Appl. Statist. (1992)
41, *No.* 1, *pp.* 31–38

Approximating Point Process Likelihoods with GLIM

By MARK BERMAN† and T. ROLF TURNER

CSIRO, Sydney, Australia

[Received August 1987. Final revision July 1990]

SUMMARY

This paper shows how approximate maximum likelihood estimation for fairly general point processes on the line can be performed with GLIM. The approximation is based on a weighted sum approximation to an integral in the likelihood. Various weighting schemes are briefly examined. The methodology is illustrated with an example, and its extension to Poisson processes in higher dimensions is briefly described.

Keywords: GLIM; Likelihood; Point process; Poisson process; Weights

1. Introduction

Although models for point processes on the line and their statistical analysis abound (Cox and Lewis, 1966; Cox and Isham, 1980), there has been, until recently, relatively little modelling of the dependence of point processes on covariates and their associated statistical analysis. Probably the earliest major models of this type were Cox's (1972) modulated Poisson processes and their generalization, and Berman's (1981) modulated gamma processes. Both papers only seriously addressed questions of hypothesis testing (via the likelihood ratio test), but did not satisfactorily solve the problem of parameter estimation. More recently, there have been several papers devoted to seismological applications, e.g. Ogata and Akaike (1982), Ogata (1983, 1988), Ogata and Katsura (1986) and Vere-Jones and Ozaki (1982). In these papers, the relationship between the point process and the covariates, while being numerically convenient, can give rise to likelihoods which can be negative for certain parameter values. To overcome this problem, most of the above-mentioned researchers use a penalized likelihood approach to estimate the parameters. As a result, the statistical properties of these estimators are not clear.

Another recent development in this field is the use of generalized linear models (GLMs). Becker (1983, 1986) and Brillinger and Preisler (1986) use the GLM framework for specific inhomogeneous Poisson processes. However, they do not exploit the full potential of this framework. In this paper, we shall show how a fairly large class of point processes on the line, which incorporate dependence on covariates, can be modelled as GLMs; this class includes models (or slight variants of them) used in all

†*Address for correspondence*: CSIRO Divison of Mathematics and Statistics, PO Box 218, Lindfield, New South Wales 2070, Australia.

0035–9254/92/41031 $2.00

the above-mentioned papers, except for the modulated gamma process. The practical advantage of using a properly defined GLM formulation is that it enables the maximization of an approximation to the likelihood by means of the well-known computer package GLIM (Payne, 1985), rather than for instance maximizing a penalized likelihood as is done in the above-mentioned seismological papers. This, in turn, enables the use of the asymptotically optimal statistical properties of maximum likelihood (ML) estimation for large classes of point processes (Ogata, 1978). Our hope is that, not only will point process specialists see how certain technical problems can be overcome with the aid of GLIM, but also that other statisticians, already familiar with GLIM, will feel more comfortable when presented with point process problems.

At this point, we must also mention the work of Hurley (1990) which has developed concurrently with ours, has strong connections with the present paper and has come to our attention during the refereeing process.

The present paper is organized as follows. Section 2 contains some basic point process theory, and in particular a general formula for the likelihood of a point process on the line. This is illustrated with an example taken from Ogata (1983). In Section 3, we show how GLIM can be used to approximate the likelihood, and we use this approximation to analyse a modified version of Ogata's model. The theory of Section 3 is extended to spatial Poisson processes in Section 4, while Section 5 contains a summary and conclusions.

2. Basic Theory and an Example

Suppose that we observe an orderly point process (Cox and Isham, 1980) on $(0, T]$ with conditional intensity function (CIF)

$$\lambda_\theta(x|F_x) = \lim_{\Delta\to 0}(\Pr\{\geqslant 1 \text{ event in } [x, x+\Delta)|F_x\}/\Delta), \tag{2.1}$$

where F_x represents the history of the process (and possibly other explanatory stochastic processes) prior to x, and $\theta = (\theta_1, \ldots, \theta_p)$ is a vector of unknown parameters requiring estimation. Assume that the CIF is of the form

$$\lambda_\theta(x|F_x) = g\left\{\sum_{k=1}^{p} \theta_k\, Q_k(x|F_x)\right\}, \tag{2.2}$$

where $g(\ )$ is a monotonic continuous function and $Q_k(x|F_x)$, $k=1, \ldots, p$, is a sequence of covariates whose values are known for all $x \in [0, T]$. Each Q_k may be purely a function of x (e.g. x^k) or an observed stochastic process. When $g(y) = \exp y$ and all the Q_k are independent of F_x, equation (2.2) represents the intensity function of Cox's (1972) modulated Poisson process.

