Available online at www.sciencedirect.com
SCIENCE DIRECT®

C. R. Biologies 327 (2004) 181–192

COMPTES RENDUS
BIOLOGIES

Biological modelling / Biomodélisation

Simulation-based estimation of stochastic process parameters in tumor growth

James R. Thompson

Department of Statistics, Rice University, 6100 South Main Street, Houston, TX 77001-1892, USA

Received 15 April 2003; accepted 24 June 2003

Presented by Pierre Auger

Abstract

Models are generally developed at the micro level. Data are generally gathered at the macro level. Obtaining the macromodel which is the natural consequence of the underlying micro model is generally not feasible. SIMEST gives a means whereby the micromodel is used to generate, for a given assumed set of parameters, simulated sets of macro data. These data are compared with the actual clinical macro data. The parameters are then adjusted to obtain concordance with the clinical data. In this manner, simulation gives us a means of parameter estimation without the necessity of generating the macro model. ***To cite this article: J.R. Thompson, C. R. Biologies 327 (2004).***

Keywords: SIMEST; simulation; tumor growth; pseudo-data

1. The SIMEST paradigm

Following the argument developed in [1], we shall be creating pseudo-datasets for given assumed parameters of a micromodel for a set of input model parameters. As we note from Fig. 1, these pseudo-data will be compared with actual clinical data and the assumed parameters adjusted to bring the simulated pseudo-data in concordance with the clinical data.

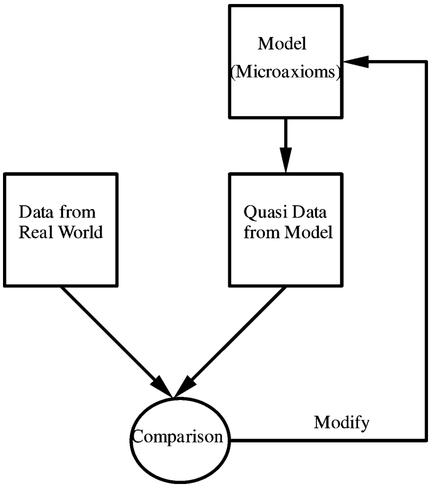

Fig. 1. The SIMEST paradigm.

2. Poisson process modeling

In 1837, well before there were the plethora of technological processes which suit his modeling strategy,

E-mail address: thomp@rice.edu (J.R. Thompson).

1631-0691/$ – see front matter © 2004 Published by Elsevier SAS on behalf of Académie des sciences.
doi:10.1016/j.crvi.2003.06.001

Poisson [2] proposed the following model to deal with $P_k(t) = Prob[k \text{ events in } [0,t)]$. Everything flows from the following four axioms:

(1) $Pr[1 \text{ event in } [t, t+h)] = \lambda h + \mathrm{o}(h)$
(2) $Pr[2 \text{ or more events in } [t, t+h)] = \mathrm{o}(h)$
(3) $Pr[j \text{ events in } [t_1, s_1) \text{ and } k \text{ in } [t_2, s_2)] = Pr[j \text{ in } [t_1, s_1)]Pr[k \text{ in } [t_2, s_2)]$ if $[t_1, s_1) \cap [t_2, s_2)] = \phi$
(4) λ is constant over time.

Then

$$\begin{aligned} P_k(t+h) &= P_k(t)P_0(h) + P_{k-1}(t)P_1(h) \\ &= P_k(t)\big[1 - \lambda h + \mathrm{o}(h)\big] \\ &\quad + P_{k-1}(t)\big[\lambda h + \mathrm{o}(h)\big] \end{aligned}$$

so

$$P_k(t+h) - P_k(t) = \lambda h\big[P_{k-1}(t) - P_k(t)\big] + \mathrm{o}(h) \quad (1)$$

Dividing by h and letting $h \to \infty$, we have

$$\frac{\mathrm{d}P_k(t)}{\mathrm{d}t} = \lambda\big[P_{k-1}(t) - P_k(t)\big] \quad (2)$$

Simple substitution in (2) verifies Poisson's solution:

$$P_k(t) = \frac{\mathrm{e}^{-\lambda t}(\lambda t)^k}{k!} \quad (3)$$

The mean and variance of k are easily shown both to be equal to λt.

Let us consider an early application of Poisson's model. The German statistician von Bortkiewicz examined the number of suicides of women in eight German States in 14 years. His results are shown in Table 1 [3]. Now, there it is an interesting question as to whether it can plausibly be claimed that the suicide data follows Poisson's model. If we compute the sample mean of the number of suicides per year, we find that it is 3.473. We can then use this value as an estimate for λt. In Table 1, we also show the expected numbers of suicides using the Poisson model.

One of the oldest statistical tests is Karl Pearson's *goodness of fit*. When data is naturally categorized, as it is here, in k bins (the number of suicides per state per year), if the number observed in a bin is X_i and the expected number, according to a model, is E_i, then

$$\sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i} \approx \chi^2(k-1) \quad (4)$$

For the von Bortkiewicz data, we compute a value of χ^2 of 54.9. This is well beyond the limit of $\chi^2_{0.990}(10)$ value of 23.21, so we might reject the applicability of the Poisson model. On the other hand, the Pearson approximation is asymptotic. We require a minimum number for each E_i of 5. In the present example, that would mean that we would have to pool the first two bins and the last four. That would give the revised Table 2.

