Ph.D dissertation, Dept. of Elec. Eng., Stanford University, Stanford, CA, May 1976.
[3] J. Rissanen and G. G. Landgon, Jr., "Arithmetic coding," *IBM J. Res. Dev.*, vol. 23, no. 2, pp. 149–162, Mar. 1979.
[4] J. P. M. Schalkwijk, "An algorithm for source coding," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 3, pp. 395–398, May 1972.
[5] T. M. Cover, "Enumerative source encoding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 1, pp. 73–77, Jan. 1973.
[6] N. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1969.
[7] C. B. Jones, "An efficient coding system for long source sequences," (submitted to *IEEE Trans. on Inform. Theory*).
[8] G. N. N. Martin, "Range encoding: An algorithm for removing redundancy from a digitized message," presented at the Video & Data Recording Conf., Southampton, July 1979.
[9] G. G. Landgon, Jr. and J. Rissanen, "Compression of black-white images with binary arithmetic coding," IBM Res. Rep., Dec. 1979 (submitted to *IEEE Trans. Commun.*).
[10] A. Kolmogorov, "Three approaches to the quantitative definition of information," *Prob. Peredach. Inform.*, vol. 1, no. 1, pp. 3–11, 1965, (Russian).
[11] J. Rissanen, "Arithmetic codings as number representations," *Acta Polytech. Scandinavica*, Math. 31, pp. 44–51, Dec. 1979.
[12] S. K. Leung-Yan-Cheong and T. Cover, "Some equivalences between shannon entropy and Kolmogorov complexity," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 3, pp. 331–338, May 1978.
[13] T. J. Lynch, "Sequence time coding for data compression," *Proc. IEEE* (Lett.), vol. 54, pp. 1490–1491, Oct. 1966.
[14] L. D. Davisson, "Comments on 'Sequence time coding for data compression,'" *Proc. IEEE (Lett.), vol. 54, p. 2010, Dec. 1966.*

On Lewis' Simulation Method for Point Processes

YOSIHIKO OGATA

***Abstract*—A simple and efficient method of simulation is discussed for point processes that are specified by their conditional intensities. The method is based on the thinning algorithm which was introduced recently by Lewis and Shedler for the simulation of nonhomogeneous Poisson processes. Algorithms are given for past dependent point processes containing multivariate processes. The simulations are performed for some parametric conditional intensity functions, and the accuracy of the simulated data is demonstrated by the likelihood ratio test and the minimum Akaike information criterion (AIC) procedure.**

I. Introduction

ANY point process (N_t, F_t, P) on a finite interval $(0, T]$ is a submartingale and therefore by the Doob–Meyer decomposition may be written as $N_t = m_t + A_t$, where m_t is an (F_t, P) martingale and A_t is the natural increasing process. It is known that there is a predictable process (λ_t, F_t), such that $A_t = \int_0^t \lambda_s\, ds$, if and only if P is absolutely continuous with respect to the standard Poisson process P_0; furthermore $\lambda = \{\lambda_t,\ 0 < t \leqslant T\}$ corresponds uniquely to the process P, and the Radon–Nikodym derivative is given by

$$\frac{dP}{dP_0} = \exp\left\{ \int_0^T \log \lambda_t\, dN_t + \int_0^T (1-\lambda_t)\, dt \right\}.$$

Manuscript received April 16, 1979; revised April 22, 1980. This paper was presented at the Annual Meeting of the Japan Mathematical Society, Nagoya, Japan, April 1979.

The author is with the Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan.

Similar results hold for multivariate or marked point processes [7], [8].

The main object of this paper is to discuss the applications of Lewis' thinning simulation algorithm to any point process which is absolutely continuous with respect to the standard Poisson process. Recently Ozaki [11] generated simulation data for Hawkes' self-exciting processes by making use of a recursive structure. However his method is not fast enough unless the process has a simple structure, because given a past history of the process $t_1, t_2, \cdots, t_n$ and a uniform random number U_{n+1} from the interval $(0, 1)$, we have to solve the equation $U_{n+1} = S(t_{n+1} | t_1, \cdots, t_n)$ by Newton's iterative method to get the next point t_{n+1}, where S is the conditional survivor function

$$S(t | t_1, \cdots, t_n) = \exp\left\{ -\int_{t_n}^{t} \lambda(s | t_1, \cdots, t_n)\, ds \right\}.$$

We do not need to solve this equation to get the next point. The idea of simulating these point processes by thinning is developed using algorithms due to Lewis and Shedler [9] for the simulation of nonhomogeneous Poisson processes.

In Section II we give the simulation method and a proof for past dependent point processes containing multivariate processes. Some typical algorithms also will be given. In Section III we give some examples of parametric intensity functions for the simulation and obtain their

0018-9448/81/0100-0023$00.75

maximum-likelihood estimates from the simulated data. The accuracy of the simulated data will be discussed by the likelihood ratio test or the minimum Akaike information criterion (AIC) procedure. Numerical results are given in the Appendix.

II. Simulation of Point Processes

Consider a point process $(N, F, P)=\{N_t, F_t,\ 0<t\leqslant T, P\}$ on a fixed interval with its F-predictable intensity process $\lambda=\{\lambda_t\}$, where $F=\{F_t\}$ is a family of right continuously increasing σ-fields. Suppose we obtain a positive F-predictable piecewise constant process $\lambda^*=\{\lambda_t^*\}$ which is constructed pathwise in such a way that $\lambda_t\leqslant\lambda_t^*$, almost surely (a.s.), $0<t\leqslant T$. Then λ^* can be an intensity process of a locally homogeneous Poisson process $(N^*, F, P)=\{N_t^*, F_t, P\}$ with piecewise constant intensity changing its rate according to the past history F_t. The main result, which is formally similar to the one given in [9], is as follows.

Let $t_1^*<t_2^*<,\cdots,<t_{N_T^*}^*$ be the points in $(0, T]$ of the process (N^*, F, P). Delete the points t_j^* with probability $1-\lambda_{t_j}/\lambda_{t_j}^*$ for $j=1,2,\cdots,N_T^*$. Then the remaining points $\{t_i\}$ form a point process (N, F, P) with conditional intensity $\lambda=\{\lambda_t\}$ in the interval $(0, T]$.

