Scand J Statist 16: 177–225, 1989

Survival Models and Martingale Dynamics

ELJA ARJAS
University of Oulu

ABSTRACT. This paper reviews some ideas and techniques which appear central in the statistical modelling of longitudinal observational data. The presentation is based entirely on the framework of marked point processes, with an emphasis on the dynamical aspects of the modelling. The concepts of hazard and prediction are given special attention, together with the information which forms their natural basis. Finally, we discuss a number of structural questions relating to likelihood-based inference from marked point processes.

Key words: marked point process, hazard, prediction process, filtering, event history analysis, partial model specification, unobservable random variation

1. Introduction

1.1 Motivation

The terms "event history analysis" and "life history analysis" have become very common in the recent discussion on the methodologies in social sciences (e.g. Allison, 1984; Thygesen, 1988; Manton & Stallard, 1988). This is quite understandable: cross-sectional methods can never provide the same kind of understanding as longitudinal ones about how the individuals or objects in question develop. In fact, the importance of the longitudinal aspect has long been recognized in some related areas, for example, in medical research and epidemiology, and in engineering pertaining to the operation or reliability of technical systems.

Although the benefits from a longitudinal approach can be easily understood, there are also reasons why their use is not more widespread:

(a) The statistical methodologies for analysing longitudinal data are not yet fully developed; there are conceptual as well as mathematical difficulties in modelling, and the experience about how to interpret the results from empirical work is largely insufficient.
(b) Often the quality of the data prevents the execution of a meaningful longitudinal analysis; and finally,
(c) The advanced methods tend to be computationally involved. The available statistical packages have still considerable weaknesses and the time requirements may also prevent their efficient full-scale use.

Presently, there are many developments which contribute to the elimination of (b) and (c). That (c) becomes less and less crucial is of course largely a consequence of the improving technologies in computing. Objection (b), on the other hand, has received a great deal of attention recently and many data registers, some of enormous potential interest, are being compiled or revised accordingly. As a result, it seems that the bottleneck is formed by the lack of understanding about how to model and analyse longitudinal observational data and draw valid conclusions from such studies.

This paper is not intended as a systematic review. Rather, it could be described as a mixed collection of ideas about modelling observational study data, reflecting my own research (and taste). A reader who is more interested in the statistical and computational aspects should read the recent reviews by Andersen & Borgan (1985) and Clayton (1988). The monograph of

Kalbfleisch & Prentice (1980) provides an excellent background. I will try to proceed with an almost absolute minimum of mathematical technicalities, often to the extent that the stated results can be slightly inaccurate. The references given will then provide the full details, and I will try to emphasize correct interpretations. Although this is not an "applied paper", as it concentrates more on ideas, I am making a serious attempt to illustrate everything discussed by examples.

I would also like to stress that the real difficulties in longitudinal observational studies are very often problem-specific and cannot be "solved" in general terms. The scope of this paper is to describe to a non-specialist a technical framework within which, I hope, some of these intrinsic difficulties could be formulated and discussed.

1.2. Outline of the mathematical framework

Longitudinal observational data consist of registered events which are ordered progressively in time. The time variable can be calendar time or some other quantity related to it, for example, age. The data are here viewed as a sample path of a *marked point process* (MPP) $\{(T_n, X_n); n \geq 1\}$, where $T_1 < T_2 < \ldots$ is the ordered sequence of "occurrence times" and X_n is a description of the event at T_n.

This framework is extremely flexible. There are no "states" or "boxes", or transitions between such, as are commonly encountered in applied probability models. Rather, each point (T_n, X_n) could be viewed as a "landmark" which is reached during the follow-up. We can, of course, specialize in different ways of introducing more structural assumptions, and will then obtain more conventional models such as Markov, Markov renewal, or other.

Apart from the time variable t and the marked points (T_n, X_n) there is one more basic ingredient in our approach: the pre-t-history $\mathcal{F}_t$, typically consisting of the information contained in the marked points prior to t. The idea is to consider the probability distribution of the future and/or the past marked points, conditionally on the observed *information* $\mathcal{F}_t$. In general terms, we can call these problems *prediction* and *state estimation*. Letting the time t sweep over the observational interval makes this description dynamic.