Table 2
Actual and expected numbers of suicides per year

Suicides	$\leqslant 1$	2	3	4	5	6	$\geqslant 7$	Sum
Freq.	28	17	20	15	11	8	13	112
E (Freq.)	15.6	21	24.3	21	14.6	8.5	7	112

This gives us a χ^2 value of 19.15, which is above the $\chi^2_{0.990}(6)$ value of 16.81, but below the $\chi^2_{0.998}(6)$ value of 20.79. Depending upon the use we intend to make of the Poisson model, we might choose to accept it. Yet, the relatively small sample involved might make us wish to try other approaches. For example, we know we have totals of suicides per year given in Table 1. We might decide to employ the following strategy.

Algorithm. Resampled data compared with model-generated data.

1. Create an 'urn' with nine **0** balls, nineteen **1** balls, seventeen **2** balls, and so on.
2. With replacement, sample from the urn 1000 samples of size 112, noting the results.
3. For each of the 1000 samples, compute the χ^2 statistic in (4) using the original values in Table 1 for the E_i.

Table 1
Actual and expected numbers of suicides per year

Suicides	0	1	2	3	4	5	6	7	8	9	$\geqslant 10$	Sum
Freq.	9	19	17	20	15	11	8	2	3	5	3	112
E (Freq.)	3.5	12.1	21	24.3	21	14.6	8.5	4.2	1.9	0.7	0.2	112

4. Using the estimate for λt of 3.473, divide the line segment from zero to 1 according to the Poisson model. Thus the probability of finding a state with zero suicides in a year is $\exp(-\lambda t) = 0.031$. The **0** Poisson bin then is $[0, 0.031)$. The probability of finding a state with one suicide in a year is $\exp(-\lambda t)\lambda t/1! = 0.108$. So the **1** Poisson bin is $[0.031, 0.031 + 0.108)$, and so on.
5. Repeat 1000 times 112 draws of a uniform $[0, 1]$ random variable.
6. Using the E_i values from the third row in Table 1, compute the χ^2 statistic in (4).
7. Compute histograms for both the resampling simulation and that of the Poisson model. If the overlap is, say 5%, accept the hypothesis that the Poisson model fits the data.

Uses of the Poissonian framework are seen, very frequently, in the simulation of a train of time-indexed events. Now,

$$1 - F(t) = P\big[0 \text{ events in } [0, t)\big] = \exp[-\lambda t] \qquad (5)$$

But $F(t)$, the probability that an event occurs on or before t, is a continuous cumulative distribution function and is distributed as a uniform variate on $[0, 1]$. So, also,is $1 - F(t)$. Thus, it is an easy matter, starting at time zero, to simulate the time of the next event. We generate u from $U[0, 1]$. Then the time of the next simulated event is given by

$$t = -\frac{1}{\lambda}\log(u) \qquad (6)$$

Thus, it is possible to create a series of n simulated events by simply generating $u_1, u_2, \ldots, u_n$ and then using

$$t_i = t_{i-1} - \frac{1}{\lambda}\log(u_i) \qquad (7)$$

Next, let us consider what might be done in if we relax the axiom that states that λ must be constant. We can easily do this for the special case where we are considering $P_0(t)$, that is, the probability that no events happen in the time interval $[0, t)$

$$P_0(t + h) - P_0(t) = \lambda h\big[-P_0(t)\big] + o(h) \qquad (8)$$

Dividing by h and taking the limit as h goes to zero, we have

$$\frac{1}{P_0(t)}\frac{dP_0(t)}{dt} = -\lambda(t) \qquad (9)$$

Integrating from 0 to t, we have

$$P_0(t) = \exp\left[-\int_0^t \lambda(\tau)\,d\tau\right] \qquad (10)$$

We are now able to carry out simulations in rather complicated situations. Let us suppose for example, that a tumor, starting with one cell, grows exponentially according to:

$$v(t) = c\,e^{\alpha t}, \quad \text{where } c \text{ is the volume of one cell} \quad (11)$$

Next, let us suppose that this tumor will throw off metastases at a rate a proportional to the volume of the tumor. So, then the probability a metastasis will be produced on or before time t is given by

$$F_M(t) = 1 - \exp\left[-\frac{ac}{\alpha}\,e^{\alpha t_M}\right] \qquad (12)$$

From (12) we can easily write a simulation for the origination times of metastases starting from a tumor with given values of c, α, and a.

3. SIMEST: an oncological example

The power of the computer as an aid to modeling does not get the attention it deserves. Part of the reason is that the human modeling approach tends to be analog rather than digital. Analog computers were replaced by digital computers 40 years ago. Most statisticians remain fascinated by the graphical capabilities of the digital computer. The exploratory data analysis route tends to attempt to replace modeling by visual displays which are then interpreted, in a more-or-less instinctive fashion, by an observer. Statisticians who proceed in this way are functioning somewhat like prototypical cyborgs. After over two decades of seeing data spun, colored, and graphed in a myriad of ways, I have to admit to being disappointed when comparing the promise of EDA with its reality. Its influence amongst academic statisticians has been enormous. Visualization is clearly one of the major areas in the statistical literature. But the inferences drawn from these visualizations in the real world are, relatively speaking, not so numerous. Moreover, when visualization-based inferences are drawn, they tend to give results one might have obtained by classical techniques.

Of course, as in the case of using the computer as a nonparametric smoother, some uses are better than others. It is extremely unfortunate that some are so multicultural in their outlook that they rearrange their research agenda in order to accommodate themselves to our analog-challenged friends, the digital computers. Perhaps the greatest disappointment is to see the modeling aspect of our analog friends, the human beings, being disregarded in favor of using them as gestaltic image processors. This really will not do. We need to rearrange the agenda so that the human beings can gain the maximal assistance from the computers in making inferences from data. That is the purpose of SIMEST.