It is readily seen from the predictability of λ^* that the constructions of λ^*, t_j^*, and t_i should be performed sequentially in the following manner.

1) Suppose that the last point before time t has just been obtained. Then construct λ_t^* which is F_{t_i}-measurable, piecewise constant, and $\lambda_t^*\geqslant\lambda_t$, for $t\geqslant t_i$.

2) Simulate homogeneous Poisson points $t_j^*(>t_i)$ according to the intensity λ_t^*.

3) For each of the points $\{t_j^*\}$, the probability $\lambda_{t_j^*}/\lambda_{t_j^*}^*$ is given conditionally independent of t_j^* under the past history $F_{t_j^*}$.

4) t_{i+1} is the first accepted point among $t_j^*(>t_i)$.

Details of the algorithm will be given later. By generalizing this result to a multivariate point process, we have the following proposition.

Proposition 1: Consider a multivariate point process (N^p, F, P), $p=1,2,\cdots,m$, on an interval $(0, T]$ with joint intensity $(\lambda, F)=\{\lambda_t^p, F_t\}$, $p=1,\cdots,m$. Suppose we can find a one-dimensional F-predictable process λ_t^* which is defined pathwise satisfying

$$\sum_{p=1}^{m}\lambda_t^p\leqslant\lambda_t^*,\qquad 0<t\leqslant T,$$

P-almost surely, and set

$$\lambda_t^0=\lambda_t^*-\sum_{p=1}^{m}\lambda_t^p.$$

Let $t_1^*, t_2^*,\cdots,t_{N_T^*}^*\in(0, T]$ be the points of the process (N^*, F, P) with intensity process λ_t^*. For each of the points, attach a mark $p=0,1,\cdots,m$ with probability $\lambda_{t_j^*}^p/\lambda_{t_j^*}^*$. Then the points with marks $p=1,2,\cdots,m$, provide a multivariate point process which is the same as that given above.

Proof: Define a random measure for the finite marked process (using the notation in [7]) by

$$M(dt, p)=N^*(dt)I(t, p),\qquad p=1,\cdots,m,$$

where $I(t, p)$ is the transition random measure of the marks under the condition that there is a point at t. As a result of the conditions of the proposition, $I(t, p)$ has the following properties.

i) $I(t, q)=\delta_p(q)$ with probability λ_t^p/λ_t^* for $p, q=0,1,2,\cdots,m$, where $\delta_p(q)$ is a Dirac delta function.

ii) For fixed t, $N^*(dt)$ and $\{I(t, p), p=0,1,\cdots,m\}$ are conditionally independent given F_t.

Then for each mark $p=1,2,\cdots,m$, the intensity measure of the marked point process is given by

$$\begin{aligned}\nu(dt, p)&=E[\,M(dt, p)|F_t]\\&=E[\,N^*(dt)I(t, p)|F_t]\\&=E[\,N^*(dt)|F_t]\,E[\,I(t, p)|F_t].\end{aligned}$$

By definition

$$E[\,N^*(dt)|F_t]=\lambda_t^*\,dt,$$

and also

$$\begin{aligned}E[\,I(t, p)|F_t]&=1\cdot P\{I(t, p)=1|F_t\}\\&\quad+0\cdot P\{I(t, p)=0|F_t\}\\&=\lambda_t^p/\lambda_t^*.\end{aligned}$$

Therefore for each $p=1,2,\cdots,m$, we get $\nu(dt, p)=\lambda_t^p\,dt$. Since the predictable random measure corresponds uniquely to the multivariate process (see [7], [8]), this completes the proof. □

We now give some typical algorithms based on this proposition.

Algorithm 1: A bivariate (doubly) Poisson process with intensity process $\{X_t^p(\omega), F_t\}$, $p=1,2$.

1) Obtain a path function of the process $\omega_p(t)=X_t^p(\omega)$, $0<t\leqslant T$, $p=1,2$.
2) Take a piecewise constant function $\omega^*(t)$ such that $\omega_1(t)+\omega_2(t)\leqslant\omega^*(t)$. For efficiency of simulation we should take $\omega^*(t)$ as close as possible to $\omega_1(t)+\omega_2(t)$.
3) Simulate stationary Poisson processes for each interval of constant intensity. Denote the points by $t_1^*, t_2^*,\cdots,t_{N_T^*}^*$.
4) Set $k=1$, $i=0$, and $j=0$.
5) Independently generate a uniform random number U_k on $(0,1)$.
6) If $U_k\leqslant\omega_1(t_k^*)/\omega^*(t_k^*)$, set i equal to $i+1$ and $t_i^{(1)}=t_k^*$.
7) If $U_k\leqslant\{\omega_1(t_k^*)+\omega_2(t_k^*)\}/\omega^*(t_k^*)$, set j equal to $j+1$ and $t_j^{(2)}=t_k^*$.
8) Set k equal to $k+1$. If $N_T^*<k$, then stop. Otherwise go to step 5.

Consider the case of a univariate self-exciting process. Since $F_t=\sigma\{N_s,\ 0<s\leqslant t\}$, the intensity of the process is given by a function of t and the points t_i before t, i.e. $\lambda_t=\lambda(t|t_1,\cdots,t_n)$. There are two types of intensity processes.

Consider first the case where the path function of the intensity process is decreasing if no more points occur. The predictability of λ_t implies left-continuity of the path functions. We assume that the minimum value of the intensity function is μ, the jump size at each point is not larger than α, and the Λ_i^* are values, of the piecewise constant function such that $\lambda(t|t_1,\cdots,t_n)\leqslant\Lambda_i^*$ for $t_n\leqslant s_i\leqslant t<s_{i+1}\leqslant t_{n+1}$.

Algorithm 2:

1) Set $\Lambda_0^*=\mu$ and put $s_0=0$.
2) Generate U_0 and put $u_0=-\log(U_0/\Lambda_0^*)$.
3) If $u_0\leqslant T$ then put $t_1=u_0$. Otherwise stop.
4) Set $i=j=k=0$ and $n=1$.
5) Set k equal to $k+1$ and put $\Lambda_k^*=\lambda(t_n|t_1,\cdots,t_{n-1})+\alpha$.
6) Set j equal to $j+1$ and generate U_j.
7) Set i equal to $i+1$ and put $u_i=-\log(U_j/\Lambda_k^*)$.
8) Put $s_i=s_{i-1}+u_i$. If $s_i>T$, stop.
9) Set j equal to $j+1$, and generate U_j.
10) If $U_j\leqslant\lambda(s_i|t_1,\cdots,t_{n-1})/\Lambda_k^*$, set n equal to $n+1$, put $t_n=s_i$ and go to step 5.
11) Set k equal to $k+1$, put $\Lambda_k^*=\lambda(s_i|t_1,\cdots,t_{n-1})$ and go to step 6.

