SIMEST: An Algorithm for Simulation Based Estimation of Parameters Characterizing a Stochastic Process

by

James R. Thompson[1]
E. Neely Atkinson[2]
Barry Brown[2]

Technical Report 86-20, August 1986

[1]Department of Mathematical Sciences, Rice University, Houston, Texas

[2]Department of Biomathematics, The University of Texas System Cancer Center, M.D. Anderson Hospital and Tumor Institute, Houston, Texas

SIMEST: An Algorithm for Simulation Based Estimation of Parameters Characterizing a Stochastic Process

James R. Thompson[1]
E. Neely Atkinson[2]
Barry W. Brown[2]

[1]Department of Mathematical Sciences
Rice University
Houston, Texas
[2]Department of Biomathematics
The University of Texas System Cancer Center
M. D. Anderson Hospital and Tumor Institute

SIMEST: An Algorithm for Simulation Based Estimation of Parameters Characterizing a Stochastic Process

James R. Thompson[1], E. Neely Atkinson[2], and Barry W. Brown[2]
[1]Department of Mathematical Sciences, Rice University and
[2]Department of Biomathematics, The University of Texas System Cancer Center, M. D. Anderson Hospital

I. INTRODUCTION

The axioms defining stochastic processes are generally simple. However, estimation of the parameters of a process from data is extremely difficult if customary techniques are used. This is due to the complexities involved in obtaining closed forms of likelihoods and evaluating them. The authors develop an estimation technique which selects those parameters which produce simulations that best mimic the data. SIMEST makes stochastic process modelling in oncology (and other fields) an attractive alternative to such currently popular alternatives as *ad hoc* regression models.

The first published work dealing with the estimation of a cumulative distribution function was John Graunt's tabulation of the probability of survival until given ages [1]. Thus the first analysis of probabilities based on the real number system used time data. This initial use may be connected with the empirical ordering of time. Unlike the three spatial variables which are unordered in direction, the time dimension is not only ordered but irreversible. It is likely that this psy-

chologically strong ordering property of time drew Graunt to use failure analysis as the world's first example of the representation of continous data.

Graunt's approach was empirical, representational, and without axiomatic underpinings. It was over 150 years later that Poisson [2] began the study of the underlying mechanisms which generated stochastic processes. For the simplest cases, the parameters characterizing these mechanisms lend themselves readily to estimation from data; but the more difficult models that can be used to model, for example, an economy or a tumor system, have proved generally intractable to such estimation. Hence, workers in these areas tend to use empirical models, which are frequently linear, to describe phenomena. This empirical approach has been generally frustrating, frequently unstable, and all too often misleading. The inability to perform satisfactory parameter estimation using the axiomatic approach of Poisson has been taken for granted for so long that stochastic process modelling is little used in fields where it is *the* natural approach.

The position we take here is that it is the emphasis on obtaining closed form solutions of differential, difference, or integral equations resulting from the axiomatic approach that has caused its lack of utilization. The usual axioms of stochastic processes describe changes in the order in which they occur in time. In contrast, the derivation of a likelihood function requires a backwards approach;

for each end state of the process, all paths which might have led to it must be traced backwards in time. We propose as an alternative to closed form solutions the simulation of stochastic processes directly from the assumptions used in their definition. Parameter estimation is accomplished by systematically varying the values of the parameters of the process under investigation until the simulated values are maximally concordant with the data.

Two sets of questions arise from this brief statement of the method. First, how are the simulations to be performed and how is the degree of agreement between simulated values and data to be assessed? Second, what methods for systematically varying parameter values lead to an optimal accord with the data? Most of this investigation is concerned with the numerous ramifications of the first question. The second question, which can be restated as the problem of minimizing a function whose values can only be estimated with error, is under intensive investigation. Some preliminary suggestions are offered.

The primary advantage of the simulation method over closed form solutions is that it can be used even in cases in which closed form solutions are unknown. In addition, the intellectual effort necessary to perform simulation from assumptions is much less than that required to derive exact likelihood equations. This is of great importance for cases in which a series of related models is examined. The similarity

of the models generally assures that the simulation algorithms implementing them are also similar; in contrast, the exact likelihood equation may change greatly with minor changes in the model.

Simulation methods are computationally intensive and require a digital computer. As a practical matter, however, a computer is also required for maximizing exactly obtained non-linear likelihoods.

The methods described here were first presented in Atkinson, Bartoszyński, Brown and Thompson in 1983 [3,4]. A statement of the overall problem very similar to that presented here appears in Diggle and Gratton, 1984 [5]. Their simulation solution involved density estimation at each set of parameter values. A likelihood is then calculated from the estimated density. We feel that the density estimation is unnecessary (and it is computationally expensive); in addition a bad choice of the smoothing parameter in the density estimation can lead to instability in the estimation. There are simpler, cheaper, and more direct alternatives.

II. DISCUSSION

A. Poisson Modeling

Let us consider Poisson's simplest model. We shall be interested here in the number of failures as a function of time. Following Poisson, we shall use the following axioms:

(A1) The probability that one failure takes place in a time interval $[t,t+\Delta t)$ is given by $\theta\Delta t$.

(A2) The probability two or more failures take place in $[t,t+\Delta t)$ is given by $o(\Delta t)$ (where $\lim_{\Delta t \to o} o(\Delta t)/\Delta t = 0$).

(A3) (First order stationarity)

$\Pr[k \text{ failures in } [t,t+s)] = \Pr[k \text{ failures in } [u,u+s)]$

for all k, t, u and s.

(A4) $\Pr[k \text{ failures in } [t,t+s) \text{ and } l \text{ failures in } [u,u+v)] =$

$\Pr[k \text{ failures in } [t,t+s)] \Pr[l \text{ failures in } [u,u+v)]$

when the two intervals have no points in common.

We denote by $x(t)$ the number of failures in $[o,t)$.

Then

$$Pr[x(t+\Delta t)=k] \tag{1}$$

$$= Pr[x(t)=k]Pr[x(\Delta t)=0] + Pr[x(t)=k-1]P[x(\Delta t)=1] + o(\Delta t)$$

$$= Pr[x(t)=k][1-\theta\Delta t] + Pr[x(t)=k-1]\theta\Delta t + o(\Delta t)$$

This yields

$$\frac{Pr[x(t+\Delta t)=k] - Pr[x(t)=k]}{\Delta t} = \theta[Pr[x(t)=k-1] - Pr[x(t)=k]] + o(\Delta t)/\Delta t \tag{2}$$

Letting $\Delta t \rightarrow$ O, we obtain the differential-difference equation.

$$\frac{dP[x(t)=k]}{dt} = \theta[Pr[x(t)=k-1] - Pr[x(t)=k]] \tag{3}$$

By using integrating factors for k=0, 1, 2 and 3 we can guess the solution,

$$Pr[x(t)=k] = \frac{e^{-\theta t}(\theta t)^k}{k!} \tag{4}$$

which can be checked as correct by noting that (4) satisfies (3). We can use (4) to obtain the probability of at least one failure on or before t via the cumulative distribution function

$$F(t \mid \theta) = 1 - Pr[x(t)=0] = 1 - e^{-\theta t} \tag{5}$$

The density function for the first failure is readily obtained by differentiation to

give

$$f(t \mid \theta) = F'(t|\theta) = \theta e^{-\theta t} \tag{6}$$

We note that the expectation (average) of t is given by

$$\mu = E(t) = \int_0^\infty t f(t \mid \theta) dt = \int_0^\infty t\theta e^{-\theta t} dt = \frac{1}{\theta} \tag{7}$$

In many situations we will have n independent observations $\{t_1 \leq t_2 \leq \ldots \leq t_n\}$ from which we wish to estimate the characterizing parameter(s) (in the example given, θ). There are a number of procedures available for this purpose.