Let the event times in $(0, T]$ of the observed point process be $x_1 < x_2 < \ldots < x_n$. Suppose that we wish to compute the ML estimators of the θ_j. The log-likelihood for any orderly point process can be written

$$L(\theta) \equiv L(\theta; x_1, \ldots, x_n; T) = \sum_{i=1}^{n} \log \lambda_\theta(x_i|F_{x_i}) - \int_0^T \lambda_\theta(x|F_x)\, \mathrm{d}x. \tag{2.3}$$

To make this theory more concrete, we introduce an example studied by Ogata

(1983), section 4.2. He examined the relationship between 'shallow' and 'deep' earthquakes in a region covering the North Island of New Zealand from 1946 to 1980. There were 58 shallow earthquakes (at times $x_1, \ldots, x_{58}$) and 84 deep earthquakes (at times $u_1, \ldots, u_{84}$). Ogata modelled the dependence of shallow earthquakes on deep earthquakes using the CIF

$$\lambda_\theta(x|F_x) = \mu + \sum_{j=1}^{J} \alpha_j \phi_j(2x/T-1) + \sum_{k=1}^{K} \beta_k \sum_{u_l<x} (x-u_l)^{k-1} \exp\{-d(x-u_l)\}, \qquad (2.4)$$

where J, K and $\theta = \{d, \mu, \alpha_j, j=1, \ldots, J, \beta_k, k=1, \ldots, K\}$ are unknown and $\phi_j(\ )$ is the jth Legendre polynomial. The first summation in equation (2.4) represents orthogonal time trend components, while the second represents directly the dependence of shallow earthquakes on deep earthquakes.

For fixed J, K and d, equation (2.4) is of the form (2.2) with $g(y)=y$. If we wish to compute the ML estimators for a model of the form (2.2), it is at first glance *numerically* convenient to use $g(y)=y$, for then the unknown parameters can be taken outside the integral sign in equation (2.3) and only the integrals $\int_0^T Q_k(x|F_x)\,dx$, $k=1, \ldots, p$, need be computed. Unfortunately, the use of $g(y)=y$ does not guarantee the non-negativity of the CIF and hence the non-negativity of its likelihood (see equations (2.1)–(2.3)). Therefore, for each (J, K, d) combination examined, Ogata maximized

$$L_p(\theta, J, K) = L(\theta, J, K) - P(\theta, J, K), \qquad (2.5)$$

where $P(\theta, J, K)$ is a non-negative *penalty function* designed to keep the CIF non-negative for $x \in [0, T]$. The non-linear parameter d was varied over a grid of candidate values, and equation (2.5) was maximized over the linear parameters for each d. The orders J and K of the model were chosen to minimize a modified Akaike information criterion (AIC), namely

$$\mathrm{AIC}_p = -2\, L_p(\hat{\theta}_{J,K}, J, K) + 2(J+K+1), \qquad (2.6)$$

where $\hat{\theta}_{J,K}$ is the value of θ maximizing equation (2.5) given J and K; the usual AIC has L instead of L_p in equation (2.6). The model chosen using AIC_p had $J=7$, $K=0$. Thus the model chosen contained only trend components; this suggests that shallow earthquakes are independent of deep earthquakes occurring beforehand.

Most of the seismologically oriented papers mentioned earlier considered models with a CIF of the form (2.2) and with $g(y)=y$. Because of the consequent possibility of the CIF becoming negative, they also maximized a penalized log-likelihood. Unfortunately, this obscures both the numerical and the statistical properties of the estimators. In the next section, we show how GLIM can be used in conjunction with a non-negative $g(\ )$ (e.g. $g(y)=\exp y$) to maximize directly an approximation to the log-likelihood. The method will be applied to Ogata's data and compared with his original analysis.