Often in this framework, it is convenient to distinguish between a *response* event and a set of *covariates*, used to explain the occurrence or non-occurrence of the response. However, this division is not clear-cut; in fact, the response event can sometimes repeat itself many times (say, spells of unemployment, or hospitalizations in a chronic disease), and then it can be that the earlier occurrences of the same event have concrete predictive value. On the other hand, the conditioning covariate information is in no way canonical, and different levels of conditioning in the statistical models will typically provide answers to questions of a different nature. The role of information is one of the central themes in this paper.

2. Probability, information and hazard

2.1. The "single-point" process and hazard

2.1.1. The ingredients

We consider a positive real random variable T, defined on a probability space $(\Omega, \mathcal{F}, P)$. In applications T is typically the *life length* of an object, or, more generally, the waiting time until some response of interest occurs. T can be naturally identified with the simple counting process $N=(N_t)$ defined by

$$N_t = 1_{\{T \leq t\}}, \qquad t \geq 0, \tag{1}$$

which counts "one" at T. Thus, for $\omega \in \Omega$, the sample path $T \to N_t(\omega)$ looks as follows:

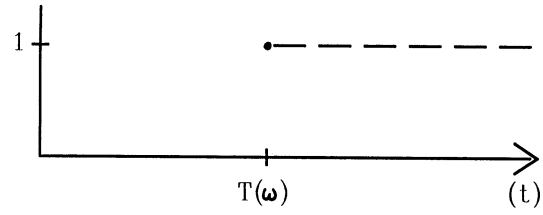


Fig. 1. A sample path of the "single-point" counting process.

Note that the sample path is right continuous at the jump point $T(\omega)$.

As a second key ingredient, we consider an increasing family of σ-fields $\mathrm{F}=(\mathscr{F}_t)$: for all $0 \leqslant s \leqslant t$, $\mathscr{F}_s \subset \mathscr{F}_t \subset \mathscr{F}$. We call F a *history* (or filtration), and view $\mathscr{F}_t$ as the information available to an observer at time t, calling it briefly *pre-t-information.* $\mathscr{F}_t$ will typically be generated by the pre-t evolution of some observable process (or processes), and this agrees well with our assumption that $\mathscr{F}_t$ should increase as the observation time t increases. We add the technical "usual conditions" that (i) $(\mathscr{F}_t)$ is right continuous, and (ii) $\mathscr{F}_0$ contains all null-sets.

It is important to allow for this flexibility of a general definition of F, not fixing it to be generated by some particular observable process. Different problems call for different types of conditioning, and sometimes it is natural to move from one level of information to another. This is, in fact, explicitly discussed in section 3.2.

2.1.2. Classical definition of hazard rate

We start our discussion of the concept of hazard by briefly reviewing the now classical notions based on the distribution function.

Let

$$F(t)=P(T \leqslant t), \qquad t \geqslant 0, \tag{1}$$

be the distribution function of T, and

$$\bar{F}(t)=P(T>t)=1-F(t), \qquad t \geqslant 0, \tag{2}$$

the corresponding survival function. Supposing absolute continuity of F and writing f for the corresponding density, *the hazard rate* r is defined by

$$r(t)=\frac{f(t)}{\bar{F}(t)}, \qquad t \geqslant 0. \tag{3}$$

We now list some fundamental properties of the hazard rate:

(i) The differential $r(t)\,dt$ has the interpretation of "the conditional failure probability at time t, given the survival to at least t". This follows from the simple sequence of equations, where we are once using the continuity of F:

$$r(t)\,dt=\frac{f(t)\,dt}{\bar{F}(t)}=\frac{f(t)\,dt}{\bar{F}(t-)} \tag{4}$$

$$=\frac{P(T \in dt)}{P(T \geqslant t)}=P(T \in dt \mid T \geqslant t).$$

This interpretation of hazard will serve as a guideline to our later and more general definitions.

(ii) The hazard rate determines uniquely the distribution function. In fact, (3) is the same as

$$r(t)=-\frac{(d/dt)\bar{F}(t)}{\bar{F}(t)},$$

and since $\bar{F}(0)=1$, this has the solution:

$$\bar{F}(t)=\exp\left(-\int_0^t r(s)\,ds\right), \qquad t\geq 0. \tag{5}$$

This is often called the *exponential formula*. Denoting the integrated (or cumulative) hazard by $R(t)=\int_0^t r(s)\,ds$, we clearly have that $R(t)=-\log\bar{F}(t)$.