B. Method of Moments Estimation

Perhaps the oldest estimation technique, extensively investigated by Pearson [6], but in actuality used hundreds of years earlier, is the "method of moments." To explicate this view, we consider the empirical finite distribution which has probability function

$$\begin{aligned} p(t) &= \frac{1}{n} \text{ if t} = t_j \text{ for } j = \{1, 2, \ldots, n\} \\ &= 0, \text{ otherwise} \end{aligned} \tag{8}$$

The expected value of t for this distribution is given simply by the sample mean

$$\bar{t} = \sum_{j=1}^{n} t_j p(t_j) = \frac{1}{n} \sum_{j=1}^{n} t_j \tag{9}$$

If we make the oversimplifying assumption that the empirical finite distribution represents not only what has occurred but what could occur, then we could use as the estimate for μ in (7), simply $\bar{t}$. This gives immediately

$$\hat{\theta} = \frac{1}{\bar{t}} \tag{10}$$

Although the method of moments is frequently satisfactory when the parameter Θ being estimated is of low dimensionality, there are problems with its usage as the dimensionality increases. This is due, in part, to the fact that $1/n \sum_{j=1}^{n} t_j^m$ becomes less and less satisfactory as an approximation to $\int_0^\infty t^m f(t \mid \Theta) dt$ as n increases.

C. Bayesian Estimation

Another estimation technique in frequent use is that based on Bayes Theorem [7]. Here, we assume that, prior to any observation of failure times, our feelings as to the true value of $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$ can be characterized by a *prior* probability density function $\mathbf{g}(\Theta)$. The joint probability density function of Θ and $(t_1, t_2, \ldots, t_n)$ is then given by

$$g(\Theta) \prod_{j=1}^{n} f(t_j \mid \Theta) = g(\Theta) L(\Theta \mid t_1, \ldots, t_n) \tag{11}$$

The term $L(\Theta \mid t_1, \ldots, t_n)$ is called the *likelihood.*

Subsequent to the recording of the failure times, the *posterior* density of Θ is given by

$$g(\Theta \mid t_1, \ldots, t_n) = \frac{g(\Theta) \prod_{j=1}^{n} f(t_j \mid \Theta)}{\int \ldots \int g(\theta_1, \ldots, \theta_m) \prod_{j=1}^{n} f(t_j \mid \Theta) d\theta_1 \ldots d\theta_m} \tag{12}$$

For reasons of "closure" as well as computational convenience, it is frequently decided to pick $g(\Theta)$ so that $g(\Theta \mid t_1, \ldots, t_n)$ will be of the same functional form as $g(\Theta)$. In such a case, g is called a *natural conjugate* prior.

In the example considered in (6), the natural conjugate prior is the gamma density

$$g(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{\theta\beta} \theta^{\alpha-1}; 0 < \theta \tag{13}$$

The posterior density given the n failure times is given simply by:

$$g(\theta \mid t_1, \ldots, t_n) = \frac{g(\theta) \prod_{j=1}^{n} f(t_j \mid \theta)}{\int_0^{\infty} g(\theta) \prod_{j-1}^{n} f(t_j \mid \theta) d\theta} = \frac{(n\bar{t} + \beta)^{n+\alpha}}{\Gamma(n - \alpha)} e^{-\theta(n\bar{t}+\beta)} \theta^{n+\alpha-1} \tag{14}$$

Using the posterior distribution in (14), we have several candidates for estimating θ. For example, the posterior mean is given by

$$\begin{aligned} E(\theta \mid t_1, \ldots, t_n) &= \int_0^{\infty} \theta g(\theta \mid t_1, \ldots, t_n) d\theta \\ &= \frac{n + \alpha}{n\bar{t} + \beta} \end{aligned} \tag{15}$$

The value of θ which maximizes $g(\theta \mid t_1, \ldots, t_n)$ is the posterior mode.

$$PM(\theta \mid t_1, \ldots, t_n) = \frac{n + \alpha - 1}{n\bar{t} + \beta} \tag{16}$$

We note that as n becomes very large, both these estimators become approximately equal to $1/\bar{t}$.

One problem with Bayesian estimators is the fact that it is frequently difficult to incorporate "prior information" into a prior density $g(\Theta)$. Another difficulty is the fact that if we do not have a ready conjugate prior for Θ, then Bayesian estimation frequently becomes cumbersome — both computationally and perceptually.

D. Maximum Likelihood Estimation

If our information about Θ is very vague we may use (12) to derive an estimator which is based purely on the likelihood. We assume that the prior density $g(\Theta)$ is given by

$$\begin{aligned} g(\theta_1, \ldots, \theta_m) &= \prod_{i=1}^{m} \frac{1}{b_i - a_i}, \text{when } a_i < \theta_i < b_i \text{ for all i} \\ &= 0, \text{otherwise} \end{aligned} \tag{17}$$

We take a_i to be so small and b_i to be so large that we are practically certain that θ_i as contained in the interval (a_i, b_i). Then (12) becomes

$$\begin{aligned} g(\Theta \mid t_1, \ldots, t_n) &= \frac{\prod_{i=1}^{m}(\frac{1}{b_i - a_i}) \prod_{j=1}^{n} f(t_j \mid \Theta)}{\prod_{i=1}^{m}(\frac{1}{b_i - a_i}) \int \cdots \int \prod_{j=1}^{n} f(t_j \mid \Theta) d\Theta} \\ &= C(t_1, \ldots, t_n) \prod_{j=1}^{n} f(t_j \mid \Theta) \\ &= C(t_1, \ldots, t_n) L(\Theta \mid t_1, \ldots, t_n) \end{aligned} \tag{18}$$

Clearly, then, to maximize $g(\Theta \mid t_1, \ldots, t_n)$, (i.e., to obtain the posterior mode of Θ), we find the value of Θ which maximizes the likelihood of $L(\Theta \mid t_1, \ldots, t_n)$. Such an estimatior $\hat{\Theta}$ is called a *maximum likelihood* estimator for Θ. This type of derivation of the method of maximum likelihood is attributed by Fisher [8] to Gauss. For the example given in (6), then, the likelihood is given by:

$$L(\theta \mid t_1, \ldots, t_n) = \theta^n \exp[(-n\bar{t}\theta)] \tag{19}$$

The maximum likelihood estimator for θ is simply equal to $1/\bar{t}$.

A major difficulty with all the estimation procedures considered so far is their dependence on the assumption that the density $f(t \mid \Theta)$ is known. As we shall demonstrate shortly, such an assumption is frequently unjustified. It is interesting to note that in his famous attacks on Bayesian estimation, R. A. Fisher [8] rejected the reasonableness of the Bayesian assumption of knowledge of a prior density $g(\Theta)$. However, he was quite willing to presuppose that $f(t \mid \Theta)$ would be available.

E. Simulation Based Estimation

We return to the case in which $f(t \mid \Theta)$ is not readily available, but there is a means of simulating failure times according to axioms presumed to govern the data. Using the sample $0 < t_1 < t_2 < \ldots < t_n$, we divide the time axis into k bins, the ℓ'th of which contains n_ℓ observations. Assuming a value for Θ, we use the

simulation mechanism SM(Θ) to generate a large number N of simulated failures $0 < s_1 < s_2 < \ldots < s_N$. The number of these observations which fall into the ℓ'th bin will be denoted by $\nu_{k\ell}$. If our selection of Θ was close to the truth, then the simulated bin probabilities

$$\hat{p}_{k\ell}(\Theta) = \frac{\nu_{k\ell}}{N} \tag{20}$$

should approximate the corresponding proportion of data in the same bin,

$$\hat{p}_\ell = \frac{n_\ell}{n} \tag{21}$$

We shall call the asymptotic value of $\hat{p}_{k\ell}(\Theta)$ as N goes to infinity, $p_{k\ell}(\Theta)$.

