

**Gasbarra, Dario, Bayesian inference for models based on point processes,  
by using Markov chain methods.**

Department of Mathematical Sciences, University of Oulu, Linnanmaa FIN-90571  
Oulu, Finland

*Acta Univ. Oul. A 302, 1997*

Oulu, Finland

(Received 13 November 1997)

**Abstract**

This work consists of an introduction to Bayesian inference and Markov chain sampling methods for point processes, and four applied original papers. The first two papers study the nonparametric Bayesian estimation of intensities under different assumptions. The third paper is on the implementation of the predictive-sequential approach in model-checking to counting processes. The fourth paper is concerned with the Bayesian estimation of a cumulative hazard process and a probability-valued regression function.

*Keywords:* survival analysis, Markov chain Monte Carlo, nonparametric Bayesian statistics.

## Foreword

I would like to say Thanks to those who have been determinant in the course of this work.

To my supervisor, Professor Elja Arjas, for his enthusiasm and patience, in proposing, discussing, and sometimes accepting, all sorts of ideas.

To Mr. Pekka Kangas, who introduced me to several computer systems and packages, he was always ready to give a hand.

To Dr. Sangita Karia, I do not know how I would have completed this thesis without her collaboration.

I wish also to thank Antti Penttinen and Professor Esa Uusipaikka who read and commented on the manuscript, Arnaldo Frigessi who kindly has accepted to be my opponent in the public defense, and all my colleagues at the Department of Mathematical Sciences of the University of Oulu and at the Rolf Nevanlinna Institute, where this study was carried out.

The financial support of the Academy of Finland and the University of Helsinki is gratefully acknowledged.

A different kind of support was given by my beloved Margit-Kristi, with love and understanding. Thank you my darling.

I dedicate this work to my parents.

Helsinki, November 1997

Dario Gasbarra

## List of original papers

- I. Arjas E & Gasbarra D (1994) Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statistica Sinica* 4: 505-524.
- II. Arjas E & Gasbarra D (1996) Bayesian inference of survival probabilities, under stochastic ordering constraints. *JASA* 91: 1101-1109.
- III. Arjas E & Gasbarra D (1997) On prequential model assessment in life history analysis. *Biometrika* 84: 505-522.
- IV. Gasbarra D & Karia S (1997) Analysis of competing risks by using Bayesian smoothing. Submitted for publication.

## Contents

1. Introduction .....	11
1.1. Bayes' formula .....	11
1.2. Markov chains .....	12
1.3. Metropolis-Hastings transition kernels .....	14
1.4. Marked point processes on the real line .....	15
1.5. Change of measure .....	19
1.6. Smoothing Lévy process priors .....	20
1.7. Personal views and perspectives .....	24
2. Summaries of original papers .....	25
3. References .....	27

*“Only a subtle layer between trivial and unattainable things exists at every given time. It is in this layer that mathematical discoveries are being made of. That is why an ordered applied problem has for the most part either a trivial solution or none at all.” ( A.N. Kolmogorov, 14.09.1943, diary remarks)*

# 1. Introduction

This dissertation does not contain mathematical discoveries. It is an applied work based on two formulæ. One is Bayes' formula, the other one is the definition of the Metropolis-Hastings transition kernel. We will see how a class of unattainable computations become possible, if not trivial, and give some nonparametric Bayesian applications in survival analysis.

The following sections contain some preliminaries in order to illustrate the theoretical background, to point out some important details, and to understand better the spirit of the original papers.

## 1.1. Bayes' formula

Bayes' formula tells how a completely specified probabilistic model (a coherent representation of uncertainty) updates itself in the light of new information (when uncertainty is partially removed).

To make things clear, we start with the abstract formulation.

**Theorem** [Kallianpur & Striebel, can be found in Brémaud, 1981]

Let  $(\Omega, \mathcal{F}, Q)$  be a probability space, and let  $P$  be another probability measure on the same space such that  $P \ll Q$ . Then for any  $\sigma$ -algebra  $\mathcal{G} \subseteq \mathcal{F}$  and for any  $P$ -integrable function  $f$  the following holds  $P$ -almost surely:

$$E_P[f | \mathcal{G}](\omega) = \frac{E_Q[ f \frac{dP}{dQ} | \mathcal{G} ](\omega)}{E_Q[ \frac{dP}{dQ} | \mathcal{G} ](\omega)} \quad (\text{Bayes' formula}) \quad (1)$$

Proof: apply a few times the definition of conditional expectation.

In other words, one way to compute the conditional expectation  $E_P[f | \mathcal{G}]$  under the probability  $P$ , is to compute conditional expectations under a suitable

reference probability  $Q$ .

In Bayesian analysis usually one deals with the following structure:  $\Omega = \Theta \times \Xi$ ,  $\mathcal{F} = \mathcal{B}_\Theta \otimes \mathcal{B}_\Xi$ , and the reference measure is a product measure  $Q(d\theta, dx) = \pi(d\theta)\lambda(dx)$  while the measure of interest admits the disintegration  $P(d\theta, dx) = \pi(d\theta)P_\theta(dx)$ , with  $P_\theta \ll \lambda$  for  $\pi$ -almost all  $\theta$ . Let the random variable  $X(\omega)$  be the projection  $X(\theta, x) = x$ . For  $\mathcal{G} = \sigma(X)$ , Bayes' formula assumes its classical form:

$$E_P[f(\theta, X) | X](\omega) = \frac{\int_{\Theta} f(\theta, X(\omega)) \frac{dP_\theta}{d\lambda}(X(\omega)) \pi(d\theta)}{\int_{\Theta} \frac{dP_\theta}{d\lambda}(X(\omega)) \pi(d\theta)} \quad (2)$$

The denominator on the right hand side does not depend on  $f$ , and the regular conditional probability

$$\frac{dP_\theta}{d\lambda}(X) \pi(d\theta) \delta_X(dx) \left( \int_{\Theta} \frac{dP_\theta}{d\lambda}(X) \pi(d\theta) \right)^{-1} \quad (3)$$

is the *posterior measure*. The random variable  $\theta$  can be interpreted as the unobservable part of the model and  $X$  as the observable part;  $\theta$  appears also as a mixing parameter in the marginal distribution of  $X$ , with mixing distribution  $\pi$  (the prior). When  $X$  is observed, the distribution of the unobservable  $\theta$  is updated from prior to posterior according to formula (2).

