2 - t_1, \ldots, X_n = t_n - t_{n-1}, X_{n+1} > T - t_n) \\ = \; & \left\{1 - \int_{t_n}^{T} f(u - t_n) \mathrm{d}u\right\} \prod_{i=1}^{n} f(t_i - t_{i-1}) \end{aligned}$$

where $f(\cdot)$ is the inter-arrival density function.

Renewal process likelihood

- Temporal point pattern $\{0 = t_0 < t_1 < t_2 < ... < t_n < T\}$ observed in $[0, T]$.

- For the likelihood:

$$
\begin{aligned}
& \Pr(T_1 = t_1, T_2 = t_2, ..., T_n = t_n, T_{n+1} > T) \\
= & \Pr(X_1 = t_1, X_2 = t_2 - t_1, ..., X_n = t_n - t_{n-1}, X_{n+1} > T - t_n) \\
= & \left\{1 - \int_{t_n}^{T} f(u - t_n)\mathrm{d}u\right\} \prod_{i=1}^{n} f(t_i - t_{i-1})
\end{aligned}
$$

  where $f(\cdot)$ is the inter-arrival density function.

- Density estimation problem with a twist (the likelihood normalizing constant).
  - Parametric models: Weibull or gamma inter-arrival distributions.
  - NPB prior models?

Mixture of Erlang densities model

- Model the inter-arrival density through a mixture of Erlang densities with common scale parameter

$$f(t) \equiv f(t \mid G, \theta) = \sum_{j=1}^{J} \omega_j \, \mathrm{gamma}(t \mid j, \theta)$$

where the weights are defined through a d.f. G on $\mathbb{R}^+$:

$$\omega_j = G(j\theta) - G((j-1)\theta), \; j = 1, ..., J-1; \;\; \omega_J = 1 - G((J-1)\theta)$$

Mixture of Erlang densities model

- Model the inter-arrival density through a mixture of Erlang densities with common scale parameter

$$f(t) \equiv f(t \mid G, \theta) = \sum_{j=1}^{J} \omega_j \, \text{gamma}(t \mid j, \theta)$$

where the weights are defined through a d.f. G on $\mathbb{R}^+$:

$$\omega_j = G(j\theta) - G((j-1)\theta), \; j = 1, ..., J-1; \;\; \omega_J = 1 - G((J-1)\theta)$$

- *Semiparametric* model completed with a Dirichlet process (DP) prior DP(α, G_0) for G, a gamma prior for θ, and a discrete uniform prior for J given θ.

Mixture of Erlang densities model

- Model the inter-arrival density through a mixture of Erlang densities with common scale parameter

$$f(t) \equiv f(t \mid G, \theta) = \sum_{j=1}^{J} \omega_j \, \text{gamma}(t \mid j, \theta)$$

where the weights are defined through a d.f. G on $\mathbb{R}^+$:

$$\omega_j = G(j\theta) - G((j-1)\theta), \; j = 1, ..., J-1; \;\; \omega_J = 1 - G((J-1)\theta)$$

- *Semiparametric* model completed with a Dirichlet process (DP) prior DP(α, G_0) for G, a gamma prior for θ, and a discrete uniform prior for J given θ.
  - Note that $\mathbb{E}(\sum_{k=1}^{K} \omega_k) = \mathbb{E}(G(K\theta)) = G_0(K\theta)$.
  - With a Weibull for G_0, $G_0(K\theta) = 1 - \exp(-(K\theta/\tau)^{\phi})$.

Mixture of Erlang densities model

- Posterior simulation via Gibbs sampling based on augmented model with latent configuration variables $y_1, \ldots, y_n, y_{n+1} \mid G \overset{i.i.d.}{\sim} G$:

$$\prod_{i=1}^{n} \sum_{j=1}^{J} \text{gamma}(t_i - t_{i-1} \mid j, \theta) \mathbf{1}_{((j-1)\theta, j\theta]}(y_i)$$
$$\times \quad \sum_{j=1}^{J} \left\{ 1 - \int_{t_n}^{T} \text{gamma}(u - t_n \mid j, \theta) \mathrm{d}u \right\} \mathbf{1}_{((j-1)\theta, j\theta]}(y_{n+1})$$

- Key feature of the Erlang mixture structure: through the additional latent variable y_{n+1}, the normalizing constant can be handled in similar fashion with the other likelihood components.