Criteria are needed for assessing the deviation of the $\hat{p}_{k\ell}(\Theta)$ from the $\hat{p}_\ell$. One criterion is the multinomial log likelihood:

$$S_1^1(\Theta) = \ln n! - \sum_{j=1}^{k} \ln n_j! + \sum_{j=1}^{k} n_j \ln \hat{p}_{kj} \tag{22}$$

The simulated observations are used to estimate the probability, $\hat{p}_{kj}$, that an observation will fall in the j'th bin. The expression shown is then the logarithm of the probability that for each j, the j'th bin will contain n_j observations. The first two terms in the expression for S_1 do not depend in any way on Θ, but only on the binned observations n_j. Consequently, we can drop them and use as the criterion the equivalent expression,

$$S_1(\Theta) = \sum_{j=1}^{k} n_j \ln \hat{p}_{kj} \tag{23}$$

This log of the multinomial likelihood is maximized when

$$\hat{p}_{kj} = \frac{n_j}{n} = \hat{p}_j \tag{24}$$

i.e., when the simulated cell probabilities match those from the original sample.

The determination of the relative sizes of the k bins used to discretize the data has not been specified. There are a number of reasons for using a binning scheme with equal numbers of observations in each bin. For example, this minimizes the chance of empty cells in the simulation. Moreover, setting $p_j = 1/k$ gives the Min Max Var$\{\hat{p}_k\}$. And equal binning (setting $p_j = 1/k$) guarantees that the expectation of $S_1(\Theta)$ does not increase for small perturbations of Θ from truth (i.e., the Gateaux variation is not positive). However, there are circumstances in which equal sized binning is not practical. For example, the values of the dependent variable may be clustered because of inaccuracies in measurement or rounding. In this case, the estimation procedure is enhanced if the division points for discretizing the data are chosen between clusters.

Now, expanding $\ln(\hat{p}_{ki})$ in a Taylor's series about p_{ki}, we have, discarding terms of $O(1/N^2)$, the following formula for the asymptotic variance of S_1:

$$\begin{aligned} Var(S_1(\Theta)) &= \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial S_1}{\partial \hat{p}_{ki}} \mid_{p_{ki}} \frac{\partial S_1}{\partial \hat{p}_{kj}} \mid_{p_{kj}} \mathrm{Cov}(\hat{p}_{ki}, \hat{p}_{kj}) \mid_{p_{ki}, p_{kj}} \\ &= \sum_{i=1}^{k} \frac{(1-p_{ki})}{N p_{ki}} n_i^2 - 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \frac{p_{ki}p_{kj}}{N p_{ki}p_{kj}} n_i n_j \end{aligned} \tag{25}$$

$$= \frac{1}{N}\left[\sum_{i=1}^{k}\frac{(1-p_{ki})}{p_{ki}}n_i^2 - (n^2 - n_1^2 - n_2^2 \quad \ldots - n_k^2)\right]$$

which is minimized for $p_{ki} = n_i/n$ for all i.

Suppose for the remainder of this discussion that the bins are chosen so that $\hat{p}_i = 1/k$ for all i; then if Θ is close to the truth, the simulated bin probabilities should each approximate $1/k$. If for a given k, there is only one value of Θ such that

$$\lim \hat{p}_{k\ell} = \frac{1}{k}, \text{ as } n \to \infty, N \to \infty \tag{26}$$

then we shall say that Θ is *k-identifiable.*

For the simple Poisson case given in (5), the bin boundaries $b_{n0}, b_{n1}, \ldots, b_{nk}$ converge almost surely to

$$b_\ell = \frac{-\ln(1-\ell/k)}{\theta_0} \tag{27}$$

where θ_0 is the true value of θ. Suppose there is only one bin ($b_0 = 0$ and $b_1 = \infty$). Now for any value of θ, all the simulated failures will fall into the bin. Consequently, in this case, θ is not 1-identifiable. For two bins, $b_0 = 0, b_2 = \infty$ and

$$b_1 = \frac{-\ln(\frac{1}{2})}{\theta_0} \tag{28}$$

There is no value of θ other than θ_0 for which

$$\lim_{N\to\infty} \hat{p}_{k1}(\theta) = \frac{1}{2} \tag{29}$$

Consequently for the simple Poisson distribution, θ is 2-identifiable. Moreover, for values of $k \geq 2$, θ is k-identifiable. If SM(Θ) is k-identifiable, then a natural procedure for estimating Θ is to pick the value that maximizes

$$S_2(\Theta) = \sum_{\ell=1}^{k} \ln \hat{p}_{k\ell}(\Theta) \tag{30}$$

Note that maximizing S_2 is equivalent to maximizing S_1 when all of the n_j are equal.

Typically the size of the sample, n, will be relatively small compared to N, the size of the simulation. It is clear that the number of bins, k, is a natural smoothing parameter. For example, for k=1, $\mathrm{Var}(\hat{p}_{11}) = 0$ for all n and N. The variability of $S_2(\Theta)$ can be approximated via the asymptotic formula

$$\begin{aligned} Var(S_2(\Theta)) &= \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial S_2}{\partial \hat{p}_{ki}} \mid_{p_{ki}} \frac{\partial S_2}{\partial \hat{p}_{kj}} \mid_{p_{kj}} Cov(\hat{p}_{ki}, \hat{p}_{kj}) \mid_{p_{ki},p_{kj}} \\ &= \sum_{i=1}^{k} \frac{1}{p_{ki}^2(\Theta)} \frac{p_{ki}(\Theta)(1-p_{ki}(\Theta))}{N} \\ &\quad -2\sum_{i=1}^{k-1}\sum_{j=k+1}^{k} \frac{p_{ki}(\Theta)p_{kj}(\Theta)}{N} \frac{1}{p_{ki}(\Theta)p_{kj}(\Theta)} \\ &\sim \frac{1}{N}\left[\sum_{i=1}^{k} \frac{1-\hat{p}_{ki}}{\hat{p}_{ki}} - k(k-1)\right] \end{aligned} \tag{31}$$

Now,

$$E(S_2(\Theta)) \sim \sum_{i=1}^{k} \ln \hat{p}_{ki} \tag{32}$$

Thus we have a ready measure of a signal to noise ratio via

$$SN_2(k) \quad - \quad | \, E(S_2(\Theta)) \, | \, / \sqrt{Var(S_2(\Theta))} \tag{33}$$

$$\sim \quad \frac{| \, \sqrt{N} \sum_{i=1}^{k} \ln \hat{p}_{ki} \, |}{\sqrt{\sum_{i=1}^{k} (1 - \hat{p}_{ki}) / \hat{p}_{ki} - k(k-1)}}$$

Let us suppose that

$$\frac{Max\{p_{k\ell}\}}{Min\{p_{k\ell}\}} = M \tag{34}$$

Suppose ℓ of the bins have probability η and k-ℓ have probability Mη. Then

$$Var(S_2(\Theta)) = \frac{1}{N} \left[\frac{(M-1)\ell + k}{M\eta} - k^2 \right] \tag{35}$$

This variance is maximized for ℓ = k/2. So for the "worst case,"

$$Var(S_2(\Theta)) = \frac{k^2}{N} \left[\frac{(M+1)^2}{4M} - 1 \right] \tag{36}$$

and

$$E(S_2(\Theta)) = k/2 \ln M - k \ln \left[(M+1)k/2 \right] \tag{37}$$

Thus

$$SN_2(k, M) = \frac{\sqrt{N} \, | \ln[2\sqrt{M}/(k(M+1))] \, |}{\sqrt{(M+1)^2/(4M) - 1}} \tag{38}$$

In Table 1 below, we show values of $SN_2(k, M)/\sqrt{N}$ for various k and M. (38) gives us an indication of instabilities introduced by the simulation process for values of Θ away from truth. For example, for M=100 and 20 bins, a simulation size of

11,562 will be required to achieve a signal to noise ratio of 100. For M=10, a simulation size of 1,014 will achieve SN_2 of 100. At any stage of the optimization algorithm, we can use $\text{Max}\{\hat{p}_{kj}\}/\text{Min}\{\hat{p}_{kj}\}$ as a pessimistic estimate of M in order to achieve conservative estimates of the signal to noise ratio.