If a sample function of the intensity function $\lambda(t|t_1,\cdots,t_n)$ is not always decreasing but only has a decreasing tail, then we can define a process $\lambda^{**}(t|t_1,\cdots,t_n)$ which is always decreasing and satisfies $\lambda(t|t_1,\cdots,t_n)\leqslant\lambda^{**}(t|t_1,\cdots,t_n)$ for $t_n\leqslant t$ (see Example 1). Replace $\lambda(t_n|t_1,\cdots,t_{n-1})$ with $\lambda^{**}(t_n|t_1,\cdots,t_{n-1})$ in step 5 and $\lambda(s_i|t_1,\cdots,t_{n-1})$ with $\lambda^{**}(s_i|t_1,\cdots,t_{n-1})$ in step 11. Then we obtain a point process with intensity λ_t by using the modified Algorithm 2.

Assume now that the intensity function $\lambda_t(\omega)=\lambda(t|t_1,\cdots,t_n)$ is monotonically increasing if no more points occur. In the following algorithm, the interval $(0,T]$ is divided equally into subintervals $(kr,(k+1)r]$, for some appropriate choice of the length r.

Algorithm 3:

1) Set $i=n=1$.
2) Put $\lambda_i^*=\lambda((i+1)r|t_1,\cdots,t_n)$.
3) Generate a homogeneous Poisson process with intensity λ_i^* on the interval $(kr,(k+1)r]$.
4) If the number of the points on the interval, say N_i^*, is zero, go to step 11.
5) Denote the ordered points on the interval $(ir,(i+1)r]$ by $s_1^*,s_2^*,\cdots,s_{N_i^*}^*$.
6) Set $j=1$.
7) Generate U_j uniformly distributed between zero and one.
8) If $U_j>\lambda(s_j^*|t_1,\cdots,t_n)/\lambda_i^*$, go to step 7.
9) Put $t_n=s_j^*$ and set n equal to $n+1$.
10) Set j equal to $j+1$. If $j\leqslant N_i^*$ go to step 7.
11) Set i equal to $i+1$. If $(i+1)r\leqslant T$ go to step 2.
12) Stop.

Thus $t_1,t_2,\cdots$ are the data which are required. It is recommended for numerical accuracy in step 4 to adopt a method which generates a Poisson random number N_i^* and then uniform random numbers from $(ir,(i+1)]$ according to N_i^*.

If $\lambda_t=\lambda_t(t|t_1,\cdots,t_n)$ is only eventually increasing (if no more points occur), a modification similar to that for Algorithm 2 is possible. That is to say, construct a function $\lambda^{**}(t|t_1,\cdots,t_n)$ which is increasing in $(ir,(i+1)r]$, and satisfies $\lambda(t|t_1,\cdots,t_n)\leqslant\lambda^{**}(t|t_1,\cdots,t_n)$ for $ir<t\leqslant(i+1)r$. Then change step 2 of Algorithm 2 to

$$2^*\cdot\lambda_i^*=\lambda^{**}((i+1)r|t_1,\cdots,t_n).$$

It is not difficult to construct simulation algorithms for multivariate mutually exciting point processes, or mixed doubly Poisson and self-exciting point processes from the above algorithms.

III. Some Examples and Discussions

Hawkes' Self-Exciting Process

The intensity function is given by

$$\lambda(t)=\lambda(t|t_1,\cdots,t_n)=\mu+\int_0^t\nu(t-s)\,dN(s),$$

where $\mu>0$, $\nu(s)\geqslant 0$, and $\int_0^\infty\nu(s)\,ds<1$ for asymptotic stationarity of the process. Hawkes and Oakes [5] first gave the author the idea that this process may be simulated through nonhomogeneous Poisson processes. Indeed they say that a Hawkes' self-exciting process is nothing but an immigrant-birth process which is composed of an homogeneous Poisson immigrant with rate μ and nonhomogeneous Poisson descendants with rate $\nu(s)$.

As a parametrization of $\nu(s)$, Hawkes [4] used an exponential $\nu(s)=\alpha e^{-\beta s}$. In this case we can apply Algorithm 2 for the simulation. Ozaki and Akaike [12] suggested a generalized parametrization $\nu(s)=\sum_{j=0}^p\alpha_j s^j e^{-\beta s}$, where the α and β (>0) are restricted to satisfy $\nu(s)\geqslant 0$ and $\int_0^\infty\nu(s)\,ds<1$. This is a decreasing function for sufficiently large s. Thus we have a function $\nu^{**}(s)$ which is always decreasing and $\nu(s)\leqslant\nu^{**}(s)$ for $s\geqslant 0$, say $\nu^{**}(s)=\sum_{j=0}^p\alpha_j^+\max\{(j/\beta)^j e^{-j}\cdot 1_{[0,j/\beta]},s^j e^{-\beta s}\}$ where $\alpha_j^+=\max(\alpha_j,0)$. Therefore making use of a predictable intensity $\lambda^{**}(s)=\mu+\int_0^t\nu^{**}(s)\,dN(s)$, where $N(s)$ is the point process generated by the intensity function $\lambda(t)=\mu+\int_0^t\nu(t-s)\,dN(s)$, the modified Algorithm 2 can be applied for the simulation. The jump size for this case is $\sum_{j=0}^p\alpha_j^+(j/\beta)^j e^{-j}$.

Suppose $t_1,t_2,\cdots,t_n$ in $(0,T]$ are the simulated data. Then the log-likelihood function is given by

$$L_T(\alpha_0,\cdots,\alpha_p,\beta)=\sum_{i=1}^n\log\left\{\mu+\sum_{j=0}^p\alpha_j R_j(i)\right\}-\mu T-\sum_{i=1}^n\sum_{j=0}^p\alpha_j S_j(T-t_i),$$

where $R_j(i)$ and $S_j(t)$ are given recursively in the following way.