There is an old adage to the effect that quantitative change carried far enough may produce qualitative change. The fact is that we now have computers so fast and cheap that we can proceed (almost) as though computation were free and instantaneous (with infinite accessible memory thrown in as well). This should change, fundamentally, the way we approach data analysis in the light of models.

There are now a number of examples in several fields where SIMEST has been used to obtain estimates of the parameters characterizing a market-related applied stochastic process. Below we consider an oncological application to motivate and to explicate SIMEST. We shall first show a traditional model-based data analysis, note the serious (generally insurmountable) difficulties involved, and then give a simulation-based, highly computer-intensive way to get what we require to understand the process and act upon that understanding.

3.1. An exploratory prelude

In the late 1970s, my colleague Barry W. Brown, of the University of Texas M.D. Anderson Cancer Center, and I had started to investigate some conjectures concerning reasons for the relatively poor performance of oncology in the American 'War on Cancer'. Huge amounts of resources had been spent with less encouraging results than one might have hoped. It was my view that part of the reason might be that the basic orthodoxy for cancer progression was, somehow, flawed.

This basic orthodoxy can be summarized briefly as follows.

> At some time, for some reason, a single cell goes wild. It, and its progeny, multiply at rates greater than that required for replacement. The tumor thus formed grows more or less exponentially. From time to time, a cell may break off (metastasize) from the tumor and start up a new tumor at some distance from the primary (original) tumor. The objective of treatment is to find and excise the primary before it has had a chance to form metastases. If this is done, then the surgeon (or radiologist) will have "gotten it all" and the patient is cured. If metastases are formed before the primary is removed, then a cure is unlikely, but the life of the patient may be extended and ameliorated by aggressive administration of chemotherapeutic agents which will kill tumor cells more vigorously than normal cells. Unfortunately, since the agents do attack normal cells as well, a cure of metastasized cancer is unlikely, since the patient's body cannot sustain the dosage required to kill all the cancer cells.

For some cancers, breast cancer, for example, long-term cure rates had not improved very much for many years.

3.2. Model and algorithms

One conjecture, consistent with a roughly constant intensity of display of secondary tumors, is that a patient with a tumor of a particular type is not displaying breakaway colonies only, but also new primary tumors due to suppression of a patient's immune system to attack tumors of a particular type. We can formulate axioms at the micro level which will incorporate the mechanism of new primaries.

Such an axiomitization has been formulated by Bartoszyński et al. [4]. The first five axioms are consistent with the classical view as to metastatic progression. Hypothesis 6 is the mechanism we introduce to explain the nonincreasing intensity function of secondary tumor display.

Hypothesis 1. For any patient, each tumor originates from a single cell and grows at exponential rate α.

Hypothesis 2. The probability that the primary tumor will be detected and removed in $[t, t + \Delta t)$ is given by $bY_0(t)\Delta t + o(\Delta t)$, and until the removal of the primary, the probability of a metastasis in $[t, t + \Delta t)$ is $aY_0(t)\Delta t + \mathrm{o}(\Delta t)$, where $Y_0(t)$ is the size of the primary tumor at time t.

Hypothesis 3. For patients with no discovery of secondary tumors in the time of observation, S, put $m_1(t) = Y_1(t) + Y_2(t) + \cdots$, where $Y_i(t)$ is the size of the ith originating tumor. After removal of the primary, the probability of a metastasis in $[t, t + \Delta t)$ equals $am_1(t) + \mathrm{o}(\Delta t)$, and the probability of detection of a new tumor in $[t, t + \Delta t)$, is $bm_1(t) + \mathrm{o}(\Delta t)$.

Hypothesis 4. For patients who do display a secondary tumor, after removal of the primary and before removal of Y_1, the probability of detection of a tumor in $[t, t + \Delta t)$ equals $bY_1(t) + \mathrm{o}(\Delta t)$, while the probability of detection of a metastasis is $aY_1(t) + \mathrm{o}(\Delta t)$.

Hypothesis 5. For patients who do display a secondary tumor, the probability of a metastasis in $[t, t + \Delta t)$ is $am_2(t)\Delta t + \mathrm{o}(\Delta t)$, while the probability of detection of a tumor is $bm_2(t)\Delta t + \mathrm{o}(\Delta t)$, where $m_2(t) = Y_2(t) + \cdots$.

Hypothesis 6. The probability of a systemic occurrence of a tumor in $[t, t + \Delta t)$ equals $\lambda\Delta t + \mathrm{o}(\Delta t)$, independent of the prior history of the patient.

Essentially, we shall attempt to develop the likelihood function for this model so that we can find the values of a, b, α, and λ which maximize the likelihood of the data set observed. It turns out that this is a formidable task indeed. The SIMEST algorithm which we develop later gives a quick alternative to finding the likelihood function. However, to give the reader some feel as to the complexity associated with model aggregation from seemingly innocent axioms, we shall give some of the details of getting the likelihood function. First of all, it turns out that in order to have any hope of obtaining a reasonable approximation to the likelihood function, we will have to make some further simplifying assumptions. We shall refer to the period prior to detection of the primary as Phase 0. Phase 1 is the period from detection of the primary to S', the first time of detection of a secondary tumor. For those patients without a secondary tumor, Phase 1 is the time of observation, S. Phase 2 is the time, if any, between S' and S. Now for the two simplifying axioms. T_0 is defined to be the (unobservable) time between the origination of the primary and the time when it is detected and removed (at time $t = 0$). T_1 and T_2 are the times until detection and removal of the first and second of the subsequent tumors (times to be counted from $t = 0$). We shall let X be the total mass of all tumors other than the primary at $t = 0$.

Hypothesis 7. For patients who do not display a secondary tumor, growth of the primary tumor, and of all tumors in Phase 1, is deterministically exponential with the growth of all other tumors treated as a pure birth process.