Remark: Although all the models considered in this thesis are dominated, i.e. in each situation we have a disintegration and a measure  $\lambda \gg P_\theta$  for  $\pi$ -almost all  $\theta$ , in general it is not always the case that such a measure exists. If not, the posterior measure is not necessarily absolutely continuous with respect to the prior on  $(\Theta, \mathcal{B}_\Theta)$ . One cannot write (2) as such, but has to work out (1). (For a discussion see the monograph of Florens *et al.* 1990, chapter 1).

In this thesis typically  $\theta$  and  $X$  will be marked point processes.

In order to evaluate the posterior expectation by using (2), one needs to integrate  $f$  under a probability measure (the posterior). The evaluation of such integrals has been a major problem in Bayesian theory.

## 1.2. Markov chains

We briefly recall a few definitions and results. We refer to Tierney (1994,1995,1996), and the monographs written by Nummelin (1984) and by Meyn & Tweedie (1993).

Let  $(S, \mathcal{S})$  and  $(Z, \mathcal{Z})$  be two measurable spaces. A *(stochastic) transition kernel* from  $(S, \mathcal{S})$  to  $(Z, \mathcal{Z})$  is a mapping  $K(\cdot, \cdot) : S \times \mathcal{Z} \rightarrow [0, 1]$ , such that

(i)  $K(\cdot, A)$  is  $\mathcal{S}$ -measurable for every  $A \in \mathcal{Z}$

(ii)  $K(x, \cdot)$  is a measure (a probability measure) on  $(Z, \mathcal{Z})$  for every  $x \in S$ .

A stochastic transition kernel  $K$  from  $(S, \mathcal{S})$  to  $(Z, \mathcal{Z})$ , together with a probability measure  $\mu$  on  $(S, \mathcal{S})$  define a joint probability measure  $P(dx, dz) = \mu(dx)K(x, dz)$  on  $(S \times Z, \mathcal{S} \otimes \mathcal{Z})$ .

In what follows it will be always the case that  $S = Z$  and  $\mathcal{S} = \mathcal{Z}$ ; then the previous argument can be extended inductively to construct a probability measure  $P^\infty$  on the space of sequences  $(S^\infty, \mathcal{S}^{\otimes \infty})$  (proofs can be found in Gikhman & Skorohod, 1972).

A random process  $(X_n)$  with values in  $(S, \mathcal{S})$  following such a distribution is called a *Markov chain* with initial distribution  $\mu$  and transition probability  $K$ .

**Definition:** A measure  $\pi$  is *stationary* for the transition kernel  $K$  if  $\pi K = \pi$ , where

$$(\pi K)(A) := \int_S K(x, A)\pi(dx) \quad (4)$$

Our interest will focus on the construction of Markov chains with a given stationary probability measure  $\pi$ . It follows that if the initial probability  $\mu$  is stationary for the kernel  $K$ , then the Markov chain  $(X_n)$  is a stationary process, i.e. its finite dimensional distributions are invariant under time-shifts.

For a Markov chain with stationary distribution  $\pi$ , just one more condition is needed in order to ensure that for any  $\pi$ -integrable function  $f$ , regardless of the initial distribution  $\mu$ , a strong law of large numbers holds:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i(\omega)) = E_\pi(f) \text{ with probability } 1. \quad (5)$$

Then the convergence of the sample path averages can be used in practice to evaluate the expectation  $E_\pi(f)$  from a single realization of the Markov chain.

To get the law of large numbers (5), it is necessary that almost every realization of the Markov chain visits the whole support of  $\pi$ . This property can be formalized as follows:

**Definition** A Markov chain is  *$\varphi$ -irreducible* for a measure  $\varphi$  on  $(S, \mathcal{S})$  when for every  $A \in \mathcal{S}$  such that  $\varphi(A) > 0$  and for all  $x \in S$

$$P_x(X_n \in A \text{ for some } n) > 0 \quad (6)$$

where  $P_x$  is the distribution of the Markov chain started at  $x$ .

This is quite a weak condition. For example, it is sufficient that for some  $n$  the iterated transition probabilities  $K^n(x, \cdot)$  are dominated by  $\varphi$  for all  $x$  (define  $K^1 := K$ ,  $K^n(x, A) := \int_S K(z, A)K^{n-1}(x, dz)$ ).

**Proposition** (can be found in Tierney, 1996): Suppose the Markov chain  $(X_n)$  is  $\varphi$ -irreducible for some measure  $\varphi$  and it has a stationary distribution  $\pi$ .

Then the chain is irreducible w.r.t.  $\pi$ ,  $\pi \gg \varphi$ ,  $\pi$  is the unique stationary distribution and (5) holds.

*Recipe: Markov chain Monte Carlo*

In order to compute  $E_\pi(f)$ , first find a transition kernel  $K(x, dy)$  such that  $\pi$  is stationary for  $K$ , check the irreducibility condition, choose any initial distribution you like, then just take one realization of the Markov chain and apply the law of large numbers (5).

To proceed with this plan, we need to find the stochastic kernel  $K$ .

### 1.3 Metropolis-Hastings transition kernels

**Definition:** A transition kernel  $K(x, A)$  is *reversible* with respect to a measure  $\pi$  when the operator  $T$  defined on  $L^2(S, \pi)$  by

$$(Tf)(x) = \int_S f(y)K(x, dy) \quad (7)$$

is selfadjoint.

Equivalently, on the product space  $(S \times S, \mathcal{S} \otimes \mathcal{S})$ , the *forward measure*

$$P(dx, dy) := \pi(dx)K(x, dy) \quad (8)$$

is equal to its transpose

$$P^\top(dx, dy) := \pi(dy)K(y, dx) \text{ (reverse measure)}. \quad (9)$$

It is easy to check that if the kernel  $K$  is reversible w.r.t. the measure  $\pi$ , then  $\pi$  is stationary w.r.t.  $K$  (the opposite implication is, however, not true).

