# On Linear Intensity Models for Mixed Doubly Stochastic Poisson and Self-Exciting Point Processes

Yosihiko Ogata; Hirotugu Akaike

*Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 44, No. 1 (1982), 102-107.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281982%2944%3A1%3C102%3AOLIMFM%3E2.0.CO%3B2-K

*Journal of the Royal Statistical Society. Series B (Methodological)* is currently published by Royal Statistical Society.

*J. R. Statist. Soc.* B (1982),
**44,** *No.* 1, *pp.* 102–107

# On Linear Intensity Models for Mixed Doubly Stochastic Poisson and Self-exciting Point Processes

By YOSIHIKO OGATA and HIROTUGU AKAIKE

*The Institute of Statistical Mathematics, Tokyo, Japan*

[Received August 1980. Final revision May 1981]

SUMMARY

A flexible family of parametric models for intensity processes is introduced to represent a causal relationship between a point process and another stochastic process. Algorithms for the maximum likelihood computation and the procedure of model selection are discussed.

*Keywords*: PARTIAL LOG LIKELIHOOD; INTENSITY FUNCTION; RESPONSE FUNCTIONS; LAGUERRE TYPE POLYNOMIAL; RECURSIVE RELATIONS; AIC

## 1. INTRODUCTION

IN 1978 Vere-Jones (1978) investigated the causal relationship between the eruption of a volcano and deep earthquakes by applying a doubly stochastic Poisson process model. In the present paper we intend to extend the idea to the case with a more general input process and develop a procedure of modelling which will be useful even for the analysis of a causal relation between two series of data of different kind, such as the continuous record of a geophysical quantity and the record of earthquake occurrences. The importance of this type of modelling was recognized during our discussion of some microearthquake data with Professor Oike of the Disaster Prevention Research Institute of Kyoto University.

For the observation over $[0, T]$ of a point process $N_t$ the partial log likelihood in the sense of Cox (1975) is given by

$$L_T = \int_0^T \log \Lambda_\theta(t \mid H_t)\, dN_t - \int_0^T \Lambda_\theta(t \mid H_t)\, dt, \quad \theta \in \Theta, \tag{1.1}$$

where $\Lambda_\theta(t \mid H_t)$ denotes the intensity function parameterized by $\theta$, and the family of conditioning events $H_t$ consists of the past histories of $N_t$ itself and those of another observable process $X_t$. It is assumed that the observation $N_t$ is generated by a point process whose intensity is specified by $\Lambda_{\theta_0}(t \mid H_t)$ for some $\theta_0 \in \Theta$. For this case it can be shown, analogously to the proof of the asymptotic properties of the maximum likelihood estimate of a point process discussed in Ogata (1978), that the standard large sample theory holds under usual regularity conditions, such as the stationarity and ergodicity of the joint process $\{(X_t, N_t), t \geqslant 0\}$.

The practical applicability of the present model is heavily dependent on the availability of some proper parameterization of the intensity function $\Lambda_\theta(t \mid H_t)$. In the present paper we propose a system of parametric families of $\Lambda_\theta(t \mid H_t)$ which allows efficient calculation of the (partial) likelihoods. The model selection is then realized by the minimum AIC procedure.

## 2. LINEAR INTENSITY MODELS

Throughout the present paper we consider the point process such that the relations

$$P\{N_{t+\Delta t} - N_t = 1 \mid H_t\} = \Lambda(t \mid H_t)\, \Delta t + o(\Delta t)$$

and

$$P\{N_{t+\Delta t} - N_t \geqslant 2 \mid H_t\} = o(\Delta t) \tag{2.1}$$

 0035–9246/82/44102$2.00

hold for small $\Delta t$. This implies that

$$E\{N_{t+\Delta t}-N_t \mid H_t\} = \Lambda(t \mid H_t)\,\Delta t + o(\Delta t). \tag{2.2}$$

Since $H_t$ consists of the past histories of the processes $X_s$ and $N_s$ before time $t$, we are interested in the linear model defined by

$$\Lambda(t \mid H_t) = \mu + \int_0^t g(t-s)\,dN_s + \int_0^t h(t-s)\,dX_s. \tag{2.3}$$

This type of model was first discussed by Hawkes (1971) for the case where $\{X_t\}$ is a point process. In our present model the process $\{X_t\}$ may either be a point process or a cumulative process

$$X_t = \int_0^t x(s)\,ds \tag{2.4}$$

of some stochastic process $\{x(t)\}$.