In a practical situation we will be confronted with situations where we have two values of Θ - say Θ_1 and Θ_2 - and wish to know whether

$$\lim_{N_1 \to \infty} S_2(\Theta_1) > \lim_{N_2 \to \infty} S_2(\Theta_2) \tag{39}$$

Now, suppose for a particular pair of simulated sample sizes we do have

$$S_2(\Theta_1) > S_2(\Theta_2) \tag{40}$$

How do we know that this difference is significant? From (31), we can obtain $Var(S_2(\Theta_1))$ and $Var(S_2(\Theta_2))$. The variance of the difference is given approximately by

$$Var(S_2(\Theta_1) - S_2(\Theta_2)) = Var(S_2(\Theta_1)) + Var(S_2(\Theta_2)) \tag{41}$$

Thus, if

$$S_2(\Theta_1) - S_2(\Theta_2) > 2\sqrt{Var(S_2(\Theta_1) - S_2(\Theta_2))} \tag{42}$$

we are reasonably confident that the difference is real and not due to simulation noise.

Example. Suppose we have two Θ's - Θ_1 and Θ_2. We have divided the sample of n failures into 10 bins and carried out simulations with sizes $N_1 = 900$ and $N_2 = 2500$. Suppose, moreover, the estimated cell probabilities are as shown in Table 2.

From (31) and (41), we have

$$Var(S_2(\Theta_1) - S_2(\Theta_2)) = \frac{3.5786}{900} + \frac{14.304}{2500} = .009678 \qquad (43)$$

giving

$$2\sqrt{Var(Diff)} = .196955$$

Moreover, from (30), $S_2(\Theta_1) = -23.1966$ and $S_2(\Theta_2) = -26.6579$. Since $S_2(\Theta_1) > S_2(\Theta_2) + 2\sqrt{Var(Diff)}$, we can be reasonably confident that the apparently preferred performance of Θ_1 is not simply due to simulation noise. If, however, the difference between $S_2(\Theta_1)$ and $S_2(\Theta_2)$ is not significant, we may increase the two simulation sizes to increase the signal to noise ratio. We note that if any cell is empty, $S_2(\Theta) = -\infty$ and essentially not informative. Accordingly, we need a procedure to avoid using a mesh structure which is not too fine, particularly at the beginning of the iteration procedure, when we may be far from the optimum. One such procedure is to examine the $\{\hat{p}_{kj}\}$ for a choice of k (say 100). We then

find the largest $\hat{p}_{kj}$ - say M. Then starting at the leftmost bin if $\hat{p}_{kj} < M/100$, then combine bin j with bins to the right until the combined bins have total $\hat{p}_{k,j} + \hat{p}_{k,j+1} + \cdots + \hat{p}_{k,j+\ell} \geq$ M/100. When this has been achieved, we replace each of $\hat{p}_{k,j}, \hat{p}_{k,j+1}, \cdots, \hat{p}_{k,j+\ell}$ with $(\hat{p}_{k,j} + \cdots + \hat{p}_{k,j+\ell})/(\ell + 1)$.

Another convenient criterion function is Pearson's goodness of fit

$$S_3(\Theta) = \sum_{j=1}^{k} \frac{(\hat{p}_{kj} - \hat{p}_j)^2}{\hat{p}_j} \tag{44}$$

Obviously, this function is minimized when $\hat{p}_{kj} = \hat{p}_j$ for all j.

S_3 has an advantage over both S_1 and S_2. Suppose both $\hat{p}_2 = c\hat{p}_1$ and $\hat{p}_{k2} = c\hat{p}_{k2}$. Then S_3 is unchanged when the two cells are combined into a single cell.

Now we observe that the variability of each of the three criterion functions considered is nondecreasing in the number of cells, k (in the completely noninformative case when k=1, each of the three criterion functions has zero variability). On the other hand, increasing k increases our ability to discriminate between the effectiveness of various Θ's to produce simulations which mimic the behavior of the actual sample of failure times.

To demonstrate this fact, let us suppose that we consider the effect of combining the first two cells when S_3 is used. Let us suppose the "miss" of $\hat{p}_{k1} + \hat{p}_{k2}$ from $\hat{p}_1 + \hat{p}_2$ is an amount η. Then for the pooled sample, the contribution to S_3 is

given by

$$\frac{\eta^2}{\hat{p}_1 + \hat{p}_2} \tag{45}$$

Now, for these cells uncombined, let

$$\hat{p}_{k1} = \hat{p}_1 + \frac{\hat{p}_1}{\hat{p}_1 + \hat{p}_2}\eta + \epsilon\eta; \hat{p}_{k2} = \hat{p}_2 + \frac{\hat{p}_2}{\hat{p}_1 + \hat{p}_2}\eta - \epsilon\eta \tag{46}$$

In the uncombined case, the contribution to S_3 is:

$$\frac{(\hat{p}_1/(\hat{p}_1 + \hat{p}_2) + \epsilon)^2\eta^2}{\hat{p}_1} + \frac{(\hat{p}_2/(\hat{p}_1 + \hat{p}_2) - \epsilon)^2\eta^2}{\hat{p}_2} = \frac{1 + \epsilon^2(\hat{p}_1 + \hat{p}_2)^2/(\hat{p}_1\hat{p}_2)}{\hat{p}_1 + \hat{p}_2}\eta^2 \geq \frac{\eta^2}{\hat{p}_1 + \hat{p}_2} \tag{47}$$

We note that only in the case where η is split between the two cells in proportion to $\hat{p}_1$ and $\hat{p}_2$ does a decrease in the number of cells fail to decrease S_3. Hence a decrease in the number of cells decreases our ability to tell us how well a simulation is mimicking the actual data. A similar argument holds for S_1 and S_2.

Naturally, a number of cells greater than the size of the actual sample would be a bad idea. As a practical matter, using the sufficient statistic $(t_1, t_2, \ldots, t_n)$ to give the cell boundaries $(0, t_1), (t_1, t_2), \ldots, (t_n, \infty)$ would generally be extreme. We recall that our strategy is to select a value of Θ which gives a simulation mimicking the sample. But, in a broader sense, we seek to mimic samples which *could have happened.* For n large, $F(t \leq t_j)$ will be very near j/n *except for j nearly 0 or for j close to n.* For these values of j, the t_j are poor estimators for the j/n'tiles of

F(.). Thus for the left-most and right-most bins, we might be well advised to see to it that at least 1% of the sample observations are included in each.