Given a probability  $\pi(dx)$  and a *proposal transition kernel*  $Q(x, dy)$ , there is a simple and general way to modify  $Q$  in order to obtain a kernel  $K$  reversible w.r.t.  $\pi$ . This was first established by Metropolis, and later extended and elaborated by other authors, beginning with Hastings (1970). In particular  $\pi$  will be stationary w.r.t.  $K$ . The irreducibility of the derived kernel  $K$  has to be checked case by case.

Note first that the identical kernel  $\delta_x(dy)$  is trivially reversible.

Now we proceed with a very simple idea: try a kernel of the form

$$K(x, dy) = a(x, y)Q(x, dy) + \left[1 - \int_S a(x, z)Q(x, dz)\right] \delta_x(dy) \quad (10)$$

where  $a : S \times S \rightarrow [0, 1]$  is a jointly measurable function we are allowed to choose.

It means that being at a generic time  $n$  in state  $X_n$ , we sample first a candidate state  $Y$  from the proposal distribution  $Q(X_n, \cdot)$ , and then with probability  $a(X_n, Y)$  we assign  $X_{n+1} = Y$ , otherwise  $X_{n+1} = X_n$ .

We would like to define the acceptance function  $a$  in order to satisfy the reversibility condition

$$\pi(dx)Q(x, dy)a(x, y) = \pi(dy)Q(y, dx)a(y, x) \quad (11)$$

Define on the product space  $(S \times S, \mathcal{S} \otimes \mathcal{S})$  the measures  $\rho(dx, dy) := \pi(dx)Q(x, dy)$   $\rho^\top(dx, dy) := \rho(dy, dx)$ . Consider the symmetric set  $D \in \mathcal{S} \otimes \mathcal{S}$  given by

$$\begin{aligned} & \{(x, y) \in S \times S : \frac{d\rho}{d(\rho + \rho^\top)}(x, y) > 0 \text{ and } \frac{d\rho^\top}{d(\rho + \rho^\top)}(x, y) > 0\} \\ = & \{(x, y) \in S \times S : 0 < \frac{d\rho}{d\rho^\top}(x, y) < \infty\} \end{aligned}$$

The Radon-Nikodym derivative  $\frac{d\rho^\top}{d\rho}(x, y)$  is called the *Hastings' ratio*. Condition (11) is satisfied by defining

$$\begin{aligned} a(x, y) &= \min\left(1, \frac{d\rho^\top}{d\rho}(x, y)\right) \text{ for } (x, y) \in D, \text{ and} \\ a(x, y) &= 0 \text{ for } (x, y) \notin D. \end{aligned}$$

In a dominated situation, i.e. when there is a measure  $\lambda$  on  $(S, \mathcal{S})$  such that  $dQ(x, \cdot) \ll d\lambda$  for  $\pi$ -almost all  $x$ , the Hastings-ratio can be expressed as a product of two likelihood ratios on  $S$ . Nevertheless we should not be too puzzled in the non-dominated case, just keep in mind that the Hastings ratio is defined as a likelihood ratio on the product space  $S \times S$ .

#### 1.4. Marked point processes on the real line

Let  $(E, \mathcal{E})$  be a separable metric space equipped with the  $\sigma$ -algebra of Borel sets.

We define the canonical space of marked point process histories  $H$  as the space of sequences  $h = \{(t_i, x_i)\}_{i \geq 0}$  such that  $0 \leq t_0 < t_1 < t_2 < \dots, x_i \in E$ , and endow it with the  $\sigma$ -algebra  $\mathcal{H}$  generated by the projections  $T_n(h) = t_n, X_n(h) = x_n$ .  $T_n$  are interpreted as event times, and the marks  $X_n$  describe the corresponding events.

A marked point process on the real line is a measurable map from a probability space  $(\Omega, \mathcal{F}, P)$ , into the canonical history space  $(H, \mathcal{H})$ . As usual this induces the image probability measure on  $(H, \mathcal{H})$ . We speak of canonical realization of the m.p.p. when  $\Omega = H$  and the map is the identity.

Define the counting processes  $N(\omega, t, A) = \sum_{n=1}^{\infty} 1_A(X_n(\omega))1_{[0,t]}(T_n(\omega))$ . Introduce on  $\Omega$  the filtration given by the  $\sigma$ -algebrae  $\mathcal{F}_t = \sigma(N_s(B) : s \leq t, B \in \mathcal{E})$ . We complete the  $\sigma$ -algebrae  $\mathcal{F}$  and  $\mathcal{F}_t$ ,  $t \geq 0$ , by the sets of  $P$ -measure zero. This technical step is to ensure that  $P$ -modifications of adapted processes are adapted (see Jacod & Shiryaev 1987, chapter 1). The predictable  $\sigma$ -field  $\mathcal{P}$  on  $[0, \infty) \times \Omega$  is defined as the  $\sigma$ -field generated by the sets  $\{(\omega, t) : f(\omega, t) \in B\}$  where  $f$  are  $\{\mathcal{F}_t\}$ -adapted, left-continuous, real valued processes, and  $B$  are Borel sets. Predictable processes are  $\mathcal{P}$ -measurable functions on  $[0, \infty) \times \Omega$ .

The Doob-Meyer decomposition theorem (see Jacod & Shiryaev 1987, chapter 1) tells that up to indistinguishability there is a unique predictable increasing process  $\Lambda(\omega, t, A)$  such that  $N(\omega, t, A) - \Lambda(\omega, t, A)$  is a  $(P, \{\mathcal{F}_t\})$ -martingale.  $\{\Lambda(\omega, t, A)\}$  is called the  $(P, \{\mathcal{F}_t\})$ -compensator of the counting process  $\{N(\omega, t, A)\}$ . Since  $\Delta N(\omega, t, E) \leq 1$ , it follows that  $\Delta \Lambda(\omega, t, E) \leq 1$   $P$ -almost surely.

Note that a compensator  $\Lambda(\omega, t, dx)$  is a kernel such that

- (i) for each fixed  $A$ ,  $\Lambda(\cdot, \cdot, A)$  is a predictable increasing process,
- (ii)  $\Delta \Lambda(\omega, t, E) \leq 1$ ,
- (iii) for fixed  $\omega, t$ ,  $\Lambda(\omega, t, \cdot)$  is a positive measure on the mark space  $(E, \mathcal{E})$ .