When $h(t)\equiv 0$ holds, this means that there is no causal relation between the input $\{X_t\}$ and the output $\{N_t\}$. Also $g(t)\equiv 0$ means that the output process is a doubly stochastic Poisson, while $g(t)\equiv h(t)\equiv 0$ means that the output process is a homogeneous Poisson process of rate $\mu$.

## 3. Parameterization and the Likelihood Computation

For the parameterization of $g(t)$ we propose the use of the Laguerre type polynomial

$$g(t) = \sum_{k=0}^{K} a_k t^k e^{-ct}. \tag{3.1}$$

We also adopt a similar parameterization for the response function $h(t)$ given by

$$h(t) = \sum_{k=0}^{L} b_k t^k e^{-ct}, \tag{3.2}$$

where the exponential coefficient $c$ is assumed to be equal to that of $g(t)$. The assumption of the same $c$ in both $g(t)$ and $h(t)$ is for the sake of convenience of the likelihood computation.

Given the occurrence times of two types of events $\{t_i, \tau_m;\ i = 1,\ldots, I,\ m = 1,\ldots, M\}$ during the time interval $[0, T]$, we fit the model (2.3) to the data, regarding the series $\{t_i\}$ as the output and $\{\tau_i\}$ as the input. The partial log likelihood (1.1) of the model thus defined can easily be obtained in terms of

$$\int_0^T \log \Lambda_\theta(t \mid H_t)\,dN_t = \sum_{i=1}^{I} \log \Lambda_\theta(t_i \mid H_{t_i})$$
$$= \sum_{i=1}^{I} \log\left\{\mu + \sum_{k=0}^{K} a_k P_k(i) + \sum_{k=0}^{L} b_k Q_k(i)\right\},$$

where

$$P_k(i) = \sum_{t_j<t_i} (t_i-t_j)^k \exp\{-c(t_i-t_j)\}$$
$$Q_k(i) = \sum_{\tau_m<t_i} (t_i-\tau_m)^k \exp\{-c(t_i-\tau_m)\}, \tag{3.3}$$

and

$$\int_0^T \Lambda_\theta(t \mid H_t)\,dt = \mu T + \int_0^T dt \int_0^t g(t-s)\,dN_s + \int_0^T dt \int_0^t h(t-s)\,dX_s$$
$$= \mu T + \sum_{k=0}^{K} a_k \sum_{i=1}^{I} R_k(T-t_i) + \sum_{k=0}^{L} b_k \sum_{m=0}^{M} R_k(T-\tau_m),$$

where

$$R_k(t) = \int_0^t t^k e^{-ct} dt. \tag{3.4}$$

The following recursive relations are useful for the computation of the likelihood:

$$P_k(i+1) = (t_{i+1}-t_i)^k \exp\{-c(t_{i+1}-t_i)\} + \sum_{j=0}^{k} \binom{k}{j} (t_{i+1}-t_i)^{k-j} \exp\{-c(t_{i+1}-t_i)\} P_j(i),$$

$$Q_k(i+1) = D_k(t_i, t_{i+1}) + \sum_{j=0}^{k} \binom{k}{j} (t_{i+1}-t_i)^{k-j} \exp\{-c(t_{i+1}-t_i)\} Q_j(i)$$

and

$$R_{k+1}(t) = \{(k+1) R_k(t) - t^{k+1} e^{-ct}\}/c, \tag{3.5}$$

where

$$D_k(t_i, t_{i+1}) = \sum_{t_i \leqslant \tau_m < t_{i+1}} (t_{i+1}-\tau_m)^k \exp\{-c(t_{i+1}-\tau_m)\}.$$

By definition $t_0 = 0$, we get $P_0(0) = Q_0(0) = 0$ and $R_0(t) = (1-e^{-ct})/c$.