We now address the issue of a practical means of obtaining a 95% confidence set for the true value of Θ. Once the algorithm has converged to a value - say $\hat{\Theta}$ - we then use this value to generate M simulated data sets of size n. We then determine

$$\bar{S}_j = \frac{1}{M}\sum_{i-1}^{M} S_j(\hat{\Theta}, T_i) \text{ and} \tag{48}$$

$$s^2_{S_j} = \frac{1}{M}\sum_{T=1}^{M} (S_j(\hat{\Theta}, T_i) - \bar{S}_j)^2 \tag{49}$$

where T_i denotes the ith simulated data set of size n and T_0 is the actual sample. Then with roughly 95% certainty;

$$S_j(\Theta) = \bar{S}_j \pm \frac{2}{\sqrt{M}} s_{S_j} \tag{50}$$

Next, using $\hat{\Theta}$ as the center of a rotatable design, we fit the quadratic curve

$$S_j(\Theta) = \mathbf{A} + \mathbf{B}\Theta + \mathbf{C}\Theta'\Theta \tag{51}$$

The 95% confidence set for Θ can now be approximated using

$$S_j(\hat{\Theta}, T_0) - \frac{2}{\sqrt{M}} s_{S_j} \leq \mathbf{A} + \mathbf{B}\Theta + \mathbf{C}\Theta'\Theta \leq S_j(\hat{\Theta}, T_0) + \frac{2}{\sqrt{M}} s_{S_j} \tag{52}$$

III. AN APPLICATION

A. Metastatic versus Systemic Secondary Tumor Generation

Let us now consider a stochastic process model for the description of the occurrence and growth of secondary tumors. The model was motivated by work [10] in which there was an indication that the hazard of discovery of secondary tumors after removal of the primary appeared to be nearly constant in time. This led to the postulating of a model [11] in which secondary tumors were sometimes produced by a systemic mechanism with constant intensity in addition to rather than by an accepted metastatic process whose intensity is proportional to primary tumor size. We had data sets in which we had records of the time from the removal of the primary tumor to the first discovery of a secondary tumor. The model was based on four axioms:

H1. For each patient, each tumor originates from a single cell and grows exponentially at rate α.

H2. The probability that a tumor of size $Y_j(t)$, not previously detected and removed to time t is detectable in $[t, t+\Delta)$ is $bY_j(t)\Delta + o(\Delta)$.

H3. Until the removal of the primary, the probability of metastasis in $[t, t+\Delta)$ is $aY_o(t)\Delta$.

H4. The probability of systemic occurrence of a tumor in $[t, t+\Delta)$ equals

$\lambda\Delta + o(\Delta)$. independent of the prior history of the patient.

We note that these axioms are simple - rather like the axioms of the simplest Poisson process given in A1-A4 at the beginning of this study. Clearly, the item of major interest is the estimation of the four parameters: α, b, a and λ. Now if we seek to use maximum likelihood as the estimation technique, we find ourselves confronted by a complex multiterm expression. One of these terms is given below:

$$\begin{aligned} P(T_1 = S', T_2 > S) &= \int\int e^{v(S-S')} p(t;1) p(S';e^{\alpha u})(\lambda + ae^{\alpha(t-u)}) \\ &\times \exp[-\lambda(t-u) - a/\alpha(e^{(t-u)} - 1)] \\ &\times H(v(S-S');S';e^{\alpha u}) H(v(S-S')e^{\alpha S'};u',e^{(t-u)}) du dt \\ &+ \int\int e^{v(S-S')} p(t;1) \exp[-\lambda t - a/\alpha(e^{\alpha t} - 1)] \\ &\times \lambda e^{-\lambda u} p(S'-u;1) H(v(S-S');S'-u;1) du dt \end{aligned} \tag{53}$$

where

$$H(s;t,z) = \exp\{az/\alpha e^{\alpha t}(e^{s} - 1)\log[1 + (e^{-\alpha t} - 1)e^{-s}] \tag{54}$$

$$+ \lambda s/\alpha - \lambda/\alpha \log[1 + e^{\alpha t} - 1)]\} \tag{55}$$

and

$$p(t;z) = bze^{\alpha t} \exp[-bz/\alpha(e^{\alpha t} - 1)] \tag{56}$$

and v(u) is determined from

$$u = \int_0^v (a + b + s - ae^{-s})^{-1} ds \tag{57}$$

The order of computational complexity here is roughly that of four dimensional quadrature. This is near the practical limit of contemporary main-frame computers. The CPU time required (using STEPIT [11]) in the estimation algorithm was approximately 2 hours on the CYBER 173.

The other classical estimation procedures mentioned in this paper also present enormous complexities. We see at a glance the reason that stochastic process modelling has proved so marginal as a practical device in oncology, economics, etc. Typically, the simple axioms associated with these models lead to incredible tangles in the likelihood, moment generating functions, etc. The very problem of getting to the quadrature representation of the likelihood requires enormous human labor. And if another parameter is added in the axioms, we can expect yet another quadrature dimension.

The four hypotheses H1 - H4 lend themselves very well to a simulation of the times of occurrence of secondary tumors. This is hardly surprising, since they were formulated to describe the probabilities that events will or will not occur at specified time.

To give a flow chart of the simulation process, let us first define some relevant

random variables.

$$W_D = \text{time of detection of primary} \tag{58}$$

W_M = time of origin of first metastasis

W_S = time of origin of first systemic tumor

W_R = time of origin of first recurrent tumor

W_d^* – time from W_R to detection of first recurrent tumor

W_D^* = time from W_D to detection of first recurrent tumor

We generate all random variables by first generating u from a uniform distribution on the unit interval [0,1]. Then we set $t = F^{-1}(u)$, where F is the appropriate cumulative distribution function. The tumor volume at time t is given by

$$v(t) = ce^{\alpha t} \tag{59}$$

Here, c is the volume of one cell. It follows from the hypotheses that

$$F_D(t) = 1 - \exp[-\int_0^t bce^{\alpha t}dt] = 1 - \exp[-\frac{bc}{\alpha}e^{\alpha t}] \tag{60}$$

$$F_M(t) = 1 - \exp[-\frac{ac}{\alpha}e^{\alpha t}] \tag{61}$$

$$F_S(t) = 1 - \exp[-\lambda t] \tag{62}$$

$$F_d^*(t) = 1 - \exp[-\frac{bc}{\alpha}e^{\alpha t}] \tag{63}$$

We can now write down our simulation algorithm straight away:

$\underline{SM}$ Input α, λ, a, b (64)

Repeat until $s > 0$

Generate W_D

Generate W_M

If $\mathbf{W}_M > \mathbf{W}_D$, then $W_M \leftarrow \infty$

Generate $\mathbf{W}_S$

$W_R \leftarrow \min(\mathbf{W}_M, \mathbf{W}_S)$

Generate $\mathbf{W}_d^*$

$\mathbf{W}_D^* \leftarrow \mathbf{W}_R + \mathbf{W}_d^* - \mathbf{W}_D$

$s = W_D^*$

If $s < 0$, discard

End repeat

Return s

Using the actual sample $t_1 < t_2 < \cdots < t_n$, we can generate k bins, each with apparent probabilities: $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_k$. Now letting $\Theta = (\alpha, a, b, \lambda)$, we can use SM($\Theta$) N times to generate a simulation of N recurrences $s_1 < s_2 < \cdots < s_N$.

The numbers of simulated detections in each of these bins will be denoted by $\nu_{k1}, \nu_{k2}, \ldots, \nu_{kk}$. The simulated bin probabilities $\hat{p}_{k1}, \hat{p}_{k2}, \cdots, \hat{p}_{kk}$ are then computed via

$$\hat{p}_{kj} = \nu_{kj}/N \tag{65}$$

Now, using $S_1(\Theta), S_2(\Theta), S_3(\Theta)$, or some other reasonable criterion, we are in a position to ascertain how well our guessed value of Θ mimics the behavior of the sample.

The simulation SM embodies a simplifying assumption utilized in the paper by Bartoszyński, Brown and Thompson [10] which employed a " closed form" likelihood approach. This assumption is that the first secondary tumor generated is the first secondary tumor observed. Even with this assumption, the terms in the likelihood are almost too complicated for practical computational purposes (see (53)).