3. Approximating Likelihoods with GLIM

There are two closely related ways of using GLIM to maximize equation (2.3) approximately. Both involve a discretization of the time axis. One is a *probabilistic* approximation, while the other is *numerical*. In the probabilistic approximation, the time axis is divided into intervals of *equal* width, the width being chosen sufficiently small to ensure that none of the intervals contains more than one event. The point process is then approximated by a 0–1 time series. GLIM can then be applied using binomial ERRORS with sample size 1. Sometimes these are replaced by Poisson ERRORS. Details will not be given here. Recent examples of the use of such probabilistic approximations can be found in Becker (1983, 1986) and Brillinger and Preisler (1986).

The numerical approximation, which we introduce here, is more flexible than the probabilistic approximation. In particular, it allows for the use of intervals of *unequal* widths. The generalization of this idea is particularly important when considering *spatial* Poisson processes, which will be discussed in Section 4. The idea is similar to, but goes further than, an approximation used by Clayton (1983) for survival data. The first step is to approximate the integral in equation (2.3) by a weighted sum:

$$\int_0^T \lambda_\theta(x|F_x)\,\mathrm{d}x \doteq \sum_{k=1}^{M} w_j \lambda_\theta(s_j|F_{s_j}), \tag{3.1}$$

where the w_j are weights determined by some quadrature rule and $0 \equiv s_1 < s_2 < \ldots < s_M \equiv T$. Choose the s_j so that $X \equiv \{x_1, \ldots, x_n\} \subset \{s_1, s_2, \ldots, s_M\} \equiv S$. We shall call points in S, X and $S-X$ the *design* points, the *data* points and the *dummy* points respectively. Let $N_j = 1$ if $s_j \in X$ and $N_j = 0$ otherwise. Then substitution of equation (3.1) into equation (2.3) gives

$$L(\theta) \doteq \sum_{j=1}^{M} w_j\{(N_j/w_j)\log \lambda_\theta(s_j|F_{s_j}) - \lambda_\theta(s_j|F_{s_j})\}. \tag{3.2}$$

The right-hand side of equation (3.2) is (modulo a constant not depending on θ) the weighted log-likelihood of a family of independent Poisson variates with parameters $\lambda_\theta(s_j|F_{s_j})$. The maximization of equation (3.2) can thus be accomplished in GLIM by declaring $\{N_j/w_j\}$ as the YVARIATE vector with Poisson ERRORS, WEIGHT given by $\{w_j\}$, and the LINK function to be g^{-1} (LOG is the default for Poisson errors).

Note the following points:

(a) GLIM does not object to the fact that the 'Poisson' variables N_j/w_j and the weights w_j are non-integer;
(b) GLIM does object to the w_j being *negative*;
(c) the parameter estimates, likelihood values and covariance matrices produced by GLIM when maximizing the approximate log-likelihood (3.2) are valid approximations to analogous functions of the true log-likelihood (2.3) in the sense that, if $M \to \infty$ in such a way that approximation (3.1) becomes exact, the approximating functions of equation (3.2) will converge to the analogous functions of equation (2.3).

Often, any dummy points will be chosen to be equispaced, the number being chosen according to the (numerical) accuracy of estimation required. The necessity of using the data points will impose a non-uniform spacing of the design points, however.

Simple schemes for weights are obtained either by

(a) a *linear* approximation to $\lambda_\theta(x)$ on $[s_j, s_{j+1}]$, $j = 1, \ldots, M-1$ or
(b) a *quadratic* approximation to $\lambda_\theta(x)$ on $[s_{2j-1}, s_{2j+1}]$, $j = 1, \ldots, \frac{1}{2}(M-1)$

(assuming that M is odd).

The linear approximation will always produce positive weights. However, if $s_{2j} - s_{2j-1}$ and $s_{2j+1} - s_{2j}$ are very different, the quadratic approximation leads to negative weights to which GLIM objects. This is easily overcome by adding an odd number of *equispaced* dummy points in each interval (x_i, x_{i+1}), $i=0, 1, \ldots, n$ $(x_0 \equiv 0, x_{n+1} \equiv T)$. Such a scheme ensures that $s_{2j+1} - s_{2j} = s_{2j} - s_{2j-1}$ for all j.