If the input is defined by a cumulative process $dX_t = x(t)\,dt$, $0 \leqslant t \leqslant T$, we approximate it by

$$dX_t^* = (T/M) \sum_{m=1}^{M} x(t)\,\delta(t-\sigma_m)\,dt,$$

where $M$ is a properly chosen large integer at least of the order of the sample size $I$, $\sigma_m = (m/M)T-(1/2M)T$, $m = 1, 2, ..., M$, and $\delta(t)$ denotes Dirac's delta function. With this approximation, we get an approximate partial log likelihood

$$L_T^*(\theta) = \sum_{i=1}^{I} \log\{\mu + \sum_{k=0}^{K} a_k P_k(i) + \sum_{k=0}^{L} b_k U_k(i)\} - \mu T$$

$$- \sum_{k=0}^{K} a_k \sum_{i=1}^{I} R_k(T-t_i) - \sum_{k=0}^{L} b_k \sum_{m=0}^{M} x(\sigma_m) R_k(T-\sigma_m),$$

where $P_k(i)$ and $R_k(t)$ are the same as in the previous case and $U_k(i)$ is given recursively by

$$U_k(i+1) = F_k(t_i, t_{i+1}) + \sum_{j=0}^{k} \binom{k}{j} U_j(i) (t_{i+1}-t_i)^{k-j} \exp\{-c(t_{i+1}-t_i)\}$$

and

$$F_k(t_i, t_{i+1}) = \sum_{t_i \leqslant \sigma_m < t_{i+1}} x(\sigma_m) (t_{i+1}-\sigma_m)^k \exp\{-c(t_{i+1}-\sigma_m)\}.$$

The gradient of each partial log likelihood function can easily be obtained by differentiating the above functions. Once the gradient is obtained the maximization of the likelihood function is performed by using a standard non-linear optimization technique developed by Fletcher and Powell (1963).

## 4. MODEL SELECTION

We can estimate the coefficients $a_k (k = 1, ..., K)$, $b_k (k = 1, ..., L)$ and $c$ of the polynomials (3.1) and (3.2) by the maximum likelihood method, provided that the orders $K$ and $L$ are given. By the minimum AIC procedure (Akaike, 1977), we select $(K, L)$ which minimizes

$$\text{AIC}(K, L) \equiv (-2) \max(\text{log likelihood}) + 2(\text{number of parameters})$$

$$= (-2) \max L_T(\mu, c, a_0, ..., a_k, b_0, ..., b_L) + 2(K+L+4). \tag{4.1}$$

The AIC is an estimate of the expected negentropy which is a natural measure of discrimination between the true and estimated probability law of the data. The use of the minimum AIC procedure is justified under the assumption of the standard large sample theory of the maximum likelihood estimate which is briefly touched in Introduction.

It is certainly possible to get the maximum likelihood estimates of all the parameters. However this is too much time-consuming when the sample size is large or the orders $K$ and $L$ are high. This is due to the fact that the functions $P_k(i)$, $Q_k(i)$, $U_k(i)$ and $R_k(t)$ contain the exponential parameter $c$ which changes in every step of the non-linear optimization. Also very frequently there are more than one maxima of the likelihood function.

However, once the parameter $c$ is fixed, the log likelihood function has at most one maximum. This is seen from the fact that the Hessian is everywhere non-positive definite with respect to all the other parameters, as the intensity function $\Lambda_\theta(t)$ is linearly parameterized (Ogata, 1978, p. 255). Thus for the computational convenience, it is quite advisable to repeat the maximum likelihood computation for a finite number of coarsely distributed fixed values of $c$. Accordingly we adopt the following algorithms.

For a fixed exponential coefficient $c$, compute the following statistics

$$P(k,i) \equiv P_k(i), \quad Q(k,i) \equiv Q_k(i),$$

$$U(k,i) \equiv U_k(i), \quad V(k) \equiv \sum_{i=1}^{I} R_k(T-t_i),$$

$$W(k) \equiv \sum_{m=0}^{M} R_k(T-\tau_m) \tag{4.2}$$

and

$$S(k) \equiv \sum_{m=0}^{M} x(\sigma_m)\, R_k(T-\sigma_m)$$

for $i = 0,1,\ldots,I$ and $k = 1,2,\ldots,\bar{K}$ where $\bar{K}$ is the upper bound of the orders. Compute the partial log likelihood for the point process input by

$$L_T(\theta) = \sum_{i=1}^{I} \log\left\{\mu + \sum_{k=0}^{K} a_k\, P(k,i) + \sum_{k=0}^{L} b_k\, Q(k,i)\right\} - \mu T - \sum_{k=0}^{K} a_k\, V(k) - \sum_{k=0}^{L} b_k\, W(k), \tag{4.3}$$

or that for the accumulated process input by

$$L_T^*(\theta) = \sum_{i=1}^{I} \log\left\{\mu + \sum_{k=0}^{K} a_k\, P(k,i) + \sum_{k=0}^{L} b_k\, Q(k,i)\right\} - \mu T - \sum_{k=0}^{K} a_k\, V(k) - \sum_{k=0}^{L} b_k\, S(k), \tag{4.4}$$

where $\theta$ stands for $(\mu, \{a_k\}, \{b_k\})$. Keep the statistics (4.2) and maximize the function (4.3) or (4.4) sequentially for each pair $(K, L)$ such that $K \leqslant \bar{K}$ and $L \leqslant \bar{K}$. Repeat the whole process with several other possible choices of $c$ and find the maximum of the likelihood with respect to these values of $c$.