Set $t_0=0$, $R_0(1)=0$, $S_0(t)=(1-e^{-\beta t})/\beta$, and $A_k(t)=t^k e^{-\beta t}$. Then for $j=0,1,2,\cdots$, and $i=2,3,\cdots$,

$$R_j(i)=A_j(t_i-t_{i-1})+\sum_{k=0}^{j} {}_jC_k A_{j-k}(t_i-t_{i-1})R_k(i-1),$$

and

$$S_{j+1}(t)=\{(j+1)S_j(t)-A_{j+1}(t)\}/\beta,$$

where ${}_jC_k$ denotes a binomial coefficient.

The gradient vector and Hessian matrix of the log-likelihood function also be written recursively using the above function. It is worth noting that the simulation is much faster if we make use of the recursive structure of the intensity function $\lambda(t|t_1,\cdots,t_n)$, that is,

$$\lambda(t_{n+1}|t_1,\cdots,t_n)=\mu+\sum_{j=0}^{p}\alpha_j R_j(n+1).$$

Linear Wold Process

The intensity function is given by

$$\lambda(t)=\mu+\alpha_1(t-t_{(1)})+\sum_{k=2}^{p}\alpha_k(t_{(k-1)}-t_{(k)}),$$

where μ and α are nonnegative parameters and $t_{(k)}$ is the kth last point before t. A point process with this intensity is always asymptotically stationary (see [3]). It is easily seen that this point process may be simulated by the simple relation

$$t_{n+1}=t_n+\frac{1}{\alpha_1}\left\{-\beta_n+(\beta_n^2-2\alpha_1\log U_{n+1})^{1/2}\right\},$$

where $\beta_n=\mu+\sum_{k=2}^{p}\alpha_k(t_{n-k+2}-t_{n-k+1})$ and U_{n+1} is a uniform random number from $(0,1)$. However, we would like to apply the modified version of Algorithm 3, setting

$$\lambda^{**}(t|t_1,\cdots,t_n)=\mu+\max_{0<i\leq k}\alpha_i(t-t_{(k)}).$$

Suppose $t_1,t_2,\cdots,t_n$ are the simulation data on the interval $(0,T]$. Then setting $t_0=0$ and $t_{n+1}=T$, the log-likelihood function is given by

$$L_T(\mu,\alpha_1,\cdots,\alpha_p)=\sum_{i=1}^{n}\log\left\{\mu+\sum_{k=1}^{p}\alpha_k(t_{i-k+1}-t_{i-k})\right\}-\mu T-\sum_{i=1}^{n+1}\left\{\alpha_1(t_i-t_{i-1})^2/2+\sum_{k=2}^{p}\alpha_k(t_i-t_{i-1})(t_{i-k+1}-t_{i-k})\right\}.$$

One of the nice properties of this model is that the Hessian matrix of the log-likelihood function is negative-definite everywhere with respect to the parameters (see [10, p. 255] for example).

Stress-Release Process

Vere-Jones [14] has suggested models for a series of strong earthquakes. One of these models is defined by the intensity function

$$\lambda(t)=e^{\alpha+\beta t-\gamma N(t-)},\qquad \text{where } N(t-)=N[0,t).$$

(A similar process is discussed by Isham and Westcott [6].) This is asymptotically stationary with a mean intensity rate γ/β. Although this process is obtained by the simple relation

$$t_{n+1}=t_n+\frac{1}{\beta}\log\left\{1-\beta e^{-\alpha-\beta t_n+\gamma n}\cdot\log U_{n+1}\right\},$$

we apply Algorithm 3 for the simulation.

Given simulated data $t_1,t_2,\cdots,t_n$ on the interval $(0,T]$, we have the log-likelihood function (setting $t_0=0$ and $t_{n+1}=T$)

$$L_T(\alpha,\beta,\gamma)=\sum_{i=1}^{n}\log\{\alpha+\beta t_i-\gamma(i-1)\}+\sum_{i=1}^{n+1}e^{\alpha-\gamma(i-1)}\cdot\{e^{\beta t_{i-1}}-e^{\beta t_i}\}/\beta.$$

Nonlinear Hawkes' Type Point Process

Consider an intensity function of the form

$$\lambda(t)=\mu+\int_0^t\nu(t-s)\,dN(s)+\int_0^t\int_0^s\pi(t-s,s-u)\,dN(s)\,dN(u),$$

where $\mu>0$, $\nu(s)\geq 0$, and $\pi(s,\mu)\geq 0$. It is necessary for asymptotic stationarity that

$$\nu+\pi_0<1,\quad \text{and}\quad (1-\nu-\pi_0)^2\geq 4\pi(\mu+\pi_C),$$

where

$$\nu=\int_0^\infty\nu(s)\,ds,$$

$$\pi_0=\int_0^\infty\pi(s,0)\,ds,$$

$$\pi=\int_0^\infty\int_0^\infty\pi(s,u)\,ds\,du,$$

$$\pi_C=\int_0^\infty\int_0^\infty\pi(s,u)C(u)\,ds\,du.$$

($C(u)$ is the autocovariance of the process). Unfortunately we can evaluate neither π_C nor $C(u)$. We can only hypothesize the domain. For example, the noise level μ of the Poisson must be small enough for asymptotic stationarity.

A parametric example of $\nu(s)$ and $\pi(s,\mu)$ is

$$\nu(s)=\alpha e^{-\beta s}\qquad \pi(s,u)=\gamma e^{-\beta(s+u)}.$$

Algorithm 2 is applicable to this case. Suppose we obtain simulation data $t_1,t_2,\cdots,t_n$ on the interval $(0,T]$. Then the log-likelihood function is

$$L_T(\mu,\alpha,\beta,\gamma)=\sum_{i=1}^{n}\log\{\mu+\alpha R_1(i)+\gamma R_3(1)\}-\mu T-\frac{\alpha}{\beta}\sum_{i=1}^{n}\{1-e^{-\beta(T-t_i)}\}-\frac{\gamma}{\beta}\sum_{i=1}^{n}\{1-e^{-\beta(T-t_i)}\}R_2(i),$$

where $R_1(i)$, $R_2(i)$, and $R_3(i)$ are given recursively by

$$R_1(1)=0,\quad R_2(1)=1,\quad R_3(i)=0,$$
$$R_1(i)=e^{-\beta(t_i-t_{i-1})}\cdot\{R_1(i-1)+1\},$$
$$R_2(i)=e^{-\beta(t_i-t_{i-1})}\cdot R_2(i-1)+1,$$
$$R_3(i)=e^{-\beta(t_i-t_{i-1})}\cdot\{R_2(i-1)+R_3(i-1)\}.$$

The gradient and Hessian are given similarly. It is worth noting that the simulation is much faster if we make use of the recursive structure.