Note that, using simulation, this assumption can be easily be eliminated using the flow chart below:

$\underline{SM2}$ Generate W_D (66)

$j = 0$

$i = 0$

Repeat until $W_M(j) > W_D$

$j = j + 1$

Generate $W_M(j)$

Generate $W^*_{dM}(j)$

$W^*_{dS}(i) \leftarrow W^*_{dS}(i) + W_S(i)$

If $W^*_{dM}(j) < W_D$, then $W^*_{dM}(j) \leftarrow \infty$

Repeat until $W_S(i) > 10W_D$

$i = i + 1$

Generate $W_S(i)$

Generate $W^*_{dS}(i)$

If $W^*_{dS}(i) < W_D$, then $W^*_{dS}(i) \leftarrow \infty$

$s \leftarrow \min[\{W^*_{dM}(j)\}, \{W^*_{dS}(i)\}]$

Return s

End Repeat

(67)

In the above, we generate an array of metastasis detection times $\{W^*_{dM}(j)\}$ and systemic detection times $\{W^*_{dS}(i)\}$ and pick the smallest of these as the first

detection time of a secondary. It should be noted the SIMEST approach enables us to increase the complexity of the underlying model at modest cost to the simulation algorithm. For example, suppose we wished to add to H1 - H4 a fifth axion:

H5 A fraction, γ, of the patients cease to be at systemic risk at the time of removal of the primary tumor if no secondary tumors exist at that time. A fraction, $1-\gamma$, of the patients remain at systemic risk throughout their lives.

The revised set of axioms H1 - H5 can then be simulated via:

SM3 Generate u from $U(0,1)$ (68)

If $u > \gamma$, then proceed as in SM2

If $u < \gamma$, then proceed as in SM2 except replace the step

"Repeat until $W_S(i) > 10W_D$"

with the step

"Repeat until $W_S(i) > W_D$"

An endless array of other modifications to the axiomatic system can be made at little cost to simulation complexity. Most of these modifications would cause an investigator essentially to start from the beginning if he wished to come up with a new "closed form" likelihood function. This is a simple manifestation of the fact that simulation follows the forward and modular nature of the axioms, whereas

the determination of the likelihood does not.

Moreover, SIMEST allows for easy collection of additional useful information. For example, by simple bookkeeping, we can observe what fraction of the discovered secondary tumors are metastatic in origin. We can also record the sizes of simulated tumors at their times of detection.

B. Optimization Methods

The problem of minimizing functions whose value can only be observed subject to noise is under intense investigation at M. D. Anderson Hospital. In the present application, parameter estimation in stochastic processes, the noise is simulation induced. Consequently, we have the option of noise reduction by increasing the simulation size. When we are using equal size bins from the original sample, the signal to noise ratio is high near the true value of Θ. When we are so far from Θ that the simulation bin probabilities are vastly different, simulation noise is a serious problem.

In this study, we have used robust, although slow, direct search optimization methods in our investigation of the methods presented. The two algorithms which have been most used are the simplex method of Nelder and Mead [12] and the more sophisticated method, STEPIT [11] of John Chandler. We are examining

modifications of these methods to handle noise, but have not progressed to the point of reporting these modifications.

More recently developed algorithms such as the optimal locally constrained methods of Gay [13] are, in general, considerably faster than direct search methods. However, these algorithms rely upon derivative information calculated either explicitly from formal differentiation or implicitly by function evaluation. The effect of noise on derivative information can be catastrophic unless regression is used instead of finite differences. Such modifications are being pursued, but are in an early stage of development. Because methods relying on derivatives are not expected to work at all well without extensive modification, they were not further pursued here.

In the examples presented in the next section, STEPIT was used, unmodified for the presence of noise. Noise can cause this method (and any other method not designed to handle noise) to veer off in incorrect directions and to converge prematurely. However, STEPIT's behavior in the presence of noise appears reasonably robust.

A brief outline of the algorithm employed by STEPIT (taken from its documentation) is as follows. At each base point, STEPIT varies each parameter value individually up and down. If either variation yields an improvement, the step

size is doubled and another step is taken. The number of such steps allowed is limited. When a local minimum has been bracketed by this process, quadratic interpolation is used to attempt to refine the position of the minimum. Should this process yield a better value, the base point is moved to this value. If varying the individual parameter values leads to no improvement in the objective function, the step size is decreased. The algorithm terminates when the step size becomes smaller than a user specified minimum.

If the examination of one parameter at a time yields a change in at least two different values, the resultant direction of the changes is calculated and steps in this direction are attempted. Again, success causes the step size to be doubled and quadratic interpolation is attempted when a minimum has been bracketed. The steps taken in this fashion sometimes oscillate and STEPIT attempts to detect and shortcut these zigzags.

1. Numeric Examples

Since the intent of this work is to explicate a class of methods rather than to demonstrate results from cancer data, simulated data from the metastatic-systemic model is used. A real cancer data base is analyzed in [3]. The parameter values chosen for this simulation are taken from a data set which was fit using the cumbersome exact likelihood equations, namely $\alpha = 0.31$, $b = 0.23e - 8$, $\beta =$

$0.17e-9$, and $\lambda = 0.003$. These values are appropriate for time in units of months and tumor volumes in units of number of cells.

With these values, about 91% of the secondary tumors are due to the systemic process and only the remaining 9% are due to the volume dependent metastatic process. This poses a somewhat difficult problem for fitting because the first three model parameters produce only a minor effect on the vast majority of the data.

In the fit, one-half was added to all bins after each simulation with a specified set of parameters in order to avoid attempts to take a logarithm of zero in calculating the criterion function value. Initial step sizes were set at half of the initial value of the parameters; convergence is achieved when these step sizes decrease to 1% of the initial parameter values. Starting values were produced by varying a subset of the parameters upward or downward by a factor of three from the values used to simulate the data. In the explorations conducted, it required 50 to 100 function evaluations to obtain convergence.

Censoring was not considered when simulating data sets to be fit, i.e., there is an unlimited follow up time and all cases eventually fail. Of course, with actual clinical data, censoring is very much a problem and there are two ways that it can be handled. One, suitable for the quasi-likelihood criterion, S_1, is to add terms to the criterion representing the probablility that the time is at least that observed.

This corresponds precisely to the modification of the usual likelihood for censoring. The second method, suitable for all of the criteria discussed, is to spread the failure time of each censored observation equally among the failure times that are larger than the censoring time. Should a censoring time exceed all failure times, it can be given an arbitrarily large value; since the last bin includes all large values the particular choice of this value is not important. It is worth noting that whether or not the data contains censored observations, the simulations performed in fitting the data should not, as such censoring would amount to throwing away information available from the axioms and choice of parameter values.

Five hundred cases were generated with the parameters stated; these were used as the data in exploring the fitting methods. The data were placed into twenty bins, the boundaries of which were chosen so that each contained 25 data points. Five thousand cases were simulated at each parameter value in order to evaluate the criterion used, which was the quasi-likelihood, S_1. Replication of the criterion evaluation at the values used to generate the data showed a standard deviation of the criterion to be 1.1.

The behavior of the fitting algorithm is quite evidently strongly determined by the starting values used for the fit. For example, initial estimates of (0.103, $0.8e-9$, $0.57e-10$, 0.003) converged to parameter values of (0.309, $0.22e-8$, $0.81e-$

10, 0.003). (The vector notation used lists parameter values in the order α, b, β, and λ.) These values are quite near those used to generate the data, the quasi-likelihood at these values was less than one greater than the mean value obtained at the values used to generate the data, and a chi-square goodness of fit test yields an acceptable value. Hence, the fit is quite good. However, starting from $(0.93, 0.69e-8, 0.51e-9, 0.003)$, the fitting algorithm converged to $(2.84, 0.42e-7, 0.26e-9, 0.0029)$. The quasi-likelihood was 26 units less than that obtained from the values used to generate the data, and the goodness of fit statistic did not show an acceptable fit. (Note that this criterion can be used with actual data for which the true parameter values are not known.)