Suppose we have some candidate values of $c$ (say $c_j$, $j = 1,2,\ldots,J$). Then we can compare

$$\mathrm{AIC}\,(c_j, K, L) = (-2)\max_{\theta} L_T(c_j, \theta) + 2(K+L+3),$$

where $\theta = (\mu, a_0, \ldots, a_K, b_0, \ldots, b_L)$, $K, L = 0,1,2,\ldots,\bar{K}$ and $j = 1,2,\ldots,J$. Thus we can perform the minimum AIC procedure quite systematically up to some considerably high degree models.

The selection of the values $c_j$ can be realized in the following manner. Firstly we assess the range of $t$ where the response functions $g(t)$ and $h(t)$ are significantly different from zero. This may be obtained by some prior information about the data or by some preliminary analysis of the data such as the auto and cross-correlogram analysis. Let a rough estimate of the range be $R$, and assume the response function of the form $ax^k e^{-ct}$ which has its peak at $x = k/c$. Then, we assume that the peak is attained in the middle of the range, i.e. we assume the relation $k/c = R/2$. If we assume $k = 2$, for example, we get $c_0 = 4/R$ as an initial guess of the exponential coefficient. Another practical way of getting an initial estimate of $c$ is to apply the direct maximum likelihood estimation procedure to the second-order polynomial model using a short subset of the data. We successively try $c_j = 2^j c_0$ or $2^{-j} c_0, j = 1, 2, \ldots, J$ and compare the AIC values sequentially until a minimum is obtained. Further search for the optimum value of $c$ can be continued within the interval defined by the two $c_j$'s, adjacent to the one which has given the minimum of AIC. By our experience the estimated response functions with different $c_j$'s but with similar AIC values are quite similar to each other in spite of their possible differences of the orders.

## 5. Discussion and Remarks

We checked the feasibility of the procedure described in Sections 3 and 4 by some artificially generated data. Also the model was applied to a pair of earthquake series which occurred in certain different seismic regions in Japan over the period 1924–74. The results of our analysis show that the earthquakes in one region are not only self-exciting but also significantly receive one-way stimulation from earthquake occurrences in the other region. The details of this earthquake example can be seen in a companion paper (Ogata, Akaike and Katsura, 1981) to which the interested reader is referred. Here we would like only to give some feeling of the estimation procedure of the response functions by Fig. 1 obtained by applying the

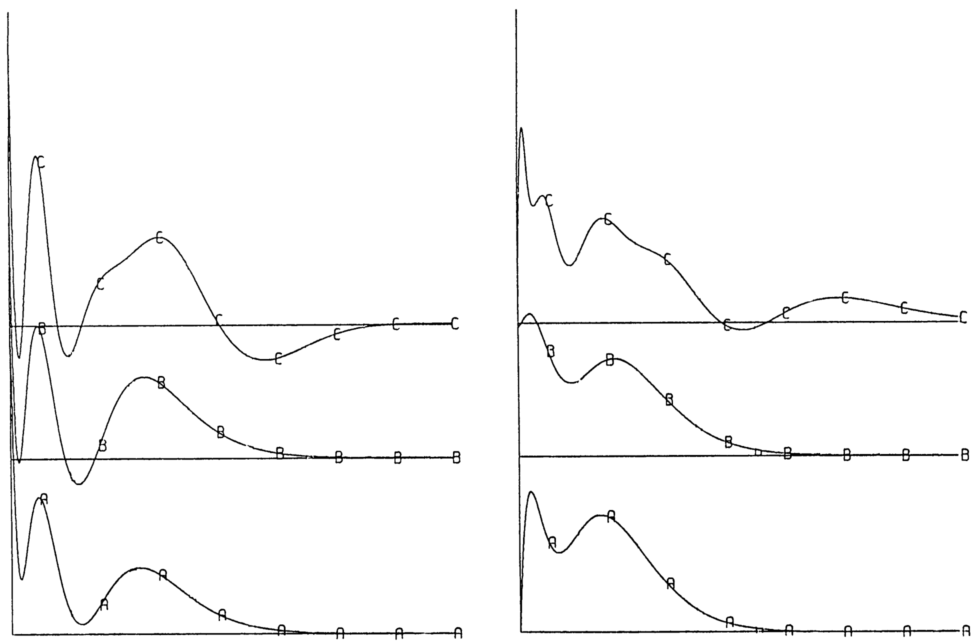

FIG. 1.