3.1. *Return to Example*

We generalized Ogata's CIF (2.4) to

$$\lambda_\theta(x|F_x) = g\left[\mu + \sum_{j=1}^{J} \alpha_j \phi_j(2x|T-1) + \sum_{k=1}^{K} \sum_{u_l<x} (x-u_l)^{k-1} \exp\{-d(x-u_l)\}\right]. \quad (3.3)$$

We tried both the IDENTITY link $g(y)=y$ (i.e. Ogata's model but without any penalty function) and the LOG link $g(y)=\exp y$. Because of its complexity, we have only considered versions of equation (3.3) with the value of d (1.07×10^{-4}) chosen by Ogata. For a range of (J, K) values, we used GLIM to maximize log-likelihood (3.2) with $\lambda_\theta(x|F_x)$ given by equation (3.3). We used both linear and quadratic weights. The total number M of design points used varied between the minimum possible, 60, and 1093. With the IDENTITY link, we found that GLIM stopped for some (J, K, M) combinations because $\lambda_{\hat\theta}(s_j|F_{s_j})$ was negative for some j. However, no problems were encountered using the LOG link. With this link function, we calculated the AIC for each (J, K) combination (i.e. equation (2.6) with L_p replaced by L). Often there are several models with AIC values very close to the minimum. It is important in such cases to estimate the log-likelihood accurately; however, if our model is fixed and we are only interested in parameter estimates and their standard errors, accurate estimation of the log-likelihood may be less important. A CIF such as equation (3.3) will tend to oscillate more as J or K increases, by analogy with ordinary polynomials. Hence, as a general rule, the larger J or K is, the more design points will be needed to estimate $L(\theta)$ accurately. This is borne out by the example considered here. In Table 1, we give the AIC values for various (J, K) combinations using the LOG link and linear weights with $M=60$ and $M=1061$.

Table 1 shows clearly how the two approximations to the AIC for each (J, K) combination become more discrepant broadly speaking as J or K increases. For all

TABLE 1
AIC values for a model with CIF given by equation (3.3) using $g(x)=\exp x$, $d=1.07\times10^{-4}$ and linear weights with $M=60$ and $M=1061$

K	*AIC value for the following values of J and M:*									
	J									
	0		*2*		*4*		*8*		*12*	
	M									
	60	*1061*	*60*	*1061*	*60*	*1061*	*60*	*1061*	*60*	*1061*
0	743.9	743.9	727.6	727.4	730.4	729.5	730.5	723.6	735.5	727.9
1	725.7	725.3	729.5	729.1	732.1	731.2	732.0	724.5	737.4	729.6
2	727.6	727.1	731.5	731.1	730.6	729.3	733.9	725.8	738.7	731.6

(J, K) combinations examined, the approximations based on both linear and quadratic weights appeared to stabilize as M approached 1000.

For the model with a LOG link, it was found that the minimum AIC value achieved was 723.6, and that this was achieved when $J=8$, $K=0$. Thus, as in Ogata's analysis, the model chosen contains only time trend components. A graph of the resulting estimated CIF, together with Ogata's estimate and a yearly histogram of the data, is shown in Fig. 1. The two estimated CIFs are mostly similar in general appearance. Except for the spurious vertical asymptotes at the end points, we feel that our estimate reflects the behaviour of the data—as encapsulated by the histogram—somewhat better than does Ogata's estimate. This can be seen for instance in the levelling-off around 1000 days and the peaks near 6000 and 11000 days. The least appealing aspect of our fit is the presence of the spurious vertical asymptotes. This is largely due to the fact that $\hat{\alpha}_8$ ($=1.821$) is positive and so is ϕ_8 in both tails. Hence, $\hat{\alpha}_8\phi_8$ dominates in the tails, and exponentiating magnifies this dominance. (We confirmed this by plotting the fitted curve without the term $\hat{\alpha}_8\phi_8$. The fit was much more like Ogata's in the tails and the asymptotes disappeared. However, the minimum AIC value when $J=7$, $K=0$ was 731.7, far larger than that for $J=8$, $K=0$.) This is the unfortunate price we appear to have to pay if likelihood methods are to be used. We believe that it is worth paying because these are only edge effects (effects which also arise in many other branches of statistical modelling) and because the inferences to be drawn from the data analysis rest more securely on the asymptotic statistical properties of ML estimation than do those based on penalized likelihood methods.