This perplexing behavior caused an attempt to reduce the dimensionality of the problem in order to obtain some understanding of the behavior of the fitting algorithm.

The parameter, λ, is easily estimated from the data. It is the slope of the logarithm of the survival curve for large times. Experimentation with the stated data set and others indicated that fitting from a variety of starting points produced good estimates of this parameter even when the others were bad. Having this estimate reduces the dimension of the fitting problem from four parameters to three.

Additional information is available, both from clinical observation and the simulation; this is the size of the primary tumor at detection, and it can be used to connect parameters α and b by the following argument:

By hypothesis H1, the volume of the primary tumor at time t after its origination, $Y_0(t)$, is

$$Y_0(t) = \exp(c\alpha t) \tag{69}$$

where c is the volume of a single cell (which is 1 in the units being used). By hypothesis **H2**, the detection of the primary tumor is an inhomogeneous Poisson process with intensity $bY_0(t)$. Thus the probability of no detection of the primary tumor to time t is given by

$$\exp(-\int_0^t bY_0(u)\,du) = \exp(-\int_0^t bc\exp(\alpha u)\,du) = \exp(-\frac{b}{\alpha}[Y_0(t) - Y_0(0)] \tag{70}$$

$Y_0(0)$, which equals c, the volume of a single cell, is negligible compared to $Y_0(t)$, the volume of the tumor at detection. Dropping this term, using the fact that in the units used, c is 1.0, and taking the negative derivative with respect to Y_0 shows that the density of the volume of the primary tumor at detection at any size y to be:

$$\frac{b}{\alpha}\exp(-\frac{b}{\alpha}y) \tag{71}$$

This is recognized as an exponential distribution with mean α/b. Consequently, the

mean observed volume of the primary tumor at detection is a maximum likelihood estimate of α/b. For the simulated data, this mean is 1.36e8 cells. (When using clinical data, we assume that there are 1.0e9 cells per cc.) For comparison, the ratio obtained from the values used to simulate the data is 1.348e8.

The reduction in dimension for starting estimates of the parameter does not solve the problem of the bad fits that were obtained above, as the starting values for the fits met the constraints of the reduction. Enforcement of these constraints is also not a solution as the behavior of the fitting method is much the same with and without such constraints. The problem appears to be fundamentally difficult.

The separate estimation of λ and α/b allows contours of the quasi-likelihood surface to be produced. Figures 1 and 2 show these contours as a function of α and β (β is multiplied by 1.0e9 in the figures). Figure 1 uses the same seed for the random number generator for each parameter value. Figure 2 allows this seed to vary from value to value, and thus shows the fluctuation of quasi-likelihood value due to the simulation process.

It is evident from these figures that the likelihood is well behaved in a reasonable sized neighborhood of the true values, particularly for small values of α. In this region, it steeply climbs to its optimal value. Outside of this region, however, the behavior is erratic and the criterion changes very slowly. This maps well to

the findings from the fits.

In an attempt to discover whether this behavior is due to random behavior in generating the data set, another data set of 20,000 observations was generated using the same parameters. This data was categorized into 99 bins. The behavior on this data set was close to that of the original. Using the same starting values that led to bad estimates in the original, the estimates obtained are (7.91, $0.59e-7$, $0.51e-9$, 0.003) and the fit was not acceptable.

2. Two Dimensional Binning

Another attempt to improve the situation was to use the volume of the primary tumor at detection as well as the time to the secondary tumor in estimating the parameter values from the 500 simulated observations; handling a two dimensional outcome is a straightforward extension of the methods presented. The observed volumes were divided into four equal sized categories and the times to secondary tumors into ten categories; the data was binned into the forty divisions formed by the cartesian product of these categories and the binned data was then fit by simulation using the quasi-likelihood criterion. This resulted in a modest improvement in performance over the time data alone, in that the estimates obtained were not as far from those used to generate the data as when time alone was used. These estimates were (3.75, $0.28e-7$, $0.54e-9$, 0.003); however these estimates were

not acceptable.

This finding would tend to indicate that most of the information in the volume data has been extracted in estimating α/b, so the use of this data does not add appreciably to time alone. However, in many cases it will not be possible to simply extract the relevent information from auxillary variables and using such variables as additional outcome information will greatly enhance the ability to estimate the process parameters.

IV. CONCLUSIONS

High speed digital computing has made little impact on the modeling process. Scientists, for the most part, continue to follow the same steps taken by their pre-computer age predecessors, generally transforming axioms into a differential-integral equation representation. From this representation, numerical techniques are used to give pointwise evaluations of a function. These, in turn, can be used to estimate the parameters which characterize the system. If these steps become too complex (and generally they do), the investigator can throw up his hands and use an *ad hoc* regression model.

In this chapter, we have shown how simulation can be used to proceed rapidly from the axioms to the estimation of the characterizing parameters. This concept is, on the one hand, a drastic departure from pre-computer age methodology. On the other hand, it simply extends the power of Karl Pearson's concept of goodness of fit. SIMEST enables the estimation of parameters in models far more complex than those tractable in Pearson's day.

We believe that we are fast approaching the time when the potential of high speed computing to change, fundamentally, the modeling process will be realized. SIMEST is, hopefully, a step in this direction.

Acknowledgements

This work was supported in part by the United States Army Research Office (Durham) and the National Cancer Institute, under DAAG 29 85K 0212 and CA11430, respectively. The authors wish to thank Ms. Betty Schwarz who created the original of this paper using the word processing language Tex.

Table 1

Values of $S\,N_2(k,M)/\sqrt{N}$

k	M			
	5	10	50	100
5	2.13	1.52	.83	.65
10	2.90	2.00	1.03	.79
20	3.68	3.14	1.23	.93
100	5.48	3.63	1.70	1.26

Table 2

Estimated Bin Probabilities

	Θ_1	Θ_2
$\hat{p}_{10,1}$	.12	.09
$\hat{p}_{10,2}$	.08	.11
$\hat{p}_{10,3}$	.10	.05
$\hat{p}_{10,4}$	.12	.08
$\hat{p}_{10,5}$	.08	.09
$\hat{p}_{10,6}$	.09	.05
$\hat{p}_{10,7}$	.11	.15
$\hat{p}_{10,8}$	.07	.11
$\hat{p}_{10,9}$	.12	.12
$\hat{p}_{10,10}$	.11	.15

References

1. John Graunt, *Natural and Political Observations on the Bills of Mortality* (1662).

2. Simeon Denis Poisson, *Recherches sur la Probabilité des Jugements* (1837).

3. E. Neely Atkinson, Robert Bartoszyński, Barry W. Brown, and James R. Thompson, Simulation techniques for parameter estimation in tumor related stochastic processes, *Proceedings of the 1983 Computer Simulation Conference*, North Holland, New York, pp. 754-757 (1983).

4. E. Neely Atkinson, Robert Bartoszyński, Barry W. Brown, and James R. Thompson, Maximum likelihood techniques, *Proceedings of the 44th Meeting of the International Statistical Institute*, Contributed Papers, **2**, pp. 494-497 (1983).

5. P. J. Diggle and R. J. Gratton, Monte Carlo methods of inference for implicit statistical models. *J. R. Statist. Soc.* B, **46**, 193-227 (1984).

6. Karl Pearson, Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material, *Philosophical Transactions of the Royal Society of London, Series H*, **186**, pp. 343-414 (1895).

7. Thomas Bayes (1764), *Facsimiles of Two Papers by Bayes*, Hafner, New York (1963).

8. Ronald A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, p. 21 (1958).

9. Robert Bartoszyński, Barry W. Brown, Charles McBride and James R. Thompson, Some nonparametric techinques for estimating the intensity function of a cancer related nonstationary Poisson process, *The Annals of Statistics*, **9**, pp. 1050-1060 (1981).