(iii) There are two ways in the literature of extending the above definitions to non-continuous distributions. An obvious possibility is to require that $R(t)=-\log\bar{F}(t)$ should continue to hold, in which case R makes jumps of size

$$\Delta R(t)=R(t)-R(t-)=\log\left(1+\frac{\Delta F(t)}{1-F(t)}\right)$$

at the jump points of F. For us here it is, however, much more convenient to require that the intuitive interpretation of hazard given in (i) continues to hold more generally. Using the "measure notation" $R(t)=\int_0^t R(ds)$ and $F(t)=\int_0^t F(ds)$ we therefore define

$$R(dt)=\frac{F(dt)}{\bar{F}(t-)}\ \{=P(T\in dt\,|\,T\geq t)\}, \qquad t\geq 0. \tag{6}$$

At the jump points of F we then simply have

$$\Delta R(t)=\frac{\Delta F(t)}{\bar{F}(t-)}=P(T=t\,|\,T\geq t). \tag{7}$$

Notice that here it is important, in order to save the interpretation on the right, to use the left limit $\bar{F}(t-)$ in the denominator.

We can then follow the well-known technique of decomposing a distribution into the discrete and the continuous part, and simply define the continuous part of the cumulative hazard by

$$R^c(t)=R(t)-\sum_{s\leq t}\Delta R(s).$$

(Note that the sum on the right has at most a countable number of non-zero terms.)

Consider finally the question which corresponds to (ii) above: does the definition (6) determine F uniquely from a given R? As a guideline we can use (ii), and the fact that for a purely discrete distribution, by an application of the chain rule of conditional probabilities, obviously

$$\bar{F}(t)=P(T>t)=\prod_{s\leq t}P(T>s\,|\,T\geq s)=\prod_{s\leq t}\{1-\Delta R(s)\}.$$

In fact, it can be shown that the solution of (6) is unique also in general, and that

$$\bar{F}(t)=\exp\{-R^c(t)\}\prod_{s\leq t}\{1-\Delta R(s)\}, \qquad t\geq 0 \tag{8a}$$

see, for example, Jacod (1975), or Lipster & Shiryayev (1978). (8a) is called Doléans–Dade's

exponential formula. It is often more convenient to write the right-hand side as a *product integral* see Gill & Johansen (1989), leading to the simpler expression

$$\bar{F}(t)=\prod_{s\leqslant t}\{1-R(ds)\}, \qquad t\geqslant 0. \tag{8b}$$

2.1.3. *Information-based hazard*

We interpret the randomness of T simply as an expression of the fact that the true value $T(\omega)$ is unknown to the observer. An assessment of the possibility of T occurring, if made at time t, is naturally expressed in terms of the conditional probability, given the information then available. We save the notation $R(dt)$ for the classical notion of hazard discussed above and denote the F-history based notion by $A^{\mathrm{F}}(dt)$. Throughout, our guideline is that $A^{\mathrm{F}}(dt)$ should have the interpretation

$$A^{\mathrm{F}}(dt)=P(T\in dt|\mathscr{F}_{t-}), \tag{1}$$

and that the cumulative hazard A_t^{F} is the integral

$$A_t^{\mathrm{F}}=\int_0^t A^{\mathrm{F}}(dt). \tag{2}$$

Intuitively, this corresponds to predicting whether T is going to occur "now", on the basis of all observations available up to (but not including) the present. In practice this means that $A^{\mathrm{F}}(dt)$ can be expressed as a function of the strict pre-t sample path of the process which generates F.

Example (*i*). If $\mathscr{F}_t=\{\Phi, \Omega\}$="trivial history" for all $t\geqslant 0$, then obviously (almost surely)

$$A^{\mathrm{F}}(dt)=P(T\in dt|\mathscr{F}_{t-})=P(T\in dt)=F(dt).$$

This is the case when "nothing is observed, not even T when it occurs".

Example (*ii*). Considering then $\mathscr{F}_t^N=\sigma\{T; T\leqslant t\}=\sigma\{N_s; s\leqslant t\}$ we have the case where "one observes T when it occurs but nothing else". (We do not show explicitly the inclusion of the null-sets.) Now the definition of $A^{\mathrm{F}}(dt)$ splits into two cases depending on whether T actually occurred before t or not (we write $\mathrm{F}^N=(\mathscr{F}_t^N)$):

$$A^{\mathrm{F}^N}(dt)=P(T\in dt|\mathscr{F}_{t-})=\begin{cases}P(T\in dt|\,T\geqslant t)=R(dt) & \text{on } \{T\geqslant t\}\\ P(T\in dt|\,T<t)=0 & \text{on } \{T<t\}\end{cases}. \tag{3}$$