If the compensator  $\Lambda(\omega, dt, A)$  is absolutely continuous w.r.t. the Lebesgue measure on the real line for each  $A \in \mathcal{E}$ , we define the *intensity measure process* as

$$\lambda(t, \omega, A) := \frac{d\Lambda([0, t], \omega, A)}{dt} \quad (12)$$

The intensity process is a nonnegative predictable process, and it has the following interpretation:

$$\lambda(t, \omega, dx) = \lim_{h \rightarrow 0} \frac{P(N_{t+h}(dx) - N_t(dx) > 0 \mid \mathcal{F}_t)(\omega)}{dt}. \quad (13)$$

A useful expression is

$$\lambda(\omega, t, dx) = \lambda_t(\omega) \Phi_t(\omega, dx) \quad (14)$$

where  $\lambda_t(\omega) = \lambda(\omega, t, E)$  and  $\Phi_t(\omega, dx) = \lambda(\omega, t, dx) / \lambda(\omega, t, E)^{-1}$  is a probability measure on  $(E, \mathcal{E})$ . It follows that  $\Phi_{T_n}(\omega, A) = P(X_n \in A \mid \mathcal{F}_{T_n-})(\omega)$  (see Brémaud 1981, chapter VIII).

The Doob-Meyer decomposition's existence theorem has the following counterpart for marked point processes:

**Theorem** Given a kernel  $\Lambda(\omega, t, dx)$  defined on a canonical history space  $\Omega = H$  satisfying (i), (ii), (iii), there is a unique probability distribution  $P$  on  $\Omega$  such that  $\{\Lambda\}$  is the  $(P, \{\mathcal{F}_t\})$ -compensator of the marked point process  $\{N\}$ . The construction is given in Jacod (1975).

Typically we will assign a probability measure  $P$  to the m.p.p. by giving its compensator (its intensity measure, when we assume that the compensator is absolutely continuous).

Consider now the following situation: the mark space is the union of two disjoint sets  $E = \hat{E} \cup \tilde{E}$ , where the marked points with mark in  $\hat{E}$  are observable and those with mark in  $\tilde{E}$  are unobservable. In other words, the information available to the observer is carried by the smaller filtration  $\{\mathcal{G}_t\}$ , where  $\mathcal{G}_t = \sigma(N_s(B) : s \leq t, B \subseteq \hat{E}) \subset \mathcal{F}_t$ .

Our goal will be to evaluate the conditional expectation  $E_P[f | \mathcal{G}_t](\omega)$  of some  $P$ -integrable function  $f$  on  $\Omega$  for fixed  $t$  and  $\omega$ .

At this point we take a brief technical digression:

**Proposition** (Disintegration theorem, can be found in Dellacherie & Meyer 1975, chapter III)

Let  $\pi$  be a probability measure on a separable metric space  $(S, \mathcal{S})$ , let  $T$  a measurable map from  $S$  into  $(U, \mathcal{U})$ . Let  $\pi T^{-1}$  be the image probability measure on  $(U, \mathcal{U})$ . If  $\mathcal{U}$  is countably generated and contains all the singletons  $\{u\}$ , then  $\pi$  has a  $T$ -disintegration, i.e. there are measures  $\{\pi_u : u \in T(S)\}$ , such that for  $\pi T^{-1}$ -almost all  $u$ ,  $\pi_u$  is a probability measure concentrated on  $\{x : T(x) = u\}$  such that for any  $f \in L^1_\pi(S)$  we have

$$(i) \quad \int_S f(x) \pi_u(dx) = E_\pi(f | T = u), \text{ and}$$

$$(ii) \quad \int_S f(x) \pi(dx) = \int_U \left( \int_S f(x) \pi_u(dx) \right) \pi T^{-1}(du).$$

In particular  $\pi_u(dx)$  is a regular version of the conditional probability w.r.t.  $\sigma(T)$ .

Since the mark space  $E$  is assumed to be a separable metric space, the canonical history space  $H$  endowed with the so-called Skorohod's topology becomes a separable metric space and the Borel  $\sigma$ -algebra coincides with the  $\sigma$ -algebra generated by the projections (see e.g. Pollard 1984, chapters V, VI).

This is important because then we can apply the disintegration theorem to show the existence of a regular version of the conditional probability. In our case,  $T$  is the projection map from the canonical history space  $H$  into  $\hat{H}_t = \{(t_i, x_i) : 0 \leq t_i \leq t, x_i \in \hat{E}\}$ , and  $\sigma(T) = \mathcal{G}_t$ .

In other words, for our point processes, posterior distributions  $P(d\omega | \mathcal{G}_t)$  will always exist.

When  $t$  varies over  $(0, \infty)$ , Kolmogorov's extension theorem can be applied to the family of the posterior distributions  $\{P(d\omega | \mathcal{G}_t)\}_{t \geq 0}$  to form a probability measure valued process on the space of histories  $H$ . A deeper result due to

Aldous (unpublished manuscript; see Norros 1985), concerns the existence of a version  $\{\hat{\pi}_t(d\omega)\}$  of the posterior probability process which is right continuous with left limits with respect to the topology of weak convergence of measures, and jumps together with the observed process  $\{N(t, \omega, \hat{E})\}$ . These facts are discussed in Norros (1985), Arjas *et al.* (1992). These authors have studied the filtering problem dynamically in time, deriving recursive equations and solutions for the prediction process.

At this point, we shall take a rather different approach and approximate, by Markov chain Monte Carlo, the posterior distribution of the m.p.p. given the information carried by the observed part  $\{N(s, \omega, \hat{E})\}_{(s \leq t)}$ , up to a fixed final time  $t$ .

It is somehow troublesome to write down an explicit general expression for the Hastings-ratio when the state space is a space of m.p.p. histories, without making more precise assumptions on the structure of both the m.p.p. and the proposal kernel.

Geyer and Møller (1994) give a “birth and death” version of the Metropolis-Hastings algorithm for a point process which admits a density with respect to the Poisson measure. This is simple and general enough for many applications. Similar versions of the Metropolis algorithm have been used in the original papers forming this thesis.