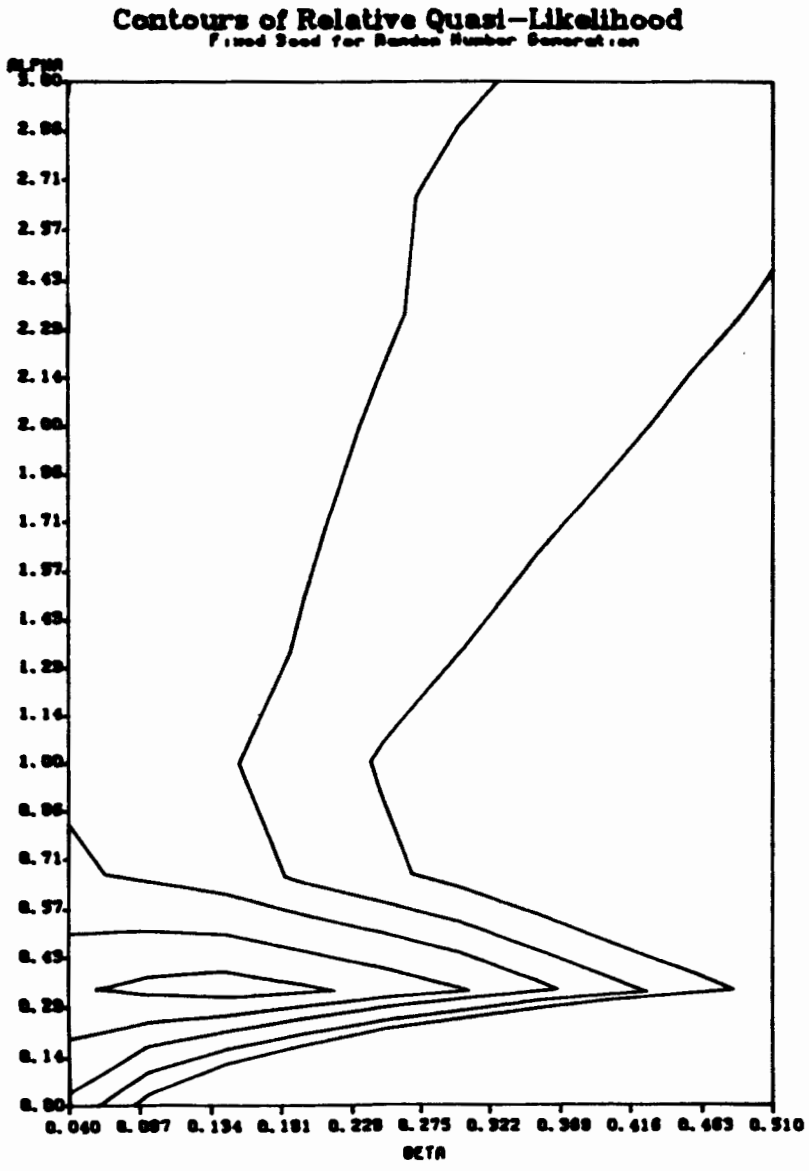
10. Robert Bartoszyński, Barry W. Brown, and James R. Thompson, Metastatic and systemic factors in neoplastic progression. *Probability Models and Cancer*, LeCam and Neyman, eds., North Holland, New York, pp. 253-264 (1982).

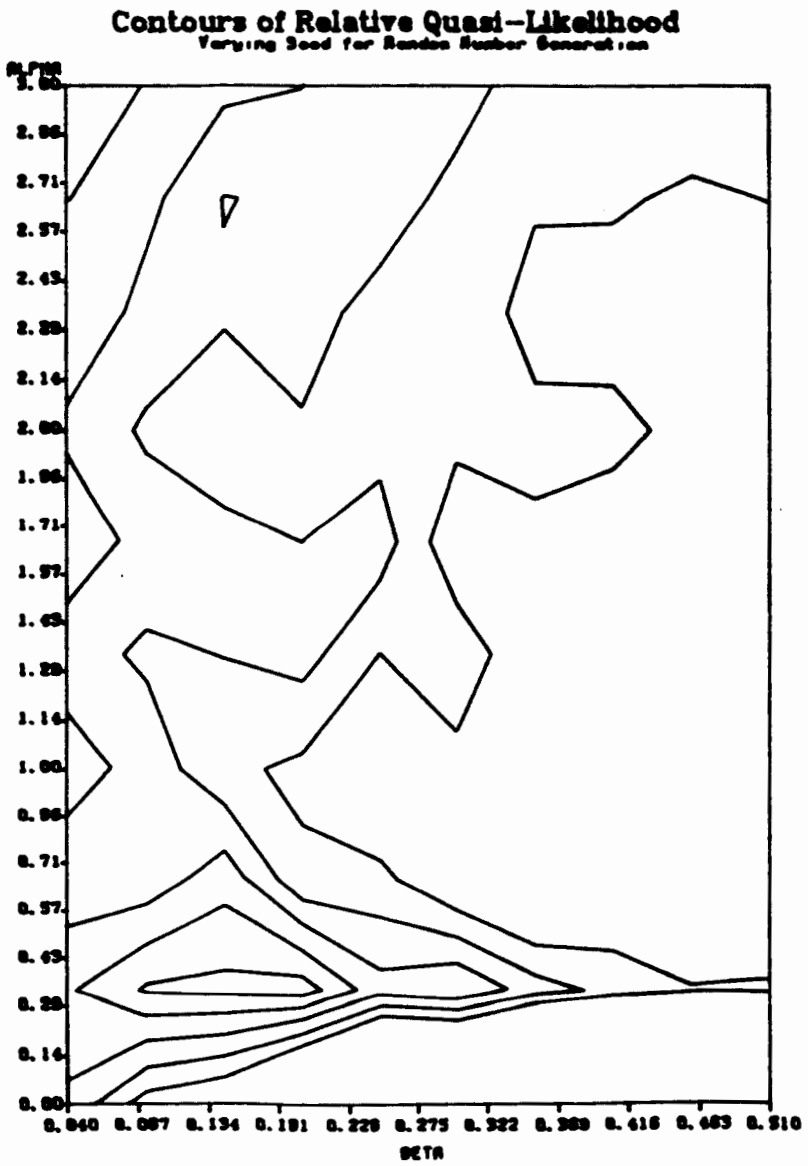
11. J. P. Chandler, STEPIT. *Behavioral Science*, **14**, p. 81 (1969).

12. J. A. Nelder and R. Mead, A Simplex Method for Function Minimization, *The Computer Journal*, **7**, 308-313 (1965).

13. D. M. Gay, Algorithm 611 - subroutines for unconstrained minimization using a model/trust-region approach. *ACM Transactions on Mathematical*

Software, **9**, 160-169 (1983).


Contours of Relative Quasi-Likelihood
Fixed Seed for Random Number Generation
ALPHA
3.00
2.86
2.71
2.57
2.43
2.29
2.14
2.00
1.86
1.71
1.57
1.43
1.29
1.14
1.00
0.86
0.71
0.57
0.43
0.29
0.14
0.00
0.040 0.087 0.134 0.181 0.228 0.275 0.322 0.369 0.416 0.463 0.510
BETA


Contours of Relative Quasi-Likelihood
Varying Seed for Random Number Generation
ALPHA
2.86
2.71
2.57
2.43
2.29
2.14
2.00
1.86
1.71
1.57
1.43
1.29
1.14
1.00
0.86
0.71
0.57
0.43
0.29
0.14
0.00
0.040 0.087 0.134 0.181 0.228 0.275 0.322 0.369 0.416 0.463 0.510
BETA