Thus this definition is closely connected to the classical one: it agrees with the latter until T occurs, and is zero thereafter. In terms of the cumulative hazard,

$$A_t^{\mathrm{F}^N}=\int_0^t 1_{\{T\geqslant s\}}R(ds)=R(t\wedge T), \qquad t\geqslant 0. \tag{4}$$

Example (*iii*). We can also consider cases where T can be predicted from F exactly. The simplest such situation is that $\sigma\{T\}\subset\mathscr{F}_0$. Then

$$A^{\mathrm{F}}(dt)=P(T\in dt|\mathscr{F}_{t-})=1_T(dt),$$

i.e.

$$A_t^{\mathrm{F}}=\int_0^t 1_T(dt)=1_{\{T\leqslant t\}}=N_t, \qquad t\geqslant 0.$$

We then discuss briefly the precise mathematical *definition of* A^F: suppose that T is an F-stopping time, i.e. $\{T\leq t\}\in\mathcal{F}_t$ for all $t\geq 0$. Then (A_t^F) with $A_0^F=0$ is defined as the unique (up to indistinguishability) F-predictable process such that the difference

$$M_t=N_t-A_t^F, \qquad t\geq 0, \tag{5}$$

is an F-martingale. (A sufficient condition for the F-predictability of a process is that it is F-adapted and left continuous.)

It is not possible for us here to cover the full technical background behind this definition. For this we have to refer to general works such as Jacod (1975), Brémaud & Jacod (1977), Brémaud (1981), Liptser & Shiryayev (1978), Karr (1986) or Brown (1988). The general definitions lacking, it is perhaps useful to think about the discrete time case $t\in\{0,1,2,\ldots\}$ as a prototype because it is so simple: (A_t^F) is predictable if A_t^F is in fact $\mathcal{F}_{t-1}$-measurable for all $t\geq 1$, and (M_t) is an F-martingale if the differences $\Delta M_t=M_t-M_{t-1}$ are integrable and satisfy the property $E(\Delta M_t|\mathcal{F}_{t-1})=0$. But both these properties are easily satisfied if we set $\Delta A_t^F=E(\Delta N_t|\mathcal{F}_{t-1})$ and $A_t^F=\Sigma_{s\leq t}\Delta A_s^F$, which is nothing else than the discrete time version of 3.1.3.(1) and (2) above. Fortunately (1) and (2), although they cannot replace the exact martingale definition in continuous time and when the history is general, will actually be adequate for our somewhat sketchy presentation.

In general theory (A_t^F) is usually called the F-*compensator* of (N_t), reflecting the idea that (A_t^F) is predictable and (M_t) is "unpredictable noise". We end with some complementary remarks:

Remark 1. If T is not an F-stopping time, as in example 1 above, the general definition still applies if N_t is replaced by $E(N_t|\mathcal{F}_t)$ (choosing a right-continuous version with left limits). This process is in general no longer increasing, but it is easily seen to be a submartingale and can therefore be "compensated" by an increasing process (A_t^F).

Remark 2. If there are more points, say at times $(0<)T_1<T_2<\ldots$, with $T_n\to\infty$ as $n\to\infty$, we can easily extend the notion of hazard by letting

$$N_t=\sum_n 1_{\{T_n\leq t\}} \tag{6}$$

count the number of points up to time t, and replacing (1) by

$$A^F(dt)=E\{N(dt)|\mathcal{F}_{t-}\}. \tag{7}$$

For the exact definition, we would then require that for every $n\geq 1$ the differences $N_{t\wedge T_n}-A^F(t\wedge T_n)$ have the F-martingale property.

Remark 3. In the case where (A_t^F) is absolutely continuous, i.e. we can write it as an integral

$$A_t^F=\int_0^t \lambda_s^F\, ds, \qquad t\geq 0, \tag{8}$$

(λ_t^F) is called the (stochastic) F-*intensity* of T (or of N). In example (ii) clearly, if F is absolutely continuous, we have

$$\lambda_t^{F^N}=r(t)\cdot 1_{\{T\geq t\}}, \qquad t\geq 0. \tag{9}$$

Remark 4. One may question whether the exponential formula 2.1.2. (5), or 2.1.2.(8a, b), extends from the classical case to general F-conditional hazards. In view of (9) above, it would