Bivariate Wold Process

The intensity functions are given by

$$\begin{bmatrix}\lambda_1(t)\\ \lambda_2(t)\end{bmatrix}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}+\begin{bmatrix}\alpha_1 & \beta_1\\ \gamma_1 & \delta_1\end{bmatrix}\begin{bmatrix}t-t_{(1)}^{(1)}\\ t-t_{(1)}^{(2)}\end{bmatrix}+\sum_{k=2}^{p}\begin{bmatrix}\alpha_k & \beta_k\\ \gamma_k & \delta_k\end{bmatrix}\begin{bmatrix}t_{(k-1)}^{(1)}-t_{(k)}^{(1)}\\ t_{(k-1)}^{(2)}-t_{(k)}^{(2)}\end{bmatrix}.$$

The coefficients must be nonnegative for asymptotic stationarity if μ_1 and μ_2 are strictly positive. If $\mu_1=0$, or $\mu_2=0$, then there are explosive cases even if the other coefficients are nonnegative. For example, consider the simplest case $p=1$. If $\mu_1>0$, $\mu_2=0$, $\alpha_1=0$, $\beta_1>0$, $\gamma_1>0$, and $\delta_1=0$, we have an explosive process.

Consider an asymptotically stationary case with $p=1$. Let $\{t_i^{(1)}\}$, $i=1,2,\cdots,n$, and $\{t_j^{(2)}\}$, $j=1,2,\cdots,m$, be the data from the model. Then the log-likelihood function is given by

$$L_T(\mu_1,\mu_2,\alpha,\beta,\gamma,\delta)=L_T^{(1)}(\mu_1,\alpha,\beta)+L_T^{(2)}(\mu_2,\gamma,\delta),$$
$$L_T^{(1)}=\sum_{i=1}^{n}\log\{\mu_1+\alpha(t_i^{(1)}-t_{i-1}^{(1)})+\beta(t_i^{(1)}-t_{(i)}^{(2)})\}-\mu_1T-\alpha\sum_{i=1}^{n}(t_i^{(1)}-t_{i-1}^{(1)})^2/2-\beta\sum_{j=1}^{m}(t_j^{(2)}-t_{j-1}^{(2)})^2/2,$$
$$L_T^{(2)}=\sum_{j=1}^{m}\log\{\mu_2+\gamma(t_j^{(2)}-t_{(j)}^{(1)})+\delta(t_j^{(2)}-t_{j-1}^{(2)})\}-\mu_2T-\delta\sum_{j=1}^{m}(t_j^{(2)}-t_{j-1}^{(2)})^2/2-\gamma\sum_{i=1}^{n}(t_i^{(1)}-t_{i-1}^{(1)}),$$

where $t_{(i)}^{(2)}$ is the last point on the line $t^{(2)}$ before $t_i^{(1)}$, $t_{(j)}^{(1)}$ is the last point on the line $t^{(1)}$ before $t_j^{(2)}$, and $t_0^{(1)}=t_0^{(2)}=0$. Notice that the minimization of L_T is equivalent to the separate minimizations of $L_T^{(1)}$ and $L_T^{(2)}$, provided the parameters are independent.

The simulation is performed using Algorithm 1 and the modified version of Algorithm 3.

Bivariate Hawkes' Mutually Exciting Process [4]

The intensity functions are given by

$$\lambda_1(t)=\mu_1+\int_0^t\nu_{11}(t-s)\,dN_1(s)+\int_0^t\nu_{12}(t-s)\,dN_2(s),$$
$$\lambda_2(t)=\mu_2+\int_0^t\nu_{21}(t-s)\,dN_1(s)+\int_0^t\nu_{22}(t-s)\,dN_2(s),$$

where $\nu_{ij}(s)\geqslant 0$. For asymptotic stationarity it is necessary that the moduli of all eigenvalues of the matrix $\{\nu_{ij}\}$ are less than one, where $\nu_{ij}=\int_0^\infty\nu_{ij}(s)\,ds$. Parametrizing the functions $\nu_{ij}(s)=\alpha_{ij}e^{-\beta_i s}$, we can simulate the data in accordance with Algorithms 1 and 2. For data $\{t_i^{(1)}\}$, $i=1,2,\cdots,n$, and $\{t_j^{(2)}\}$, $j=1,2,\cdots,m$, the log-likelihood of the model is

$$L_T(\mu_1,\mu_2,\beta_1,\beta_2,\alpha_{11},\alpha_{12},\alpha_{21},\alpha_{22})=L_T^{(1)}(\mu_1,\beta_1,\alpha_{11},\alpha_{12})+L_T^{(2)}(\mu_2,\beta_2,\alpha_{21},\alpha_{22}),$$
$$L_T^{(1)}(\mu_1,\beta_1,\alpha_{11},\alpha_{12})=\sum_{i=2}^{n}\log\{\mu_1+\alpha_{11}R_{11}(i)+\alpha_{12}R_{12}(i)\}-\mu_1T-\frac{\alpha_{11}}{\beta_1}\sum_{i=1}^{n}\{1-e^{-\beta_1(T-t_i^{(1)})}\}-\frac{\alpha_{12}}{\beta_1}\sum_{j=1}^{m}\{1-e^{-\beta_1(T-t_j^{(2)})}\},$$
$$L_T^{(2)}(\mu_2,\beta_2,\alpha_{21},\alpha_{22})=\sum_{j=2}^{m}\log\{\mu_2+\alpha_{21}R_{21}(j)+\alpha_{22}R_{22}(j)\}-\mu_2T-\frac{\alpha_{21}}{\beta_2}\sum_{j=1}^{m}\{1-e^{-\beta_2(T-t_j^{(1)})}\}-\frac{\alpha_{22}}{\beta_2}\sum_{j=1}^{m}\{1-e^{-\beta_2(T-t_j^{(2)})}\},$$