Recall that if  $\mu$  is a finite measure on a locally compact Hausdorff space  $S$ , then the  $\mu$ -Poisson measure on the space of point configurations  $S^\infty$  is

$$\text{Poisson}_\mu(dx) = \sum_{k=0}^{\infty} \frac{\exp(-\mu(S))}{k!} \delta_{k, n(x)} \mu(ds_1) \otimes \dots \otimes \mu(ds_k). \quad (15)$$

This means that, under the Poisson measure, the number of points  $n(X)$  is a  $\mu(S)$ -Poisson random variable, and conditionally on  $n(X)$ , the points  $S_1, \dots, S_{n(X)}$  are i.i.d. r.v.'s with distribution  $\mu(ds)/\mu(S)$  on  $(S, \mathcal{S})$ . The definition is extended also to the  $\sigma$ -finite driving measures  $\mu$ .

We briefly describe Geyer and Møller’s construction. Let  $X$  be a m.p.p. with density  $p(x)$  with respect to the  $\mu$ -Poisson process, where  $\mu$  is a measure on  $S = (0, \infty) \times E$ . Consider the following proposal transition kernel  $Q(x, dy)$ . Starting from a generic configuration of the marked point process, say  $x = \{s_1, \dots, s_k\}$ , with probability 1/2 add to the configuration one random marked point  $s_{k+1}$  sampled from a probability distribution  $q(\{s_1, \dots, s_k\}; s) \mu(ds)$ , proposing the configuration  $y = \{s_1, \dots, s_k, s_{k+1}\}$  (upstep). Here, for a configuration  $\{s_1, \dots, s_k\}$ , we denote by  $q(\{s_1, \dots, s_k\}; \cdot)$  the probability density of the proposed new-born marked point, w.r.t. the measure  $\mu$  on  $S$ . For example, if  $\mu(S) < \infty$ , we may use the constant density  $q \equiv \mu(S)^{-1}$ .

Otherwise, if  $n(x) > 0$ , delete the last marked point  $s_k$ , proposing the configuration  $y = \{s_1, \dots, s_{k-1}\}$  (downstep).

It is not too difficult to check that the Hastings ratio is respectively

$$\begin{aligned} \frac{p(y)}{p(x)} \frac{1}{(n(x) + 1)q(\{s_1, \dots, s_k\}; s_{k+1})} & , \text{ for the upstep,} \\ \text{and } \frac{p(y)}{p(x)} n(x)q(\{s_1, \dots, s_{k-1}\}; s_k) & , \text{ for the downstep.} \end{aligned}$$

Now consider a transition kernel given by a random transposition of  $(s_1, \dots, s_k)$ , which switches the last point  $s_k$  with  $s_I$ , where  $I$  is a random index sampled uniformly in  $\{1, \dots, k\}$ . The target distribution  $p(x)\text{Poisson}_\mu(dx)$  is invariant for this transition kernel, just because a point configuration is the same configuration after a permutation of its points. Therefore, we may as well compose the ‘‘birth and death’’ Metropolis Hastings with a random transposition and propose for deletion a randomly chosen marked point instead of the last one.

Note that in order to implement the algorithm, it is enough to know the density  $p(x)$  up to a normalizing factor. This is an useful feature of the Metropolis-Hastings algorithm.

### 1.5. Change of measure.

Let  $P$  and  $P'$  be two probability measures on the space of m.p.p histories  $(H, \mathcal{H})$ , and let  $(\Lambda)$ ,  $(\Lambda')$  be the corresponding compensators with respect to the canonical filtration generated by the m.p.p. Then  $P' \ll P$  iff  $\Lambda'(\omega, dt, dx) \ll \Lambda(\omega, dt, dx)$  as random measures on  $(0, \infty) \times E$  for  $P$ -almost every  $\omega$ , and  $\Delta\Lambda(\omega, t, A) = 1$  implies  $\Delta\Lambda'(\omega, t, A) = 1$   $P$ -almost surely. In such a case, the following expression holds for the likelihood ratio (known as Jacod’s formula, given in Jacod, 1975):

$$\begin{aligned} L_t &= \frac{dP' |_{\mathcal{F}_t}}{dP |_{\mathcal{F}_t}} \\ &= e^{-\Lambda'(\omega, t, E) + \Lambda(\omega, t, E)} \prod_{s \leq t, s \neq T_i} \frac{(1 - \Delta\Lambda'(\omega, s, E))}{(1 - \Delta\Lambda(\omega, s, E))} \prod_{T_i \leq t} \frac{d\Lambda'}{d\Lambda}(\omega, T_i, X_i), \quad (16) \end{aligned}$$

where we denote by  $P |_{\mathcal{F}_t}$  the restriction of the probability measure  $P$  to the  $\sigma$ -algebra  $\mathcal{F}_t$ . If  $\Lambda(\omega, dt, dx) = \lambda_t(\omega)dt \phi_t(\omega, x)\rho(dx)$   $P$ -almost surely, for a fixed measure  $\rho$  on  $(E, \mathcal{E})$ , then  $P |_{\mathcal{F}_t}$  is absolutely continuous with respect to the Poisson

process on  $[0, t) \times E$ , driven by the measure  $dt \otimes \rho(dx)$ , and (16) can be written as

$$\frac{dP' |_{\mathcal{F}_t}}{dP |_{\mathcal{F}_t}} = \prod_{T_i \leq t} \left[ \frac{\lambda'_{T_i} \phi'_{T_i}(X_i)}{\lambda_{T_i} \phi_{T_i}(X_i)} \right] \exp \left( - \int_0^t \lambda'_s ds + \int_0^t \lambda_s ds \right). \quad (17)$$