be a natural attempt to look for a process (ϱ_t) such that $\lambda_t^{\mathrm{F}}=\varrho_t \cdot 1_{\{T\geq t\}}$ and $P(T>t|\mathcal{F}_t^*)=\exp\{-\int_0^t \varrho_s\,ds\}$, for some conveniently defined history $(\mathcal{F}_t^*)$. For $\mathrm{F}=\mathrm{F}^N$ and $\mathcal{F}_t^*=$ the trivial σ-field we know this to be true from 2.1.2.(5), and the result continues to hold, under obvious regularity conditions, if $\mathcal{F}_t^*$ is fixed for all t, say $\mathcal{F}_t^*=\mathcal{F}_0$, and $\mathcal{F}_t=\mathcal{F}_0\vee\mathcal{F}_t^N$. If this were true more generally, we could, in the absolutely continuous case, use the easy analytical definition $-\log P(T>t|\mathcal{F}_t^*)$ for the cumulative hazard, only cutting it off at T as in (4) above. However, this attempt is not successful without some further conditions (see Yashin & Arjas, 1988). This question appears to be the source of some confusion in the literature.

2.1.4. The unit-exponentiality of total hazards

The random variable $A_T^{\mathrm{F}}=A_\infty^{\mathrm{F}}$ can be thought of as the total hazard an individual experiences during his lifetime. Although the process (A_t^{F}) depends on the history F, it is interesting that, subject to regularity conditions, A_T^{F} is always unit-exponential. We have the following result.

Proposition

Consider a life length $T<\infty$ and a history F *such that (A_t^{F}) is continuous. Then A_T^{F} is a unit-exponential random variable.*

Intuitively, this result can be understood to express a balance between the speed at which the cumulative hazard increases in time, as long as the individual is alive, and the probability that the cumulation stops because of failure at T. In the case of example (ii) where $A_t^{\mathrm{F}^N}=R(t\wedge T)$, the proposition is a simple consequence of a time change: we have, for $u\geq 0$,

$$\begin{aligned} P(A_T^{\mathrm{F}}>u) &= P\{R(T)>u\} \\ &= P\{T>R^{-1}(u)\} \\ &= \exp\left[-R\{R^{-1}(u)\}\right] \text{ (by the exponential formula)} \\ &= \exp(-u). \end{aligned}$$

Here $R(t)$ can be viewed as the reading of "the hazard clock" at real time t. According to this clock, the hazard rate is always one. This idea of a time change is the same as the well-known link between a non-homogeneous Poisson and the Poisson-(1) process. An equally simple proof of the proposition can be based on the fact that $\bar{F}(T)$ is a $(0,1)$-uniformly distributed random variable. Apparently, the simplest of the proofs which apply for general histories is based on the fact that $(1+u)^{N_t}\exp(-uA_t^{\mathrm{F}})$ $(u\geq 0)$ is an F-martingale (e.g. Norros, 1986).

It is quite remarkable that the unit-exponential total hazards experienced by different individuals are actually independent. This holds no matter how dependent the actual lifetimes are and what the history, as long as simultaneous failures are ruled out. This property is useful in simulation and it can also be applied for obtaining interesting results on stochastic comparison see Norros (1986), and Shaked & Shanthikumar (1987). For results concerning non-continuous compensators we refer to Arjas & Haara (1988) and Brown & Nair (1988).

We close this section on exponentiality by a thought which the reader can ponder about: considering only the distribution of the total hazards, everybody's "pure luck" for long life could be sampled before birth and independently from the unit-exponential distribution. It would then depend on the speed of cumulation of hazard how long this luck lasts in real time. Although the hazards depend on what history F is used, and one can indeed consider different histories, the unit exponentiality remains valid as long as the exact time of death is totally unpredictable in the sense that A^{F} is continuous.

2.2. Marked points and hazard

2.2.1. A single marked point

We now extend the above single-point framework by associating with such a point another random variable, a *mark*. We define a *marked point* as a pair (T, X) of random variables, where $T>0$ is as above and X takes values in a set E. For simplicity, we assume that E is countable, and denote its elements by x. The *mark* X has typically the role of "describing what happens at T", such as indicating the cause of failure or the failure pattern. To fix ideas, we consider the following simple example.