4. Spatial Poisson Processes

The ideas in Section 3 can also be applied to *spatial Poisson* processes, but not, in

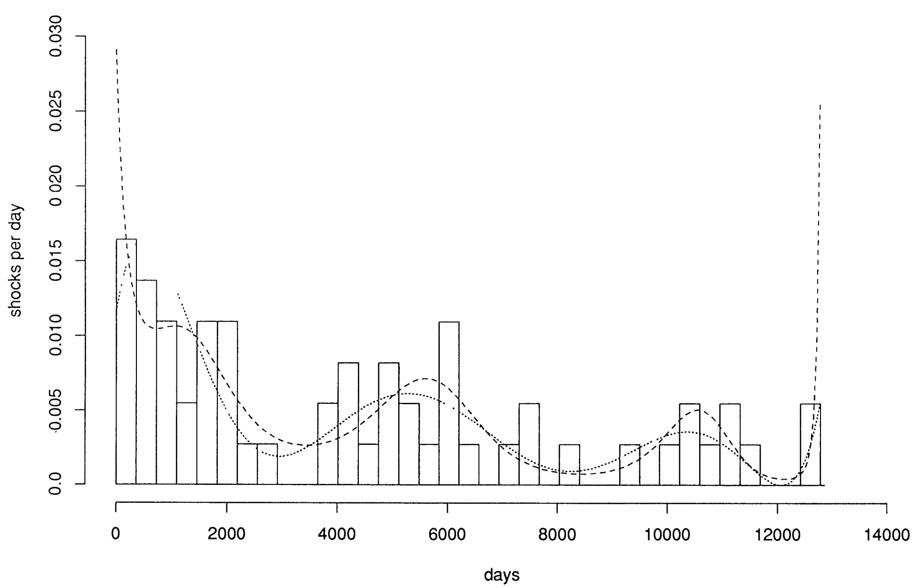

Fig. 1. Ogata (········) and Berman–Turner (--------) estimates of the CIF for Ogata's (1983) earthquake data, together with a yearly histogram (———) of the data

general, to other spatial point processes. The reason for this is that the log-likelihood for most non-Poisson spatial point processes cannot be written in a relatively simple form such as equation (2.3). In this section, we shall briefly outline the methodology for the Poisson case. For simplicity, we shall only consider the two-dimensional case; the methodology extends, in principle, to higher dimensions in an obvious way. Let A denote the region on which the Poisson process is observed and let $x_1, \ldots, x_n$ denote the event locations in A.

By analogy with equation (2.2), assume that the intensity function of the process, $\lambda_\theta(x)$, is of the form

$$\lambda_\theta(x) = g\left\{ \sum_{k=1}^{p} \theta_k Q_k(x) \right\}, \tag{4.1}$$

where $g(\ )$ is again a monotonic continuous function and $Q_k(x)$, $k=1, \ldots, p$, is a sequence of covariates whose values are known for all $x \in A$. Unlike in equation (2.2), $\lambda_\theta(x)$ and $Q_k(x)$ are independent of the point process everywhere else, and so the point process is Poisson. The log-likelihood for the Poisson process is given by

$$L(\theta; x_1, \ldots, x_n; A) = \sum_{i=1}^{n} \log \lambda_\theta(x_i) - \int_A \lambda_\theta(x)\, \mathrm{d}x. \tag{4.2}$$

Note the similarity of equation (4.2) to equation (2.3). The only example of fitting a spatial Poisson process with intensity function of the form (4.1) of which we are aware is due to Vere-Jones and Musmeci (1986). As in the papers mentioned earlier, they assume that $g(y)=y$ and circumvent the problem of a negative intensity function by using penalized likelihood estimation as described earlier.

Because of the similarity of equations (2.3) and (4.2), the theory of Section 3 leading to the use of GLIM extends easily to spatial Poisson processes. In particular, log-likelihood (3.2) holds with $\lambda_\theta(s_j \mid F_{s_j})$ being replaced by $\lambda_\theta(s_j)$, where $s_1, \ldots, s_M$ is a set of design points in A which include the data points, and the w_j are non-negative weights obtained from an appropriate two-dimensional quadrature rule. The various GLIM declarations are just as in the one-dimensional case, and the parameter estimates, likelihood values and covariance matrices produced by GLIM are valid in the sense described in Section 3.

We turn briefly to a discussion of appropriate two-dimensional quadrature rules. Perhaps the most obvious such rule (especially if A is a rectangle) is the 'product rule' (Davis and Rabinowitz, 1975) in which A is divided into a grid of smaller rectangles, the vertices of the rectangles being the design points. The associated quadrature rule has weights which are typically the products of weights obtained from univariate quadrature rules. Unfortunately, the requirement that the design points include the data points implies that, for a product rule, there are at least $(n+2)^2$ design points if A is a rectangle. This quickly strains both central processor unit (CPU) time and GLIM's storage requirements.