Hypothesis 8. For patients who display a secondary tumor, the growth of the following tumors is treated as deterministic: in Phase 0, tumors $Y_0(t)$ and $Y_1(t)$; in Phase 1, tumor $Y_1(t)$ and all tumors which originated in Phase 0; in Phase 2, all tumors. The growth of remaining tumors in Phases 0 and 1 is treated as a pure birth process.

We now define

$$\begin{aligned} H(s;t,z) = \exp\biggl\{ & \frac{az}{\alpha}\,\mathrm{e}^{\alpha t}\bigl(\mathrm{e}^{s}-1\bigr) \\ & \times \log\bigl[1+\bigl(\mathrm{e}^{-\alpha t}-1\bigr)\mathrm{e}^{-s}\bigr] \\ & + \frac{\lambda}{\alpha}s - \frac{\lambda}{\alpha}\log\bigl[1+\mathrm{e}^{\alpha t}\bigl(\mathrm{e}^{s}-1\bigr)\bigr]\biggr\} \end{aligned} \tag{13}$$

and

$$p(t;z) = bz\,\mathrm{e}^{\alpha t}\exp\biggl[-\frac{bz}{\alpha}\bigl(\mathrm{e}^{\alpha t}-1\bigr)\biggr] \tag{14}$$

Further, we shall define

$$w(y) = \lambda\Biggl[\int_0^y \mathrm{e}^{-\nu(u)}\,\mathrm{d}u - y\Biggr] \tag{15}$$

where $\nu(u)$ is determined from

$$u = \int_0^{\nu}\bigl(a+b+\alpha s - a\,\mathrm{e}^{-s}\bigr)^{-1}\,\mathrm{d}s \tag{16}$$

Then, we can establish the following propositions, and from these, the likelihood function:

$$p(T_0 > \tau) = \exp\Biggl[-b\int_0^{\tau}\mathrm{e}^{\alpha t}\,\mathrm{d}t\Biggr]$$

$$= \exp\left[-\frac{b}{\alpha}\left(e^{\alpha\tau} - 1\right)\right] \tag{17}$$

For patients who do not display a secondary tumor, we have

$$P(T_1 > S \mid X = x) = \exp\left[-x\nu(S) + w(S)\right] \tag{18}$$

For patients who develop metastases, we have

$$\begin{aligned} P(T_1 > S) &= P\big(\text{no secondary tumor in } (0, S)\big) \\ &= \int_0^\infty e^{w(s)} p(t; 1) H\big(\nu(s); t, 1\big)\, dt \end{aligned} \tag{19}$$

Similarly, for patients who do display a secondary tumor, we have

$$\begin{aligned} &P(T_1 = S', T_2 > S) \\ &= \int_0^\infty\int_0^t e^{w(S-S')} p(t; 1) p\big(S'; e^{\alpha u}\big)\big(\lambda + a\, e^{\alpha(t-u)}\big) \\ &\quad\times \exp\left[-\lambda(t-u) - \frac{a}{\alpha}\big(e^{\alpha(t-u)} - 1\big)\right] \\ &\quad\times H\big(\nu(S-S'); S', e^{\alpha u}\big) \\ &\quad\times H\big(\nu(S-S')\, e^{\alpha S'}; u,\ e^{\alpha(t-u)}\big)\, du\, dt \\ &\quad+ \int_0^\infty\int_0^{S'} e^{w(S-S')} p(t; 1) \\ &\quad\times \exp\left[-\lambda t - \frac{a}{\alpha}\big(e^{\alpha t} - 1\big)\right]\lambda\, e^{-\lambda u} p(S' - u; 1) \\ &\quad\times H\big(\nu(S-S'); S' - u, 1\big)\, du\, dt \end{aligned} \tag{20}$$

Finding the likelihood function, even a quadrature approximation to it, is more than difficult. Furthermore, current symbol manipulation programs (e.g., Mathematica, Maple) do not have the capability of doing the work. Accordingly, it must be done by hand. Approximately 1.5 person years were required to obtain a quadrature approximation to the likelihood. Before starting this activity, we had no idea of the rather practical difficulties involved. However, the activity was not without reward.

We found estimates for the parameter values using a data set consisting of 116 women who presented with primary breast cancer at the Curie-Sklodowska Cancer Institute in Warsaw (time units in months, volume units in cells): $a = 0.17 \times 10^{-9}$, $b = 0.23 \times 10^{-8}$, $\alpha = 0.31$, and $\lambda = 0.0030$. Using these parameter values, we found excellent agreement between the proportion free of metastasis versus time obtained from the data and that obtained from the model, using the parameter values given above. When we tried to fit the model to the data with the constraint that $\lambda = 0$ (that is, disregarding the systemic process as is generally done in oncology), the attempt failed.

One thing one always expects from a model-based approach is that, once the relevant parameters have been estimated, many things one had not planned to look for can be found. For example, tumor doubling time is 2.2 months. The median time from primary origination to detection is 59.2 months and at this time the tumor consists of 9.3×10^7 cells. The probability of metastasis prior to detection of the primary is 0.069, and so on. A model-based approach generally yields such serendipitous results, as a nonparametric approach generally does not. It is worth mentioning that, more frequently than one realizes, we need an analysis which is flexible, in the event that at some future time we need to answer questions different from those originally posed. The quadrature approximation of the likelihood is relatively inflexible compared to the simulation-based approach we shall develop shortly.