Now suppose we have a m.p.p. and we have specified the distribution  $P$  by assigning as its intensity a given non-negative predictable process  $\lambda(\omega, t, A)$ . Suppose the mark space is the union  $E = \hat{E} \cup \tilde{E}$  of disjoint observable and unobservable parts. Denote by  $\hat{H}$  and  $\tilde{H}$  the respective observable and unobservable histories. By construction, these history processes have a density w.r.t. to Poisson processes defined respectively on  $(0, t) \times \hat{E}$  and  $(0, t) \times \tilde{E}$ . In other words, the joint measure is proportional to

$$p(\hat{H}_t, \tilde{H}_t) \text{Poisson}(d\hat{H}_t) \text{Poisson}(d\tilde{H}_t), \quad (18)$$

where the joint density is given by Jacod's formula:  $p(\hat{H}_t, \tilde{H}_t) =$

$$\prod_{i: T_i \leq t, \hat{X}_i \in \hat{E}} \hat{\lambda}(T_i | \hat{H}_{T_i-}, \tilde{H}_{T_i-}) \phi_{T_i}(\hat{X}_i | \hat{H}_{T_i-}, \tilde{H}_{T_i-}) e^{-\int_0^\infty \hat{\lambda}(s, \hat{E} | \hat{H}_{s-}, \tilde{H}_{s-}) ds} \times \\ \prod_{i: T_i \leq t, \tilde{X}_i \in \tilde{E}} \tilde{\lambda}(T_i | \hat{H}_{T_i-}, \tilde{H}_{T_i-}) \phi_{T_i}(\tilde{X}_i | \hat{H}_{T_i-}, \tilde{H}_{T_i-}) e^{-\int_0^\infty \tilde{\lambda}(s, \tilde{E} | \hat{H}_{s-}, \tilde{H}_{s-}) ds}.$$

By Bayes' formula (2), it follows that up to a constant factor

$$P(d\tilde{H}_t | \hat{H}_t) \propto p(\hat{H}_t, \tilde{H}_t) \text{Poisson}(d\tilde{H}_t). \quad (19)$$

Therefore, the birth and death algorithm of Geyer and Møller can be applied directly to sample the unobserved history  $\tilde{H}_t$  from the posterior distribution.

A simple example is contained in the original paper [I], where the intensity rate of the observed failure process  $\hat{H}$  is the piecewise constant process

$$\hat{\lambda}(\omega, t, \mathcal{F}_{t-}) = \sum_i \mathbf{1}_{(\tilde{T}_i(\omega), \tilde{T}_{i+1}(\omega)]}(t) \tilde{\lambda}_i(\omega) R(t-, \omega, \hat{H}_{t-}), \quad (20)$$

where  $R(t-, \omega, \hat{H}_{t-}) = \#\{\text{individuals at risk just before time } t\}$ , and the marked points  $(\tilde{T}_i, \tilde{\lambda}_i)$  form the unobserved m.p.p.  $(\tilde{N}_t)$ . Under the prior distribution, the sequence  $(\tilde{T}_i)$  forms a standard Poisson process and the sequence  $(\tilde{\lambda}_i)$  is Markovian, having a density with respect to the Lebesgue measure.

## 1.6. Smoothing Lévy process priors.

An alternative and elegant way to model the cumulative hazard  $\Lambda_t(\omega)$  is by using a positive Lévy process  $(X_t)$ , which is a process with non-negative independent increments.

Processes with independent increments were first studied by De Finetti, and the main results were found by Lévy and Khintchine (see Lévy 1957, chapter 5). On this subject, we refer to the monograph of Kingman (1993), chapters 3,8,9. For a complete exposition on Lévy processes, see the books by Feller (1971), and Skorohod (1991).

When the increments are non-negative, the Lévy-Khintchine representation formula for the Laplace transform of the random measure  $X_t = X((-\infty, t])$  can be written as

$$E\left(\exp\left\{-\int_{-\infty}^t f(y)dX(y)\right\}\right) = \exp\left\{-\int_{-\infty}^t f(y)d\mu(y) + \int_{-\infty}^t \int_0^{\infty} (e^{-zf(y)} - 1)\frac{1}{z}L(dz, dy)\right\}, \quad (21)$$

where  $\mu$  is a measure on  $(-\infty, +\infty)$ , and  $L(dz, dy)$  is a measure on  $(-\infty, +\infty) \times (0, \infty)$  which is called the Lévy measure of the process  $(X_t)$ .

It follows that  $X_t$  is finite with probability one if and only if  $\mu((-\infty, t]) < \infty$  and

$$\int_{-\infty}^t \int_0^{\infty} \min(z, 1)\frac{1}{z}L(dz, dy) < \infty, \quad (22)$$

otherwise  $X_t = \infty$  almost surely. For example, a gamma process driven by the measure  $d\alpha$  with scale parameter  $\beta$  has increments  $(X_t - X_s)$  which are independent on disjoint intervals, with distribution gamma( $\alpha([s, t]), \beta$ ); the Laplace transform is

$$E\left(\exp\left\{\int_s^t f(y)X(dy)\right\}\right) = \exp\left\{-\int_s^t \int_0^{\infty} (e^{-f(y)\beta z} - 1)\frac{1}{z}e^{-z}dz \otimes \alpha(dy)\right\}, \quad (23)$$

where the Lévy measure is given by  $L(dz, dy) = e^{-z/\beta}dz \otimes \alpha(dy)$ .

Formula (21) has the following interpretation: the process  $(X_t)$  can be represented as the sum of a deterministic increasing component  $\mu((-\infty, t])$  and a compound Poisson process. In fact, let  $\xi_t = \{(Z_i, Y_i) : i = 1, \dots, N(\xi)\}$  be a Poisson process on  $(0, \infty) \times (-\infty, t)$  driven by the measure  $z^{-1}L(dz, dy)$ . Note that the number of points  $N(\xi)$  in  $(-\infty, t) \times (0, \infty)$  is possibly infinite and  $N(\xi) < \infty$  almost surely if and only if  $\int_{-\infty}^t \int_0^{\infty} z^{-1}L(dz, dy) < \infty$ . Consider the compound Poisson process

$$V_t = V((-\infty, t]) := \sum_{i=1}^{N(\xi)} Z_i \delta_{Y_i}. \quad (24)$$

Even in the case where  $N(\xi) = \infty$  a.s., under condition (22) the series  $\sum_{i=1}^{\infty} Z_i$  is still almost surely convergent. By Campbell's theorem (in Kingman 1993, chapter 3),

the Laplace transform of the process  $(V_t + \mu_t)$  coincides with the Laplace transform of  $(X_t)$  given by formula (21), and it follows that these two processes have the same law.