Mixture of Erlang densities model

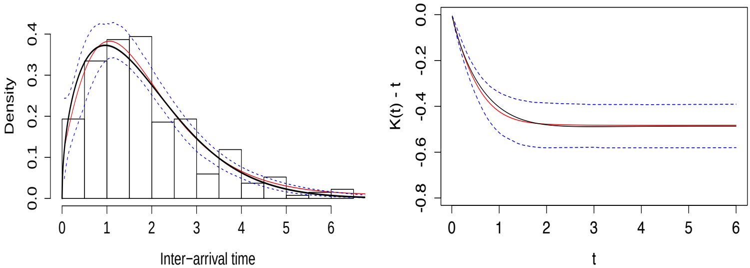
- The model enables both clustering and declustering patterns.

Mixture of Erlang densities model

- The model enables both clustering and declustering patterns.
- Inference for the K function, $K(t) = \mu M(t)$:
  - use the renewal equation, $M(t) = F(t) + \int_0^t M(t-u)\mathrm{d}F(u)$, to obtain the Laplace transform of $M(t)$ through the Laplace transform of the inter-arrival density $f(t)$;
  - obtain posterior realizations for $M(t)$, and thus for $K(t)$, through the inverse Laplace transform.

Simulation 1: Weibull inter-arrival distribution

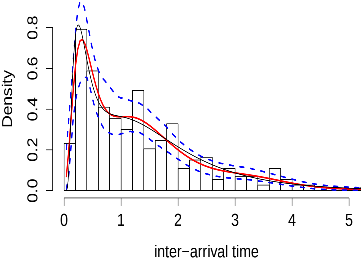
Pattern comprising 269 points in time window $(0, 500)$ generated under a Weibull inter-arrival distribution, with shape parameter > 1 (declustering).

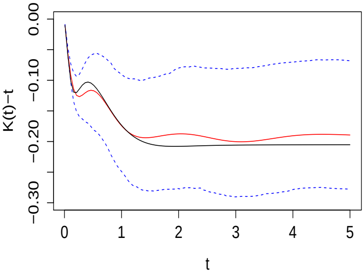


Posterior mean and 95% interval estimates for the inter-arrival density (left), and the K function (right). The black lines denote the true functions.

Simulation 2: mixture inter-arrival distribution

Pattern of 366 points in time window $(0, 500)$ generated under a two-component inverse Gaussian mixture for the inter-arrival distribution (declustering).

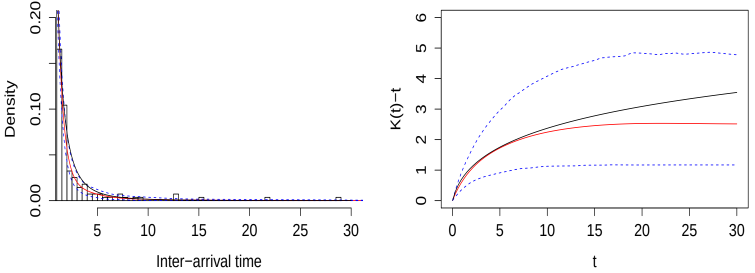




Posterior mean and 95% interval estimates for the inter-arrival density (left), and the K function (right). The black lines denote the true functions.

Simulation 3: Pareto inter-arrival distribution

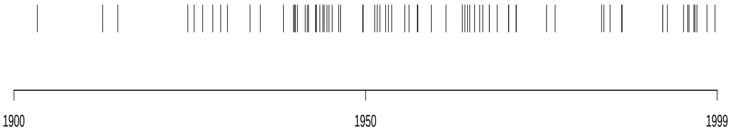
Pattern of 556 points in time window $(0, 500)$ generated under a Pareto inter-arrival distribution (clustering).



Posterior mean and 95% interval estimates for the tail of the inter-arrival density (left), and the K function (right). The black lines denote the true functions.

Earthquake data

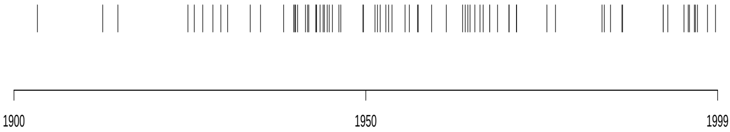
- Data on occurrence times of 76 major earthquakes (magnitude > 5.5 on the Richter scale) observed during the 20th century in North Anatolia, Turkey.