Example A. Let $S=(S_1, S_2, \ldots, S_k)$ be a vector of k life lengths, say, of the k parts in a device. Let F be the corresponding joint distribution in R^k, and suppose that $P(S_i \neq S_j)=1$ for all $i \neq j$. Then define (T, X) by

$$\begin{cases} T=\bigwedge_{i=1}^{k} S_i \\ X=i \qquad \text{on } \{T=S_i\}. \end{cases} \tag{1}$$

In other words, T is the time of a first part failure and X is the index of that part. It is rather obvious how *part-specific hazard rates*, say, $r_i(t)$, $t \geqslant 0$, $1 \leqslant i \leqslant k$, can be defined if one is to follow the ideas presented in 2.1 above. Denoting the "sub-distributions" by $F_i(dt)=P(S_i \in dt,\ S_j>t \text{ for all } j \neq i)=P(T \in dt,\ X=i)$, and supposing absolute continuity, we let

$$r_i(t)\ dt=\frac{F_i(dt)}{F([t,\infty)\times\ldots\times[t,\infty))}=\frac{P(T \in dt,\ X=i)}{P(T \geqslant t)}=P(T \in dt,\ X=i \mid T \geqslant t) \tag{2}$$

It is also obvious, from the additivity of probability, that the sum $\Sigma_{i=1}^{k} r_i(t)$ is the same as the "crude" hazard rate $r(t)$. This is an example of the additivity property of mark-specific hazards over the marks, which holds in complete generality as we shall see below.

Remark. In earlier literature the additivity property of the mark-specific hazards $r_i(t)$ and the independence of the variables S_i were often thought to be equivalent. The source of this confusion appears to be the exponential formula 2.1.2.(5): assuming absolute continuity and independence, it is certainly true that

$$\exp\left(-\int_0^t r(s)\ ds\right)=P(T>t)=\prod_{i=1}^{k} P(S_i>t)=\prod_{i=1}^{k} \exp\left(-\int_0^t r_i(s)\ ds\right)$$

$$=\exp\left(-\int_0^t \sum_{i=1}^{k} r_i(s)\ ds\right),$$

and therefore $r(t)=\Sigma_{i=1}^{k} r_i(t)$ must hold. However, additivity of hazards holds always and does not imply independence. Neither is independence something that can be decided upon on the basis of the distribution of (T, X). In fact, starting from any distribution F as in example A above, one can construct independent variables $S_1^*, S_2^*, \ldots, S_k^*$, say, such that the corresponding marked point, defined as in (1), has the same distribution as (T, X). This follows easily if one lets $P(S_i^*>t)=\exp\{-\int_0^t r_i(s)\ ds\}$. (This non-identifiability property was noted by Cox (1959, 1962) and was studied in detail by Tsiatis (1975). For extensions see, for example, Arjas & Greenwood (1981).)

After this introductory example we now take up the F-history-based notion of hazard for a marked point. Let, for $x \in E$,

$$N_t(x)=1_{\{T\leq t, X=x\}}, \qquad t\geq 0, \tag{3}$$

be the counting process which counts "one" at T if $X=x$, and nothing if $X\neq x$. Then obviously

$$N_t=\sum_{x\in E} N_t(x) \tag{4}$$

holds. We assume for simplicity that F is such that $\{T\leq t, X=x\} \in \mathcal{F}_t$ for all t and x. i.e. "(T, X) is observed in F at the time it occurs". As a straightforward extension of what was done in 2.1.3 above, we then would like to have *mark-specific hazards* $\{A_t^F(x)\}$, $x \in E$, say, such that the interpretation

$$A^F(dt; x)=P(T\in dt, X=x|\mathcal{F}_{t-}) \tag{5}$$

is valid and

$$A_t^F(x)=\int_0^t A^F(\mathrm{dt}; x), \qquad t\geq 0. \tag{6}$$

But nothing new needs to be done to define such hazards: for the formal martingale definition all we have to do is to replace N_t in 2.1.3 by $N_t(x)$, and then require that $M_t(x)=N_t(x)-A_t^F(x)$ is an F-martingale. If $\{A_t^F(x)\}$ is absolutely continuous with

$$A_t^F(x)=\int_0^t \lambda_s^F(x)\, ds, \qquad t\geq 0, \tag{7}$$

we call $\{\lambda_t^F(x)\}$ the x-specific F-intensity of (T, X).