where the R_{ij} are given recursively by

$$R_{11}(1)=R_{12}(1)=R_{21}(1)=R_{22}(1)=0,$$
$$R_{11}(i)=e^{-\beta_1(t_i^{(1)}-t_{i-1}^{(1)})}\cdot\{1+R_{11}(i-1)\},$$
$$R_{12}(1)=e^{-\beta_1(t_i^{(1)}-t_{i-1}^{(1)})}\cdot R_{12}(i-1)+\sum_{\{j:\,t_{i-1}^{(1)}\leqslant t_j^{(2)}<t_i^{(1)}\}}e^{-\beta_1(t_i^{(1)}-t_j^{(2)})},$$
$$R_{21}(j)=e^{-\beta_2(t_j^{(2)}-t_{j-1}^{(2)})}\cdot R_{21}(j-1)+\sum_{\{i:\,t_{j-1}^{(2)}\leqslant t_i^{(1)}<t_j^{(2)}\}}e^{-\beta_2(t_j^{(2)}-t_i^{(1)})},$$

and

$$R_{22}(j)=e^{-\beta_2(t_j^{(2)}-t_{j-1}^{(2)})}\cdot\{1+R_{22}(j-1)\}.$$

The gradients vector and Hessian matrix are given similarly by a recursive formula. Also the simulation algorithm should be performed recursively for the greatest efficiency.

IV. Concluding Remarks

In this section we discuss whether the simulation data are statistically accurate enough in each case. In [10] a collection of regularity conditions is given to prove the following.

1) The maximum likelihood estimator is consistent, i.e., $\hat{\theta}_T\to\theta_0$ a.s. as $T\to\infty$.

2) $\sqrt{T}\,(\hat{\theta}_T-\theta_0)$ is asymptotically normal according to $\mathcal{N}(\mathbf{0}, I(\theta_0)^{-1})$ as $T\to\infty$, where the components of Fisher's mean-rate information matrix are given by $I_{ij}(\theta_0)=E\{(1/\lambda_\theta)(d\lambda/d\theta_i)(d\lambda/d\theta_j)\}_{\theta=\theta_0}$.

3) $2\{L_T(\hat{\theta})-L_T(\theta_0)\}$ is asymptotically χ_k^2-distributed as $T\to\infty$, where k is the dimension of the parameter θ.

The examples in the preceding section basically satisfy these conditions, although the multivariate case is not treated there. So the adoption of the minimum AIC procedure [1] is justified. That is to say, we consider two competing models H_0 and H_1, where H_0 is the model supposed to have the true parameter θ_0, and H_1 is any other model with a fixed dimension k of the parameter θ. The values of the AIC for the two models are

$$\mathrm{AIC}_0=(-2)(\text{value of log-likelihood at }\theta_0),$$

since the number of unknown parameters in H_0 is zero, and

$$\mathrm{AIC}_1=(-2)(\text{maximized value of log-likelihood})+2k.$$

The AIC is an estimator of the expected negative entropy which is a natural measure of the discrimination between the true distribution for the data and the estimated probability distribution. Therefore, if the simulated data are correctly distributed according to H_0, we can expect $\mathrm{AIC}_0<\mathrm{AIC}_1$. Another useful method is to adopt the likelihood ratio test of H_0 against H_1. Under regular situations such as the nested sequence of models, the relationship between the AIC and the likelihood ratio statistic $\Delta(\theta_0,\hat{\theta}_T)$ is given by

$$\Delta(\theta_0,\hat{\theta}_T)=\mathrm{AIC}_0-\mathrm{AIC}_1+2k,$$

which is asymptotically χ_k^2-distributed. See [13] for a extensive discussion of the relation between the minimu AIC procedure and the likelihood ratio test.

Using physically generated random numbers we pe formed simulation experiments five times for each exar ple. The maximum-likelihood estimates and the negativ of the log-likelihoods are listed in the tables in the A pendix. We used the Davidon–Flecher–Powell meth for the nonlinear optimization. From the tables we can s that the maximum-likelihood estimates get more accura as the sample size (number of points) or the length of t observed interval increases. Also for each sample size interval length, the AIC and log-likelihood ratio te work well and justify the accuracy of the simulations.

ACKNOWLEDGMENT

The idea of the thinning method was communicated Professor P. A. W. Lewis, who kindly corrected the earli versions of this paper. For this the author is very gratef The author also would like to thank G. Kitagawa f providing a program in [2] and for valuable suggestio such as the recursive calculation technique, and K. Katsu for his generous help in testing the algorithms. Finally t author would like to thank Professor Okamoto for h many critical comments.

APPENDIX

TABLE I

Hawkes' self-exciting process

$$\lambda(t)=\mu+\int_0^t\left\{\left(\alpha_0+\alpha_1(t-s)+\alpha_2(t-s)^2\right\}e^{-\beta(t-s)}\,dN(s),\right.$$

where

$$\begin{pmatrix}\mu,\beta\\ \alpha_0,\alpha_1,\alpha_2\end{pmatrix}=\begin{pmatrix}0.700, 1.100\\ 0.045, -0.300, 0.500\end{pmatrix}.$$

Numbers of Data	−Log(Likelihood) at the True Parameter	−Log(Maximum of Likelihood)	Maximum Likelihood Estimates		
$n=500$	218.895	216.715	0.816	0.970	
			−0.136	−0.161	0.382
	214.112	213.181	0.760	1.125	
			0.113	−0.427	0.570
	308.358	307.341	0.794	1.234	
			0.089	−0.288	0.531
	368.521	365.038	0.923	1.213	
			0.030	−0.412	0.487
	292.764	289.974	0.540	1.095	
			0.257	−0.434	0.494
$n=50000$	26376.292	26374.593	0.721	1.115	
			0.032	−0.305	0.520
	26104.488	26099.494	0.716	1.099	
			0.045	−0.378	0.539
	25925.125	25922.638	0.707	1.134	
			0.040	−0.344	0.567
	25835.985	25832.676	0.724	1.108	
			0.036	−0.361	0.552
	25574.279	25570.103	0.715	1.107	