- Posterior inference results suggest clustering of the earthquake occurrences.

Earthquake data

- Data on occurrence times of 76 major earthquakes (magnitude > 5.5 on the Richter scale) observed during the 20th century in North Anatolia, Turkey.



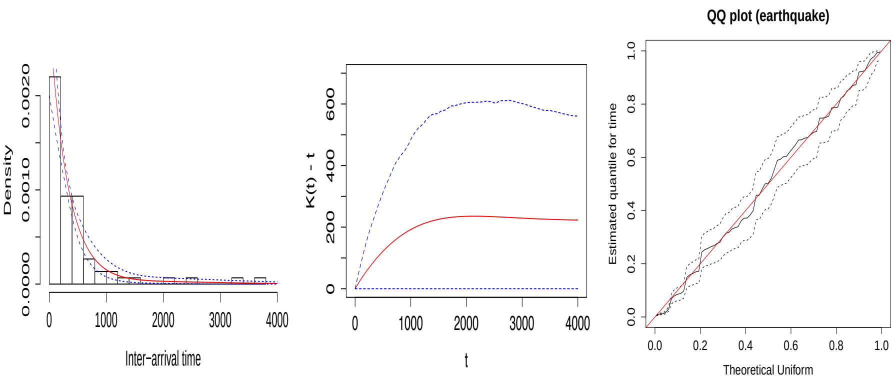
- Posterior inference results suggest clustering of the earthquake occurrences.

- Model checking through the time-rescaling theorem: if $t_1 < \ldots < t_n$ is a realization from the renewal process, the variables

$$\eta_k = -\log\{1 - \int_{t_{k-1}}^{t_k} f(u - t_{k-1}) \mathrm{d}u\}, \quad k = 1, \ldots, n$$

are i.i.d. Exp(1), and thus $1 - \exp(-\eta_k)$ are i.i.d. $U(0,1)$.

Earthquake data



Posterior mean and 95% interval estimates for the inter-arrival density (left), the K function (middle), and the Q-Q plot for model checking.

Extension to modeling for IMI processes

- Inhomogeneous Markov interval (IMI) processes can be viewed as extensions of renewal processes where the inter-arrival density depends on both the current time and the last arrival time — extend $f(t_i - t_{i-1})$ to $f(t_i \mid t_{i-1})$.

Extension to modeling for IMI processes

- Inhomogeneous Markov interval (IMI) processes can be viewed as extensions of renewal processes where the inter-arrival density depends on both the current time and the last arrival time — extend $f(t_i - t_{i-1})$ to $f(t_i \mid t_{i-1})$.

- Now, the likelihood becomes

$$\left\{1 - \int_{t_n}^{T} f_{T_{n+1}}(u \mid t_n)\mathrm{d}u\right\} \prod_{i=1}^{n} f_{T_i}(t_i \mid t_{i-1})$$

Extension to modeling for IMI processes

- Inhomogeneous Markov interval (IMI) processes can be viewed as extensions of renewal processes where the inter-arrival density depends on both the current time and the last arrival time — extend $f(t_i - t_{i-1})$ to $f(t_i \mid t_{i-1})$.

- Now, the likelihood becomes

$$\left\{1 - \int_{t_n}^{T} f_{T_{n+1}}(u \mid t_n)\mathrm{d}u\right\} \prod_{i=1}^{n} f_{T_i}(t_i \mid t_{i-1})$$

- Extensions of the Erlang mixture model to incorporate dependence on last arrival time either in the mixture weights (through a structured dependent DP prior) or in the scale parameter θ (through a Gaussian process prior).

Manuscripts

- Xiao, S., Kottas, A., and Sansó, B. "Nonparametric Bayesian modeling and inference for renewal processes." Draft manuscript.
- Xiao, S., Kottas, A., and Sansó, B. "Bayesian semiparametric modeling approaches for inhomogeneous Markov interval processes." In preparation.

Acknowledgment: Funding from *NSF, Methodology, Measurement, and Statistics* and *NSF, Statistics* programs.

MANY THANKS !!!