Example. The natural example to consider is where $F=F^N$ with $\mathcal{F}_t^N=\sigma\{(T, X); T\leq t\}$, which corresponds to observing the marked point (T, X) at T (but only knowing that $\{T>t\}$ holds at times strictly before T). It is then obvious from (2) and 2.1.3.(3) that we should have

$$A^{F^N}(dt; x)=R(dt; x)\cdot 1_{\{T\geq t\}} \tag{8}$$

where

$$R(dt; x)=P(T\in dt, X=x|T\geq t)=\frac{P(T\in dt, X=x)}{P(t\geq t)} \tag{9}$$

It is important to note that $N_t(\cdot)$ and $A_t^F(\cdot)$ are additive over the marks: if $E_0\subset E$ is an arbitrary subset of the mark space and $N_t(E_0)$ is defined as $N_t(E_0)=1_{\{T\leq t, X\in E_0\}}$, clearly $N_t(E_0)=\Sigma_{x\in E_0}N_t(x)$. But then it is obvious from the martingale definition that the F-compensator of $N_t(E_0)$ is actually $A_t^F(E_0)=\Sigma_{x\in E_0}A_t^F(x)$. (Somewhat less formally, this should be clear from the interpretation $A^F(dt; E_0)=P(T\in dt, X\in E_0|\mathcal{F}_{t-})$ and the additivity of conditional probabilities.) This shows effectively that, for fixed $t\geq 0$, both $N_t(\cdot)$ and $A_t^F(\cdot)$ can be viewed as measures on E. Even more is true: $N(dt; x)$ and $A^F(dt; x)$ are measures on $[0, \infty)\times E$.

It is sometimes useful to decompose the mark-specific hazard into the product of two factors, usually called *local characteristics*. Writing

$$A^F(dt; x)=A^F(dt)\,\frac{A^F(dt; x)}{A^F(dt)}=A^F(dt)\cdot\varphi_t(x), \tag{10}$$

say, $\varphi_t(x)$ will have the interpretation

$$\varphi_t(x)=\frac{P(T\in dt,\ X=x|\mathcal{F}_{t-})}{P(T\in dt|\mathcal{F}_{t-})}. \tag{11}$$

In other words, having observed the strict pre-t history, $\varphi_t(x)$ is the conditional probability of $\{X=x\}$ if in fact $\{T=t\}$ occurs. We shall use later a slight extension of this, writing, for $x\in E_0\subset E$,

$$A^{\mathrm{F}}(dt;\ x)=A^{\mathrm{F}}(dt;\ E_0)\frac{A^{\mathrm{F}}(\mathrm{dt};\ x)}{A^{\mathrm{F}}(dt;\ E_0)}=A^{\mathrm{F}}(dt;\ E_0)\cdot\varphi_t(x|E_0), \tag{12}$$

say. Now $\varphi_t(x|E_0)$ has the interpretation of the conditional probability of $\{X=x\}$, given that $\{T=t,\ X\in E_0\}$ occurs.

Finally, we consider the important question: do the mark-specific hazards determine (uniquely) the distribution of the marked point (T,X)? Ignoring the mark and considering only T, in the case where $\mathrm{F}=\mathrm{F}^N$, it follows at once from 2.1.3.(3) and 2.1.2.(8) that this is the case. On the other hand, in view of remark 4 in 2.1.3, if the history F is general, it may not be possible to obtain the distribution of T from (A_t^{F}). This negative result must naturally hold for marked points as well, but so does the positive one: if $\mathrm{F}=\mathrm{F}^N$, we obviously have

$$\begin{aligned}P(T\in dt,\ X=x)&=P(T\geqslant t)\cdot P(T\in dt,\ X=x|T\geqslant t)\\&=\bar{F}(t-)\cdot R(dt;x),\end{aligned} \tag{13}$$

where $\bar{F}(t-)$ is determined from 2.1.2.(8) with $R(dt)=\Sigma_{x\in E}R(dt;x)$.

2.2.2. A general marked point process

A marked point process (MPP) is a time-ordered sequence of marked points, say $(T_n,X_n)_{n\geqslant 1}$, such that $(0<)T_1<T_2<\ldots$ and $X_n\in E$ for all $n\geqslant 1$. For simplicity we consider only histories F to which the process is adapted, i.e. $\{T_n\leqslant t,\ X_n=x\}\in\mathcal{F}_t$ for all t, x and n, and assume that $T_n\to\infty$ as $n\to\infty$. In fact, in most examples we consider $T_n=\infty$ already for some finite n. Extending slightly the definition 2.2.1.(3) above, we let

$$N_t(x)=\sum_{n\geqslant 1}1_{\{T_n\leqslant t,\ X_n=x\}} \tag{1}$$

count the number of points up to time t which have mark x. Similarly, we extend the notion of internal history $\mathrm{F}^N=(\mathcal{F}_t^N)$ to be $\mathcal{F}_t^N=\sigma\{(T_n,X_n);\ T_n\leqslant t\}=\sigma\{N_s(x);\ s\leqslant t,\ x\in E\}$, corresponding to the observation of all marked points as they appear in time.