TABLE II

Linear Wold process

$$\lambda(t)=\mu+\alpha(t-t_{(1)})+\beta(t_{(1)}-t_{(2)})+\gamma(t_{(2)}-t_{(3)}),$$

where

$$(\mu,\alpha,\beta,\gamma)=(2.000,1.400,3.900,2.700).$$

Number of Data	−Log(Likelihood at the true Parameter)	−Log(Maximum of Likelihood)	Maximum Likelihood Estimates			
$n=500$	−158.717	−160.946	1.637	1.505	3.459	3.273
	−178.251	−179.852	2.540	1.067	2.624	1.814
	−165.349	−167.672	2.192	0.529	2.874	3.075
	−199.151	−203.298	1.418	3.049	3.490	4.649
	−190.100	−190.534	1.750	1.861	4.632	2.726
$n=50000$	−18794.831	−18796.546	2.008	1.386	3.792	2.802
	−18862.763	−18865.039	1.944	1.474	3.932	2.863
	−18655.569	−18656.813	2.002	1.458	3.797	2.680
	−18664.021	−18664.801	2.005	1.426	3.806	2.695
	−18766.239	−18767.236	1.978	1.358	3.953	2.785

TABLE III

Stress-release process

$$\lambda(t)=e^{\alpha+\beta t-\gamma N[0,t)},$$

where

$$(\alpha,\beta,\gamma)=(3.000,2.000,1.000).$$

Number of Data	−Log(Likelihood) at the True Parameter	−Log(Maximum of Likelihood)	Maximum Likelihood Estimates		
$n=500$	4.318	4.012	2.957	1.844	0.924
	4.419	2.621	2.662	2.002	0.995
	5.513	4.530	2.855	1.808	0.906
	−3.353	−4.842	3.374	2.529	1.262
	7.023	5.656	2.678	1.641	0.820
$n=50000$	621.435	619.599	2.993	2.017	1.009
	623.923	623.372	2.994	2.007	1.004
	580.052	578.178	3.024	2.046	1.023
	583.369	581.362	3.014	2.041	1.021
	631.726	629.010	3.003	2.007	1.004

TABLE IV

A Hawkes' type nonlinear process

$$\lambda(t)=\mu+\int_0^t \alpha e^{-\beta(t-s)}\,dN(s)+\int_0^t\int_0^s \gamma e^{-\beta(t-u)}\,dN(s)\,dN(u),$$

where

$$(\mu,\alpha,\beta,\gamma)=(0.550,0.850,4.750,0.350).$$

Number of Data	−Log(Likelihood at the True Parameter)	−Log(Maximum of Likelihood)	Maximum Likelihood Estimates			
$n=500$	591.036	590.325	0.564	0.476	5.682	0.677
	458.429	452.586	0.677	0.507	6.166	0.859
	543.007	542.315	0.554	1.326	5.106	0.149
	548.165	545.468	0.543	1.225	3.355	−0.083
	561.440	557.988	0.600	1.393	6.998	0.226
$n=50000$	56876.433	56870.358	0.560	0.778	4.758	0.372
	56632.821	56632.009	0.552	0.842	4.669	0.338
	56383.413	56382.695	0.554	0.824	4.791	0.372
	56524.242	56521.422	0.556	0.884	4.916	0.347
	56708.677	56707.395	0.551	0.804	4.653	0.260

TABLE V

Bivariate linear Wold process

$$\lambda_1(t)=2.300+10.100\left(t-t_{(1)}^{(1)}\right)+4.500\left(t-t_{(1)}^{(2)}\right),$$

$$\lambda_2(t)=0.000+7.800\left(t-t_{(1)}^{(1)}\right)+6.900\left(t-t_{(1)}^{(2)}\right).$$

Length of the Interval and Numbers of Data	−Log(Likelihood) at the True Parameter	−Log(Maximum of Likelihood)	Maximum Likelihood Estimates		
$T=100$	−424.386	−425.352	2.620	10.168	3.306
$n^{(1)}\doteqdot 500$			−0.064	8.788	6.878
	−434.336	−437.251	2.889	10.262	3.377
$n^{(2)}\doteqdot 300$			0.174	7.485	5.757
	−450.781	−454.897	2.469	10.851	4.813
			−0.134	6.339	8.894
	−424.451	−427.822	2.151	10.792	5.143
			−0.621	8.884	7.815
	−421.424	−422.401	2.231	10.475	5.120
			0.112	6.383	6.936
$T=2500$	−10568.720	−10573.727	2.216	10.618	4.710
$n^{(1)}\doteqdot 12500$			0.032	8.083	6.917
	−10373.444	−10378.002	2.130	10.465	4.875
$n^{(2)}\doteqdot 7500$			−0.083	8.156	7.096
	−10426.732	−10429.884	2.223	10.153	4.736
			−0.042	8.191	7.080
	−10549.615	−10558.059	2.193	10.817	4.738
			0.026	7.866	6.808
	−10517.033	−10523.479	2.084	10.565	5.223
			−0.064	7.993	7.106

TABLE VI

Bivariate Hawkes' mutually exciting process

$$\lambda_1(t)=\mu_1+\int_0^t a_{11}e^{-b_1(t-s)}\,dN_1(s)+\int_0^t a_{12}e^{-b_1(t-s)}\,dN_2(s),$$

$$\lambda_2(t)=\mu_2+\int_0^t a_{21}e^{-b_2(t-s)}\,dN_1(s)+\int_0^t a_{22}e^{-b_2(t-s)}\,dN_2(s),$$

where
$$\begin{pmatrix}\mu_1, a_{11}, a_{12}, b_1\\ \mu_2, a_{21}, a_{22}, b_2\end{pmatrix}=\begin{pmatrix}1.300, 0.500, 1.500, 2.800\\ 2.500, 0.001, 1.400, 2.100\end{pmatrix}.$$