By a deeper result, due to Blackwell,  $(X_t - \mu_t)$  is almost surely a purely discontinuous process, which can increase only by jumps (this is discussed in Kingman 1993, chapter 8). For example, take a gamma process driven by the Lebesgue measure with constant scaling function: with probability one, on any bounded interval it consists of a series of infinitely many point masses, where the sum of the masses converges to a gamma distributed random variable.

Note also that expectation of functionals of the Lévy process can be computed by using the moment generating function given by (21).

Now, suppose that  $(T_i)$  is an exchangeable sequence of positive random variables, which are conditionally i.i.d. given a positive Lévy process  $(X_t)$ , with common cumulative hazard  $\Lambda_t := X_t$ .

It follows that, after a sample  $T_1, \dots, T_m$  is observed, the process  $(X_t)$  is still a Lévy process under the posterior distribution. This conjugacy property has been used since the early works in Bayesian nonparametrics by Ferguson (1974), Doksum (1974), and Ferguson & Phadia (1979), where the explicit expression of the posterior Lévy measure is derived, which holds also for right-censored survival data. For example, in the gamma process case with prior Lévy measure

$$\exp\{-z/\beta(y)\}dz \otimes \alpha(dy), \quad (25)$$

the posterior Lévy measure is

$$\exp\{-z/(\beta(y) + R(y))\}dz \otimes (\alpha + N)(dy), \quad (26)$$

where  $R(y)$  is the risk set at time  $y$  and the process  $N$  counts the number of observed (uncensored) points.

A disadvantage of this approach is that often it is an unrealistic simplification to assume independence of the increments of the cumulative hazard. This is certainly not appropriate if we assume an underlying piecewise continuous hazard rate curve, which will not match with a compensator increasing only by jumps.

A way out is to take a convolution of the positive Lévy process  $(X_t)$  with a kernel  $K(t, ds)$ , obtaining the cumulative hazard process

$$\Lambda(I, \omega) := \int_{-\infty}^{\infty} K(s, I)X(ds, \omega). \quad (27)$$

If the kernel is absolutely continuous, say  $K(t, ds) = k(t, s)ds$ , then the cumulative

hazard  $(\Lambda_t)$  admits an intensity process

$$\lambda(t, \omega) := \int_{-\infty}^{\infty} k(s, t) X(ds, \omega). \quad (28)$$

We assume that, for every  $t$ ,

$$\int_{-\infty}^t \int_0^{\infty} \min\{k(t, y)z, 1\} \frac{1}{z} L(dz, dy) < \infty \quad (29)$$

so that almost surely  $\lambda(t, \omega) < \infty$ .

It follows after observing a right-censored sample of exchangeable random variables with common random hazard rate  $(\lambda_t)$ , under the posterior distribution the process  $(X_t)$  is a mixture of Lévy processes.

Namely, let  $(\hat{T}_i, \zeta_i : i = 1, \dots, M)$  be the right censored sample, where  $\hat{T}_i$  is the last time an individual was seen and  $\zeta_i$  is 1 in case of observed failure and 0 in case of right censoring.

Conditionally on the data and the Lévy process  $(X_t)$ , introduce for each  $i$  such that  $\zeta_i = 1$  independent latent random variables  $S_i$ , with respective conditional distributions

$$S_i \simeq \frac{k(s, \hat{T}_i) X(ds)}{\int_{-\infty}^{\infty} k(s', \hat{T}_i) X(ds')}. \quad (30)$$

Then the likelihood for the Lévy process  $(X_t)$ , given the augmented sample including the latent variables  $\{S_i : \zeta_i = 1\}$ , is

$$\prod_{i: \zeta_i = 1} k(S_i, \hat{T}_i) dX(S_i) \exp \left\{ - \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} k(s, u) R(u) du \right) X(ds) \right\}, \quad (31)$$

where  $R(u)$  the size of the risk-set.

Since the likelihood for the increments of  $(X_t)$  over disjoint intervals factorizes into separate terms, it follows that, conditionally on the augmented sample,  $(X_t)$  is again a Lévy process. Thus, under the (non-augmented) posterior distribution,  $(X_t)$  is a mixture of Lévy processes with mixing parameters  $\{S_i : \zeta_i = 1\}$ . Although the mixing distribution turns out to be awkward, the posterior distribution can be sampled efficiently by a Markov chain method including the mixing parameters.

This Bayesian smoothing technique was introduced in the papers by Lo & Weng (1989), Ickstadt & Wolpert (1995), and it is used in our original paper [IV].

## 1.7. Personal views and perspectives.

I would like to conclude with a few final comments:

In our first example [I], to model the intensity of the observed point process we introduced a parameter marked point process on the same space. This generates a random partition of the sample space, where the marks define the levels of the intensity process, which is assumed to be piecewise constant over the partition. Since the number of points of the parameter process is unbounded, we may well say that we are working with a countable union of finite dimensional parameter spaces. Each realization of the parameter process contains only a finite number of points.

It is quite hard to extend this approach to the estimation of the intensity of a point process on a multidimensional space, just because it is very laborious to work with partitions in a higher dimensional space (on the plane this is done by Arjas & Heikkinen 1996).

Personally, I see more possibilities in working with Lévy process priors. This is a large family of processes, with a very simple structure, which can be defined on any measurable space. A sampling algorithm is given by Iekstadt & Wolpert (1995). The basic idea is to use the compound Poisson process structure of the Lévy process, and possibly to sample the points with bigger mass first. Alternative sampling techniques have been also proposed by Bondesson (1982), and Damien *et al.* (1995).

We have also seen how to introduce dependence and to smooth a Lévy parameter process by applying a convolution with a kernel. Given the data, the original parameter process becomes a mixture of Lévy processes. The posterior can be handled by sampling in turn a conditional Lévy process, given the mixing parameters, and the conditionally independent mixing parameters, given the parameter process.

It will be interesting to continue in this direction, and apply the Lévy process and mixture of Lévy processes framework to different structures, for example to Lexis' diagrams, and beyond that, to spaces of histories.