Insofar as the relative importance of the systemic and metastatic mechanisms, in causing secondary tumors associated with breast cancer, it would appear from Fig. 2 that the systemic one is the more important. This result is surprising, but is consistent with what we have seen in our exploratory analysis of another tumor system (melanoma). Interestingly, it is by no means true that for all tumor systems the systemic term has such dominance. For primary lung cancer, for example, the metastatic term appears to be far more important.

It is not clear how to postulate, in any definitive fashion, a procedure for testing the null hypothesis of the existence of a systemic mechanism in the progression of cancer. We have already noted that when we suppress the systemic hypothesis, we cannot obtain even a poor maximum likelihood fit to the data. However, someone might argue that a different set of nonsystemic axioms should have been proposed. Obviously, we cannot state that it is simply impossible to manage a good fit without the systemic hypothesis. However, it is true that the nonsystemic axioms we

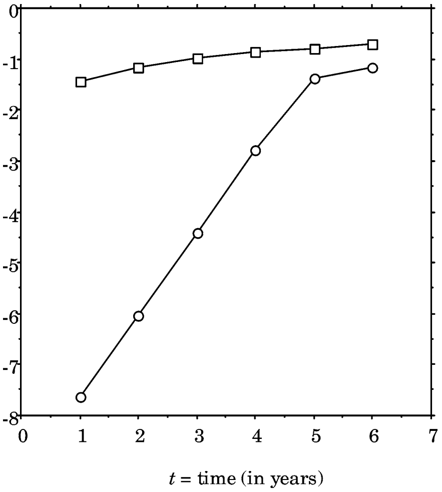

Fig. 2. Metastatic and systemic effects. ○ log[(prob(metastatic tumor originating by time t))], □ log[(prob(systemic tumor originating by time t))].

have proposed are a fair statement of traditional suppositions as to the growth and spread of cancer.

As a practical matter, we had to use data that were oriented toward the life of the patient rather than toward the life of a tumor system. This is due to the fact that human *in vivo* cancer data is seldom collected with an idea toward modeling tumor systems. For a number of reasons, including the difficulty mentioned in obtaining the likelihood function, deep stochastic modeling has not traditionally been employed by many investigators in oncology. Modeling frequently precedes the collection of the kinds of data of greatest use in the estimation of the parameters of the model. Anyone who has gone through a modeling exercise such as that covered in this section is very likely to treat such an exercise as a once in a lifetime experience. It simply is too frustrating to have to go through all the flailing around to come up with a quadrature approximation to the likelihood function. As soon as a supposed likelihood function has been found, and a corresponding parameter estimation algorithm constructed, the investigator begins a rather lengthy 'debugging' experience. The algorithm's failure to work might be due to any number of reasons (e.g., a poor approximation to the likelihood function, a poor quadrature routine, a mistake in the code of the algorithm, inappropriateness of the model, etc.). Typically, the debugging process is time consuming and difficult. If one is to have any hope for coming up with a successful model-based investigation, an alternative to the likelihood procedure for aggregation must be found.

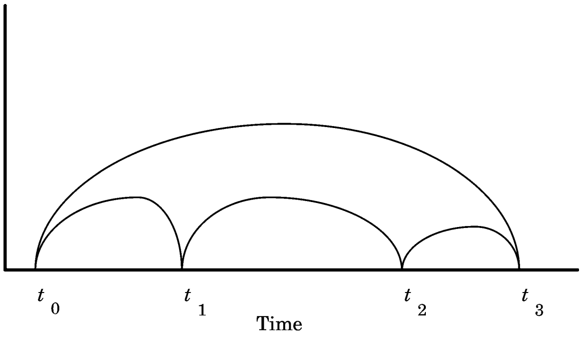

Fig. 3. Two possible paths from primary to secondary.

In order to decide how best to construct an algorithm for parameter estimation which does not have the difficulties associated with the classical closed-form approach, we should try to see just what causes the difficulty with the classical method of aggregating from the microaxioms to the macro level, where the data lives. A glance at Fig. 3 reveals the problem with the closed-form approach.

The axioms of tumor growth and spread are easy enough to implement in the forward direction. Indeed, they follow the natural forward formulation used since Poisson's work of 1837. Essentially, we are overlaying stochastic processes, one on top of the other, and interdependently to boot. But when we go through the task of finding the likelihood, we are essentially seeking all possible paths by which the observables could have been generated.

The secondary tumor, originating at time t_3, could have been thrown off from the primary at time t_3, or it could have been thrown off from a tumor which itself was thrown off from another tumor at time t_2 which itself was thrown off from a tumor at time t_1 from the primary which originated at time t_0. The number of possibilities is, of course, infinite.

In other words, the problem with the classical likelihood approach in the present context is that it is a backward look from a database generated in the forward direction. To scientists before the present generation of fast, cheap computers, the backward approach was, essentially, unavoidable unless one avoided such problems (a popular way out of the dilemma). However, we need not be so restricted.

Once we realize the difficulty when one uses a backward approach with a concatenation of forwardly axiomitized mechanisms, the way out of our difficulty is rather clear. We need to analyze the data using a forward formulation. The most obvious way to carry this out is to pick a guess for the underlying vector of parameters, put this guess in the micro-axiomitized model and simulate many times of appearance of secondary tumors. Then, we can compare the set of simulated quasidata with that of the actual data.

The greater the concordance, the better we will believe we have done in our guess for the underlying parameters. If we can quantitize this measure of concordance, then we will have a means for guiding us in our next guess. One such way to carry this out would be to order the secondary occurrences in the data set from smallest to largest and divide them into k bins, each with the same proportion of the data. Then, we could note the proportions of quasidata points in each of the bins.