Attractive alternatives to product rules are obtained using either the Dirichlet tessellation or the Delaunay triangulation (Green and Sibson, 1978; Ahuja and Schachter, 1983), especially if only approximate parameter estimates and their standard errors are required. Typically, we would use the n data points together with a coarse rectangular grid of dummy points as a base for constructing the tessellation or triangulation. Two simple associated weighting schemes are as follows. In the first scheme, approximate $\lambda_\theta(x)$ by $\lambda_\theta(s_j)$ on the Dirichlet tile surrounding s_j. In the second scheme, approximate $\lambda_\theta(x)$ by a plane on the Delaunay triangle containing x. Both schemes are guaranteed to produce non-negative weights.

Our experience has been that Dirichlet or Delaunay-based quadrature rules give answers comparable with those from product rules using far fewer design points with consequent substantial savings in CPU time. Details are available from the authors.

5. Summarizing Remarks

In this paper, we have shown how approximate ML estimation for fairly general point processes on the line and Poisson processes in space can be performed with GLIM. In addition, we have demonstrated how the use of penalty functions can be avoided, and how, for spatial data, storage and CPU time problems can be overcome by using Dirichlet tessellations or Delaunay triangulations instead of product rules. We hope that our paper will encourage the greater use of point process modelling with the aid of GLIM.

References

Ahuja, N. and Schachter, B. J. (1983) *Pattern Models*, sect. 1.3. New York: Wiley.

Becker, N. G. (1983) Analysis of data from a single epidemic. *Aust. J. Statist.*, **25**, 191–197.

——(1986) A generalized linear modelling approach to the analysis of data from a single epidemic. In *Proc. Pacific Statistical Congr., Auckland*, pp. 464-467. Amsterdam: North-Holland.

Berman, M. (1981) Inhomogeneous and modulated gamma processes. *Biometrika*, **68**, 143–152.

Brillinger, D. R. and Preisler, H. K. (1986) Two examples of quantal data analysis: (a) multivariate point process, (b) pure death process in an experimental design. *13th Int. Biometrics Conf., Seattle.*

Clayton, D. G. (1983) Fitting a general family of failure-time distributions using GLIM. *Appl. Statist.*, **32**, 102–109.

Cox, D. R. (1972) The statistical analysis of dependencies in point processes. In *Stochastic Point Processes* (ed. P. A. W. Lewis), pp. 55–66. New York: Wiley.

Cox, D. R. and Isham, V. (1980) *Point Processes*. London: Chapman and Hall.

Cox, D. R. and Lewis, P. A. W. (1966) *The Statistical Analysis of Series of Events*. London: Methuen.

Davis, P. J. and Rabinowitz, P. (1975) *Methods of Numerical Integration*, sect. 5.6. New York: Academic Press.

Green, P. J. and Sibson, R. (1978) Computing Dirichlet tessellations in the plane. *Comput. J.*, **21**, 168–173.

Hurley, M. A. (1990) Fitting some inhomogeneous point processes. To be published.

Ogata, Y. (1978) Asymptotic behaviour of the maximum likelihood estimators for the stationary point processes. *Ann. Inst. Statist. Math.*, **30**, 243–261.

——(1983) Likelihood analysis of point processes and its applications to seismological data. *Bull. Int. Statist. Inst.*, **50**, 943–961.

——(1988) Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Statist. Ass.*, **83**, 9–27.

Ogata, Y. and Akaike, H. (1982) On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes. *J. R. Statist. Soc.* B, **44**, 102–107.

Ogata, Y. and Katsura, K. (1986) Point-process models with linearly parameterized intensity for application to earthquake data. *J. Appl. Probab.* **23A**, 291–310.

Payne, C. D. (1985) *The GLIM System Release 3.77 Manual*. Oxford: Numerical Algorithms Group.

Vere-Jones, D. and Musmeci, F. (1986) Orthogonal polynomial estimates for point process densities. *Technical Report*. Institute of Statistics and Operations Research, Victoria University of Wellington.

Vere-Jones, D. and Ozaki, T. (1982) Some examples of statistical inference applied to earthquake data. *Ann. Inst. Statist. Math.*, **34**, 189–207.