### 3. Summaries of original papers

- I. Arjas E & Gasbarra D (1994) Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statistica Sinica* 4: 505-524.

This is, up to my knowledge, one of the first Bayesian works where the Markov chain Monte Carlo method is applied to a model containing a random number of parameters. I have discussed these ideas in the previous introductory section: the observed marked point process contains a right censored sample of exchangeable lifetimes. We model the individual hazard rate as a piecewise constant process, which jumps at random times to random levels. The jump times form a Poisson process while the levels form a Markov chain. A Gibbs sampler algorithm is used to sample the parameter process from the posterior distribution, and compute the posterior predictive hazard and posterior predictive survival probability of a new individual. The prior assumption of increasing hazard is also considered.

- II. Arjas E & Gasbarra D (1996) Bayesian inference of survival probabilities, under stochastic ordering constraints. *JASA* 91: 1101-1109.

Two right censored survival data samples are observed. The assumption is that individuals are exchangeable within each group, and that the lifetimes in the first subpopulation are stochastically shorter than the lifetimes in the second population. We define piecewise constant priors for the two hazard processes. By using Markov chains we first construct a coupling of the two posterior distributions and then we impose the constraint. We also prove a technical lemma on the ergodicity of a Markovian coupling of two ergodic Markov chains.

- III. Arjas E & Gasbarra D (1997) On prequential model assessment in life history analysis. *Biometrika* 84: 505-522.

In this paper the question of model checking for point processes is posed. We follow the prequential (= predictive-sequential) paradigm of P. Dawid: given a filtration of available information, a model is a good probabilistic description of the data when, given the previous observations, it gives a good prediction for the incoming observations. We discuss a continuous time extension of these ideas for the Bayesian formulation, considering some simple models. We use the fact that if  $\tau_i$  are finite stopping times, such that the compensators are continuous, and  $P(\tau_i = \tau_j) = \delta_{ij}$ , then the compensators  $\Lambda_t^{\tau_i}$  of  $\tau_i$  taken at their final times time  $t = \tau_i$ , form a sequence of i.i.d. 1-exponential random variables under  $P$ . By computing and manipulating these r.v.'s we get both quantitative and graphical goodness-of-fit methods for the model  $P$ . The main problem is the computation of these stopped compensators. A Markov chain Monte Carlo algorithm is proposed.

- IV. Gasbarra D. & Karia S (1997) Analysis of competing risks by using Bayesian smoothing. Submitted for publication.

We consider right censored competing risks data, and we model separately the overall hazard process and cause-specific conditional probability by kernel smoothing a gamma process and a probability vector valued process piecewise constant on a random grid, respectively. The idea of kernel-smoothing a gamma process prior is borrowed from Lo & Weng (1989) and Ickstadt & Wolpert (1995). The posterior distribution is approximated by a data-augmented Gibbs sampler.

### 3. References

- Arjas E, Haara P & Norros I (1992) Filtering the histories of a partially observed marked point process. *Stoch. Proc. Appl.* 40: 225-250.
- Arjas E & Heikkinen J (1996) Nonparametric Bayesian estimation of a spatial Poisson intensity. To appear in *Scandinavian Journal of Statistics*.
- Bondesson L (1982) On simulation from infinitely divisible distributions. *Adv. Appl. Prob.* 14: 855-869.
- Brémaud P (1981) *Point Processes and Queues*. Springer Verlag, New York.
- Damien P, Laud PW & Smith AFM (1995) Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *J. Royal Statist. Soc. B* 57: 547-563.
- Dellacherie C & Meyer PA (1975) *Probabilités et Potentiel A*. Hermann, Paris.
- Doksum KA (1974) Tailfree and neutral random probabilities and their posterior distributions. *Ann. Statist.* 2: 183-201.
- Feller W (1971) *An Introduction to Probability Theory and its Applications*, vol.II. Wiley, New York.
- Ferguson T (1974) Prior distributions on the space of probability measures. *Ann. Statist.* 2: 615-629.
- Ferguson T & Phadia EJ (1979) Bayesian nonparametric estimation based on censored data. *Ann. Statist.* 7: 163-186.
- Florens JP, Mouchart M & Rolin JM (1990) *Elements of Bayesian Statistics*. Dekker, New York.
- Geyer CJ & Møller J (1994) Simulation and likelihood inference for spatial point processes. *Scand. J. Statist.* 21: 359-373.
- Gikhman II & Skorohod AV (1972) *The Theory of Stochastic Processes I*. Springer Verlag, New York.
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their

- applications. *Biometrika* 57: 97-109.
- Ickstadt K & Wolpert RL (1995) Gamma/Poisson random field models for spatial statistics. *Inst. Statistics & Decision Sci.*, Duke University, Report 95-43.
- Jacod J (1975) Multivariate point processes: predictable projections, Radon-Nikodym derivatives, representation of martingales. *Z. Wahrsch. Verw. Geb.* 31: 235-253.
- Jacod J & Shiryaev AN (1987) *Limit Theorems for Stochastic Processes*. Springer Verlag, Berlin.
- Kingman JFC (1993) *Poisson Processes*. Oxford University Press, Oxford.
- Lévy P (1957) *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, Paris.
- Lo AY & Weng CS (1989) On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* 41: 227-245.
- Meyn SP & Tweedie RL (1993) *Markov Chains and Stochastic Stability*. Springer Verlag, London.
- Norros I (1985) Systems weakened by failures. *Stoch. Proc. Appl.* 20: 181-196.
- Nummelin E (1984) *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, Cambridge.
- Pollard D (1984) *Convergence of Stochastic Processes*. Springer Verlag, New York.
- Skorohod AV (1991) *Random Processes with Independent Increments*. Kluwer, Dordrecht.
- Tierney L (1994) Markov chains for exploring posterior distributions. *Ann. Statist.* 22: 1701-1762.
- Tierney L (1995) A Note on Metropolis Hastings Kernels for General State Spaces. Preprint, University of Minnesota.
- Tierney L (1996) Introduction to general state space Markov chain theory. In: Gilks WR, Richardson S & Spiegelhalter DJ (eds) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London. p 59-74.

**Original papers**