If the proportions observed for the quasidata, corresponding to parameter value Θ, were denoted by $\{\pi_j(\Theta)\}_{j=1}^{k}$, then a Pearson goodness-of-fit statistic would be given by:

$$\chi^2(\Theta) = \sum_{j=1}^{k} \frac{(\pi_j(\Theta) - 1/k)^2}{\pi_j(\Theta)} \tag{21}$$

The minimization of $\chi^2(\Theta)$ provides us with a means of estimating Θ.

Typically, the sample size, n, of the data will be much less than N, the size of the simulated quasidata. With mild regularity conditions, assuming there is only one local maximum of the likelihood function, Θ_0, as $n \to \infty$ (which function we of course do not know), then as $N \to \infty$, as n becomes large and k increases in such a way that $\lim_{n\to\infty} k = \infty$ and $\lim_{n\to\infty} k/n = 0$, the minimum χ^2 estimator for Θ_0 will have an expected mean square error which approaches the expected mean square error of the maximum likelihood estimator. This is, obviously, quite a bonus. Essentially, we will be able to forfeit the possibility of knowing the likelihood function and still obtain an estimator with asymptotic efficiency equal to that of the maximum likelihood estimator. The price to be paid is the acquisition of a computer swift enough and cheap enough to carry out a very great number, N, of simulations, say 10 000. This ability to use the computer to get us out of the ‘backward trap’ is a potent but, as yet seldom used, bonus of the computer age. Currently, the author is using SIMEST on a 400 MHz personal computer, amply adequate for the task, which now costs around $1000.

First, we observe how the forward approach enables us to eliminate those hypotheses which were, essentially a practical necessity if a likelihood function was to be obtained. Our new axioms are simply:

Hypothesis 1. For any patient, each tumor originates from a single cell and grows at exponential rate α.

Hypothesis 2. The probability that the primary tumor will be detected and removed in $[t, t + \Delta t)$ is given by $bY_0(t)\Delta t + \mathrm{o}(\Delta t)$. The probability that a tumor of size $Y(t)$ will be detected in $[t, t + \Delta t)$ is given by $bY(t)\Delta t + \mathrm{o}(\Delta t)$.

Hypothesis 3. The probability of a metastasis in $[t, t + \Delta)$ is $a\Delta t \times$ (total tumor mass present).

Hypothesis 4. The probability of a systemic occurrence of a tumor in $[t, t + \Delta t)$ equals $\lambda\Delta t + \mathrm{o}(\Delta t)$, independent of the prior history of the patient.

In order to simulate, for a given value of (α, a, b, λ), a quasidata set of secondary tumors, we must first define:

t_{D} = time of detection of primary tumor;
t_{M} = time of origin of first metastasis;
t_{S} = time of origin of first systemic tumor;
t_{R} = time of origin of first recurrent tumor;
t_{d} = time from t_{R} to detection of first recurrent tumor;
t_{DR} = time from t_{D} to detection of first recurrent tumor.

Now, generating a random number u from the uniform distribution on the unit interval, if $F(\cdot)$ is the appropriate cumulative distribution function for a time, t, we set $t = F^{-1}(u)$. Then, assuming that the tumor volume at time t is:

$$v(t) = c\,\mathrm{e}^{\alpha t}, \quad \text{where } c \text{ is the volume of one cell} \tag{22}$$

we have

$$F_{\mathrm{M}}(t) = 1 - \exp\left(-\frac{a\,c}{\alpha}\,\mathrm{e}^{\alpha t_{\mathrm{M}}}\right) \tag{23}$$

Similarly, we have

$$\begin{aligned} F_{\mathrm{D}}(t_{\mathrm{D}}) &= 1 - \exp\left(-\int_0^{t_{\mathrm{D}}} b\,c\,\mathrm{e}^{\alpha\tau}\,\mathrm{d}\tau\right) \\ &= 1 - \exp\left(-\frac{b\,c}{\alpha}\,\mathrm{e}^{\alpha t_{\mathrm{D}}}\right), \end{aligned} \tag{24}$$

$$F_{\mathrm{S}} = 1 - \mathrm{e}^{-\lambda t_{\mathrm{S}}} \tag{25}$$

and

$$F_{\mathrm{d}}(t_{\mathrm{d}}) = 1 - \exp\left(-\frac{b\,c}{\alpha}\,\mathrm{e}^{\alpha t_{\mathrm{d}}}\right) \tag{26}$$

Using the actual times of discovery of secondary tumors $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_n$ we generate k bins. In actual tumor situations, because of recording protocols, we may not be able to put the same number of secondary tumors in each bin. Let us suppose that the observed proportions are given by $(p_1, p_2, \ldots, p_k)$. We shall generate N recurrences $s_1 < s_2 < \cdots < s_N$. The observed proportions in each of the bins will be denoted $\pi_1, \pi_2, \ldots, \pi_k$. The goodness of fit corresponding to (α, λ, a, b) will be given by:

$$\chi^2(\alpha, \lambda, a, b) = \sum_{j=1}^{k} \frac{(\pi_j(\alpha, \lambda, a, b) - p_j)^2}{\pi_j(\alpha, \lambda, a, b)} \tag{27}$$

As a practical matter, we may replace $\pi_j(\alpha, \lambda, a, b)$ by p_j, since with (α, λ, a, b) far away from truth, $\pi_j(\alpha, \lambda, a, b)$ may well be zero. Then the following algorithm generates the times of detection of quasi-secondary tumors for the particular parameter value (α, λ, a, b).