–A–A–A–: The true response functions; $C = 1{\cdot}1, K = L = 5$.

–B–B–B–: The estimated response functions with the minimum $\text{AIC}(2^{1/4}, 5, 5) = 3281{\cdot}1$.

–C–C–C–: The estimated response functions with $\text{AIC}(2^{1/4}, 10, 10) = 3289{\cdot}9$.

procedure of Section 4 to artificially generated bivariate series of events. The numbers of the events in the input and output series were 992 and 769 over the time interval [0, 2000], respectively. The result clearly shows the practical utility of the procedure.

Hawkes (1971) assumed the response functions $g(t)$ and $h(t)$ in (2.3) to be non-negative to ensure the non-negativity of the intensity process with probability one. The point process of this type has a certain kind of clustering property (see Hawkes and Oakes, 1974). However, there are examples of series of events with some inhibitory property like spike trains of nerve-system data. If we fit the parametric model of Section 3 to such data, we often get response functions having negative values in some parts of $t$ (Nakamura, Ogata and Oomura, 1980). If the negativity of the response functions does not affect the non-negativity of the estimated intensity process throughout the observed time interval, the negativity could be understood as an indication of the existence of an inhibitory range. To handle such possibility properly we will have to replace the parametric intensity function in (1.3) by $\Lambda_\theta^+(t \mid H_t) = \max\{\Lambda_\theta(t \mid H_t), 0\}$ where $\Lambda_\theta(t \mid H_t)$ is given in (2.3).

If $\mu > 0$ holds in (2.3), it is easily seen that the existence of the negative valued parts in the estimated response functions does not always imply the existence of negative parts of the intensity process. Given a series of events $\{t_i, \tau_m, i = 1, ..., I, m = 1, ..., M\}$ over the time interval $[0, T]$, if the intensity $\Lambda_\theta(t \mid H_t)$ with $\theta$ obtained by the method of maximum likelihood is non-negative in $[0, T]$, then that estimate also maximizes the log likelihood function

$$L_T(\theta) = \int_0^T \log \Lambda_\theta^+(t \mid H_t)\, dN_t - \int_0^T \Lambda_\theta^+(t \mid H_t)\, dt.$$

## Acknowledgements

The authors would like to thank Professor K. Oike of Kyoto University for valuable suggestions about earthquakes data. The authors are also grateful to Professor D. Vere-Jones and Dr T. Ozaki for very helpful suggestions during the preparation of this paper. The idea of using Laguerre type polynomials for modelling the response functions was obtained during the discussions with Ozaki. K. Katsura and G. Kitagawa generously helped us with the programming of the necessary computations. The authors are particularly grateful to the referees whose comments led to a significant revision of the paper.

## References

AKAIKE, H. (1977). On Entropy Maximization principle. In *Application of Statistics* (P. R. Krishnaiah, ed.), Amsterdam: North-Holland. 27–41.

COX, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

FLETCHER, R. and POWELL, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer J.*, **6**, 163–168.

HAWKES, A. G. (1971). Spectra of some mutually exciting point processes. *Biometrika*, **58**, 83–90.

HAWKES, A. G. and OAKES, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Prob.*, **11**, 493–503.

NAKAMURA, M., OGATA, Y. and OOMURA, Y. (1980). Evaluation of neuronal spike train data by new methodologies of wave form discrimination and point process analysis. In *Proceedings of the 7th International CODATA Conference* (Y. Mashiko, ed.). Oxford: Pergamon.

OGATA, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.*, **30**, A, 243–261.

OGATA, Y., AKAIKE, H. and KATSURA, K. (1981). The application of linear intensity models to the investigation of causal relations between a point process and another stochastic process, *Research Memo.*, No. 203, The Inst. Statist. Math., Tokyo (Submitted to *Ann. Inst. Statist. Math. B.*)

VERE-JONES, D. (1978). Personal communication.