Length of the Interval and Numbers of Data	−Log(Likelihood) at the True Parameter	−Log(Maximum of Likelihood)	Maximum Likelihood Estimates			
$T=100$	−1573.376	−1578.276	0.899	0.530	1.504	2.610
$n^{(1)}\doteqdot 650$			3.138	0.225	1.364	2.588
	−1385.130	−1387.372	1.949	0.500	1.453	2.955
$n^{(2)}\doteqdot 750$			2.702	0.167	1.251	2.238
	−2331.594	−2336.610	1.695	0.394	1.383	2.538
			2.519	0.066	1.169	1.648
	−1159.968	−1162.594	0.661	0.513	1.374	2.344
			2.888	−0.168	1.271	1.965
	−1380.468	−1381.538	1.150	0.580	1.486	2.791
			2.079	0.104	1.138	1.725
$T=2500$	−36844.227	−36847.240	1.312	0.425	1.452	2.651
$n^{(1)}\doteqdot 16000$			2.483	0.029	1.481	2.247
	−35538.868	−35542.371	1.340	0.479	1.529	2.796
$n^{(2)}\doteqdot 19000$			2.539	0.058	1.393	2.209
	−37916.281	−37921.286	1.433	0.504	1.530	2.865
			2.644	−0.011	1.331	2.023
	−38305.112	−38311.857	1.144	0.429	1.438	2.531
			2.628	−0.082	1.441	2.083
	−36965.147	−36969.812	1.286	0.472	1.352	2.546
			2.508	0.004	1.381	2.071

REFERENCES

[1] H. Akaike, "On entropy maximization principle," *Application of Statistics*, P. R. Krishnaiah, Ed. Amsterdam: North-Holland, 1976, pp. 27–41.

[2] H. Akaike, G. Kitagawa, E. Arahata, and F. Tada, "TIMSAC-78," *Computer Science Monographs*, Institute of Statistical Mathematics, Tokyo, 1979.

[3] K. B. Athreya, R. L. Tweedie, and D. Vere-Jones, "Asymptotic behavior of point processes with Markov-dependent intervals," *Math. Nachrichten*, 1981, to be published.

[4] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, pp. 83–95, 1971.

[5] A. G. Hawkes, and D. A. Oakes, "A cluster process representation of a self-exciting process," *J. Appl. Prob.*, vol. 11, pp. 493–503, 1974.

[6] V. Isham, and M. Westcott, "A self-correcting point process," *Stoc. Proc. Appl.*, vol. 8, pp. 335–348, 1979.

[7] J. Jacod, "Multivariate point processes: Predictable projection, Radon–Nikodym derivatives, representation of martingales," *Z. f. Wahr. verw. Geb.*, vol. 31, pp. 235–253, 1975.

[8] Yu. M. Kabanov, R. S. Lipcer, and A. N. Shiryaev, "Criteria of absolute continuity of measures corresponding to multivariate point processes," in *Proc. 3rd U.S.S.R. and Japan Symposium on Probability Theory*. Berlin: Springer, pp. 232–252, 1976.

[9] P. A. W. Lewis, and G. S. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Res. Logistics Quart.*, vol. 26, no. 3, pp. 403–413, 1979.

[10] Y. Ogata, "The asymptotic behavior of maximum likelihood estimators for stationary point processes," *Ann. Inst. Stat. Math.*, vol. 30, no. 2,A, pp. 243–261, 1978.

[11] T. Ozaki, "Maximum likelihood estimation of Hawkes' self-exciting point processes," *Ann. Inst. Stat. Math.*, vol. 30, no. 3,B, pp. 145–155, 1978.

[12] T. Ozaki, and H. Akaike, "On a generalized parametric model for Hawkes' self-exciting point processes," Institute of Statistical Mathematics, Tokyo, Japan, *Res. Memo.*, no. 149, 1979.

[13] Y. Sakamoto and H. Akaike, "Analysis of cross classified data by AIC," *Ann. Inst. Stat. Math.*, vol. 30, no. 1,B, pp. 185–197, 1978.

[14] D. Vere-Jones, "Earthquake prediction—A statistician's view," *J. Phys. Earth*, vol. 26, pp. 129–146, 1978.

Universal Tree Encoding for Speech

YASUO MATSUYAMA, MEMBER, IEEE, AND ROBERT M. GRAY, FELLOW, IEEE

***Abstract*—A low-rate (about one bit per sample) waveform coder for speech compression is designed using techniques from universal source coding, fake process tree encoding, and linear predictive coding (LPC). The system does not require on-line adaptation or LPC analysis, yet it yields a fidelity that compares well with the best existing adaptive-waveform coder of the same rate.**

INTRODUCTION

SPEECH COMPRESSION systems usually fall into one of two general classes—*waveform coders* and *speech coders*. Waveform coders include traditional schemes such as fixed and adaptive scalar quantizers, delta modulators, predictive quantizers, and more recent systems based on information theoretic ideas such as the fixed tree encoding systems of Anderson *et al.*, [1], [2], [3], and the adaptive tree encoding systems of Jayant and Christensen [4], and Wilson and Husain [5], [6]. Speech coders estimate or model the process producing the observed waveform and then send a digitized representation of this model rather than the waveform itself. Examples of such systems are formant vocoders and linear predictive coding (LPC) systems [7], [8].

Existing fixed-rate waveform coders operate at bit rates from about one bit per sample to over ten bits per sample. For the sampling rate of 8 kHz considered here, this yields transmission rates of 8000 to 80 000 (bit/s). Speech coders operate at lower bit rates–typically from one bit per sample down to very low rates such as 1/6 bit per sample [9].

Speech coders provide better quality for low to very low bit rates, but are generally far more complex than the waveform coders. Furthermore, waveform coders are generally more robust against speaker variations, background noise, and channel errors. Hence, waveform coders are of interest when the available channel data rate is adequate, although of course one still wishes to keep the data rate as low as possible. Here we use recent techniques from speech coding and universal source coding to develop a waveform coder that operates at the relatively low rate of about one bit per sample or 8000 bit/s. This is effectively the lowest rate at which there exist intelligible waveform coders with fair subjective fidelity. In particular, the adaptive tree coding system of Wilson and Husain [5], [6]—which uses LPC techniques for on-line adaptation—provides intelligible fair-quality speech at slightly over one bit per sample. Our principal goal is to obtain a system with comparable fidelity by using universal coding techniques instead of on-line adaptation. We use LPC techniques off-line to design a code which requires more

Manuscript received February 14, 1979; revised May 20, 1980. This work was supported by the Japan Society for the Promotion of Science, Murata Overseas Scholarship Foundation, and by the Joint Services Electronics Program at Stanford University, CA. This paper was presented at the 1979 IEEE International Symposium on Information Theory, Grignano, Italy, June 25–29, 1979

Y. Matsuyama is with the Information Science Division, College of General Education, Ibaraki University, Mito, Japan.

R. M. Gray is with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

0018-9448/81/0100-0031$00.75