Algorithm. Secondary tumor simulation (α, λ, a, b).

```
Generate $t_{\mathrm{D}}$
$j = 0$
$i = 0$
Repeat until $t_{\mathrm{M}}(j) > t_{\mathrm{D}}$
$j = j + 1$
Generate $t_{\mathrm{M}}(j)$
Generate $t_{\mathrm{dM}}(j)$
$t_{\mathrm{dM}}(j) \leftarrow t_{\mathrm{dM}}(j) + t_{\mathrm{M}}(j)$
If $t_{\mathrm{dM}}(j) < t_{\mathrm{D}}$, then $t_{\mathrm{dM}}(j) \leftarrow \infty$
Repeat until $t_{\mathrm{S}} > 10 t_{\mathrm{D}}$
$i = i + 1$
Generate $t_{\mathrm{dS}}(i)$
$t_{\mathrm{dS}}(i) \leftarrow t_{\mathrm{dS}}(i) + t_{\mathrm{S}}(i)$
$s \leftarrow \min[t_{\mathrm{dM}}(j), t_{\mathrm{dS}}(i)]$
Return $s$
End repeat
```

The algorithm above does still have some simplifying assumptions. For example, we assume that metastases of metastases will probably not be detected before the metastases themselves. We assume that the primary will be detected before a metastasis, and so on. Note, however, that the algorithm utilizes much less restrictive simplifying assumptions than those which led to the terms of the closed-form likelihood. Even more importantly, the Secondary Tumor Simulation algorithm can be discerned in a few minutes, whereas a likelihood argument is frequently the work of months.

Another advantage of the forward simulation approach is its ease of modification. Those who are familiar with 'backward' approaches based on the likelihood or the moment generating function are only too familiar with the experience of a slight modification causing the investigator to go back to the start and begin anew. This is again a consequence of the tangles required to be examined if a backward approach is used. However, a modification of the axioms generally causes slight inconvenience to the forward simulator.

For example, we might add the following

Hypothesis 5. A fraction γ of the patients ceases 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.

Clearly, adding Hypothesis 5 will cause considerable work if we insist on using the classical aggregation approach of maximum likelihood. However, in the forward simulation method we simply add the following lines to the secondary tumor simulation code:

```
Generate $u$ from $U(0, 1)$
If $u > \gamma$, then proceed as in the secondary tumor simulation code
```

> If $u < \gamma$, then proceed as in the secondary tumor simulation code except replace the step 'Repeat until $t_S > 10t_D$' with the step 'Repeat until $t_S(i) > t_D$'.

In the discussion of metastasis and systemic occurrence of secondary tumors, we have used a model supported by data to try to gain some insight into a part of the complexities of the progression of cancer in a patient. Perhaps this sort of approach should be termed *speculative data analysis.*

In the current example, we were guided by a nonparametric intensity function estimate, which was surprisingly nonincreasing, to conjecture a model, which enabled us to test systemic origin against metastatic origin on something like a level playing field. The fit without the systemic term was so bad that anything like a comparison of goodness-of-fit statistics was unnecessary.

It is interesting to note that the implementation of SIMEST is generally faster on the computer than working through the estimation with the closed-form likelihood. In the four-parameter oncological example we have considered here, the running time of SIMEST was 10% of the likelihood approach. As a very practical matter, then, the simulation-based approach would appear to majorize that of the closed-form likelihood method in virtually all particulars. The running time for SIMEST can begin to become a problem as the dimensionality of the response variable increases past one. Up to this point, we have been working with the situation where the data consists of failure times. In the systemic versus metastatic oncogenesis example, we managed to estimate four parameters based on this kind of one-dimensional data. As a practical matter, for tumor data, the estimation of five or six parameters for failure time data is the most one can hope for. Indeed, in the oncogenesis example, we begin to observe the beginnings of singularity for four parameters, due to a near trade-off between the parameters a and b. Clearly, it is to our advantage to be able to increase the dimensionality of our observables. For example, with cancer data, it would be to our advantage to utilize not only the time from primary diagnosis and removal to secondary discovery and removal, but also the tumor volumes of the primary and the secondary. Such information enables one to postulate more individual growth rates for each patient. Thus, it is now appropriate to address the question of dealing with multivariate response data.

Gaussian template criterion. In many cases, it will be possible to employ a procedure using a criterion function. Such a procedure has proved quite successful in another context (see [5], pp. 275–280). First, we transform the data $\{X_i\}_{i=1}^n$ by a linear transformation such that for the transformed data set $\{U_i\}_{i=1}^n$ the mean vector becomes zero and the covariance matrix becomes I:

$$U = AX + b \tag{28}$$

Then, for the current best guess for Θ, we simulate a quasidata set of size N. Next, we apply the same transformation to the quasidata set $\{Y_j(\Theta)\}_{j=1}^N$, yielding $\{Z_j(\Theta)\}_{j=1}^N$. Assuming that both the actual data set and the simulated data set come from the same density, the likelihood ratio $\Lambda(\Theta)$ should increase as Θ gets closer to the value of Θ, say Θ_0, which gave rise to the actual data, where,

$$\Lambda(\Theta) = \frac{\prod_{i=1}^n \exp[-\frac{1}{2}(u_{1i}^2 + \cdots + u_{pi}^2)]}{\prod_{i=1}^N \exp[-\frac{1}{2}(z_{1i}^2 + \cdots + z_{pi}^2)]} \tag{29}$$

As soon as we have a criterion function, we are able to develop an algorithm for estimating Θ_0. The closer Θ is to Θ_0, the smaller will $\Lambda(\Theta)$ tend to be.

The procedure above which uses a single Gaussian template will work well in many cases where the data has one distinguishable center and a falling off away from that center which is not too 'taily'. However, there will be cases where we cannot quite get away with such a simple approach. For example, it is possible that a data set may have several distinguishable modes and/or exhibit very heavy tails. In such a case, we may be well advised to try a more local approach. Suppose that we pick one of the n data points at random – say x_1 – and find the m nearest neighbors amongst the data.

We then treat this m nearest-neighbor cloud as if it came from a Gaussian distribution centered at the sample mean of the cloud and with covariance matrix estimated from the cloud. We transform these $m + 1$ points to zero mean and identity covariance matrix, via

$$U = A_1X + b_1 \tag{30}$$

Now, from our simulated set of N points, we find the $N(m + 1)/n$ simulated points nearest to the mean of the $m+1$ actual data points. This will give us an expression like

$$\Lambda_1(\Theta) = \frac{\prod_{i=1}^{m+1} \exp[-\frac{1}{2}(u_{1i}^2 + \cdots + u_{pi}^2)]}{\prod_{i=1}^{N(m+1)/n} \exp[-\frac{1}{2}(z_{1i}^2 + \cdots + z_{pi}^2)]} \quad (31)$$

If we repeat this operation for each of the n data points, we will have a set of local likelihood ratios $\{\Lambda_1, \Lambda_2, \ldots, \Lambda_n\}$. Then one natural measure of concordance of the simulated data with the actual data would be

$$\boldsymbol{\Lambda}(\Theta) = \sum_{i=1}^{n} \log\bigl(\Lambda_i(\Theta)\bigr) \quad (32)$$

We note that this procedure is not equivalent to one based on density estimation, since the nearest-neighbor ellipsoids are not disjoint. Nevertheless, we have a level playing field for each of the guesses for Θ and the resulting simulated data sets.

A simple counting criterion. Fast computing notwithstanding, with n in the 1000 range and N around 10,000, the template procedure can become prohibitively time consuming. Accordingly, we may opt for a subset counting procedure:

For data size n, pick a smaller value, say nn.

Pick a random subset of the data points of size nn.

Pick a nearest neighbor outreach parameter m, typically $0.02n$.

For each of the nn data points, X_j, find the Euclidean distance to the mth nearest neighbor, say $d_{j,m}$.

For an assumed value of the vector parameter Θ, generate N simulated observations.

For each of the data points in the random subset of the data, find the number of simulated observations within $d_{j,m}$, say $N_{j,m}$.

Then the criterion function becomes

$$\chi^2(\Theta) = \sum_{j=1}^{nn} \frac{((m+1)/n - N_{j,m}/N)^2}{(m+1)/n}$$

Experience indicates that whatever nn size subset of the data points is selected should be retained throughout the changes of Θ. Otherwise, practical instability may obscure the path to the minimum value of the criterion function.

A SIMDAT-SIMEST stopping rule. We may use the resampling algorithm SIMDAT to compare the results from resampled data points with those from model-based simulations. SIMDAT is not a simple resampling so much as it is a stochastic interpolator. We can take the original data and use SIMDAT to generate a SIMDAT pseudodata set of N values.

Then, for a particular guess of Θ, we can compute a SIMEST pseudodata set of N values. For any region of the space of the vector observable, the number of SIMEST-generated points should be approximately equal to the number of SIMDAT-generated points. For example, let us suppose that we pick nn of the n original data points and find the radius $d_{j,m}$ of the hypersphere which includes m of the data points for, say, point X_j. Let $N_{j,\mathrm{SD}}$ be the number of SIMDAT-generated points falling inside the hypersphere and $N_{j,\mathrm{SE}}$ be the number of SIMEST-generated points falling inside the hypersphere. Consider the empirical goodness-of-fit statistic for the SIMDAT cloud about point X_j:

$$\chi^2_{j,\mathrm{SD}}(\Theta) = \frac{((m+1)/n - N_{j,\mathrm{SD}}/N)^2}{(m+1)/n}$$

For the SIMEST cloud, we have

$$\chi^2_{j,\mathrm{SE}}(\Theta) = \frac{((m+1)/n - N_{j,\mathrm{SE}}/N)^2}{(m+1)/n}$$

If the model is correct and if our estimate for Θ is correct, then $\chi^2_{j,\mathrm{SE}}(\Theta)$ should be, on the average, distributed similarly to $\chi^2_{j,\mathrm{SE}}(\Theta)$. Accordingly, we can construct a sign test. To do so, let:

$$\begin{aligned} W_j &= +1 \text{ if } \chi^2_{j,\mathrm{SD}}(\Theta) \geqslant \chi^2_{j,\mathrm{SE}}(\Theta) \\ &= -1 \text{ if } \chi^2_{j,\mathrm{SD}}(\Theta) < \chi^2_{j,\mathrm{SE}}(\Theta) \end{aligned}$$

So, if we let:

$$Z = \frac{\sum_{j=1}^{nn} W_j}{\sqrt{nn}}$$

we might decide to terminate our search for estimating Θ when the absolute value of Z falls below 3 or 4.

References

[1] J.R. Thompson, Simulation: A Modeler's Approach, Wiley, New York, 2000.

[2] S.D. Poisson, Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précedées des règles générales du calcul des probabilités, Paris, 1837.

[3] A. Stuart, J.K. Ord, in: Kendall's Advanced Theory of Statistics, vol. 1, fifth ed., Oxford University Press, New York, 1987, p. 7.

[4] R. Bartoszyński, B.W. Brown, J.R. Thompson, Metastatic and systemic factors in neoplastic progression, in: L. LeCam, J. Neyman (Eds.), Probability Models and Cancer, Academic Press, New York, 1982, pp. 253–264, 283–285.

[5] J.R. Thompson, R.A. Tapia, in: Nonparametric Function Estimation, Modeling, and Simulation, SIAM, Philadelphia, 1990, pp. 214–226.