Example A (continuation from 2.2.1). We let $(T_1,X_1)=(T,X)$ as defined previously, and let $(T_2,X_2),\ldots,(T_k,X_k)$ be the subsequent part failure times and the corresponding "diagnoses". In other words e.g. Arjas (1981b), Norros (1986):

$$\begin{aligned}&T_1=\bigwedge_{i=1}^{k}S_i;\qquad X_1=i_1\quad\text{on}\quad\{T_1=S_{i_1}\}\\&T_2=\bigwedge_{\{i:\,S_i>T_1\}}S_i;\qquad X_2=i_2\quad\text{on}\quad\{T_2=S_{i_2}\}\\&\quad\ldots\\&T_k=\bigwedge_{\{i:\,S_i>T_{k-1}\}};\qquad X_k=i_k\quad\text{on}\quad\{T_k=S_{i_k}\}.\end{aligned} \tag{2}$$

Finally, as a convention, we set $T_{k+1}=T_{k+2}=\ldots=\infty$ and $X_{k+1}=X_{k+2}=\ldots=\Delta$, where Δ is a fictitious mark not in $(1, 2, \ldots, k)$.

Having assumed that the marked points are observed in the F-history, we can actually again go back to single marked points and use iteratively the corresponding definitions. Denoting

$$N_t^{(n)}(x)=1_{\{T_n\leqslant t,\, X_n=x\}}, \tag{3}$$

we obviously have from (1) that $N_t(x)=\Sigma_{n\geqslant 1}N_t^{(n)}(x)$.

Therefore the corresponding equality

$$A_t^{\mathrm{F}}(x)=\sum_{n\geqslant 1} A_t^{\mathrm{F},(n)}(x) \tag{4}$$

must hold for the F-hazards where $A_t^{\mathrm{F},(n)}(x)$ is the F-compensator of (3). But $N_t^{(n)}(x)$ can only "count" on the time interval $(T_{n-1}, T_n]$ (i.e. $N_t^{(n)}(x)\equiv 0$ for $t\leqslant T_{n-1}$, and $N_t^{(n)}(x)\equiv N_{T_n}^{(n)}(x)$ for $t>T_n$) so that the corresponding F-hazard $A^{\mathrm{F},(n)}(dt;x)$ must vanish outside that interval. We therefore see that (4) effectively gives us a piecewise definition of $A_t^{\mathrm{F}}(x)$, each $A_t^{\mathrm{F},(n)}(x)$ contributing to the cumulative hazard only on $(T_{n-1},\ T_n]$. The interpretation is now (cf. 2.2.1.(5))

$$A^{\mathrm{F}}(dt;x)=P(T_n\in dt,\, X_n=x\,|\,\mathscr{F}_{t-}) \quad \text{on} \quad \{T_{n-1}<t\leqslant T_n\}. \tag{5}$$

The case $\mathrm{F}=\mathrm{F}^N$, the internal history, is naturally of special importance. Then it will often be most convenient to work on the *canonical space* of marked points (see, for example, Jacobsen (1982)): let $\Omega=\{(t_n,x_n)_{n\geqslant 1};\ 0<t_1<t_2<\ldots,x_n\in E (n\geqslant 1)\}$, and define, for $\omega=(t_n,x_n)_{n\geqslant 1}$,

$$T_n(\omega)=t_n \quad \text{and} \quad X_n(\omega)=x_n. \tag{6}$$

We can then think of the pre-t marked points as forming a set-valued stochastic process (Norros, 1986): let *the history process* (H_t) be defined by

$$H_t=\{(T_n,\, X_n);\ T_n\leqslant t\} \tag{7}$$

(i.e. $H_t(\omega)=\{(t_n,x_n);\ t_n\leqslant t\}$). Clearly $\mathrm{F}_t^N=\sigma\{H_t\}$, which means that all $\mathscr{F}_t^N$-conditional probabilities can be expressed as functions of H_t. On the other hand, $H_t=H_{T_{n-1}}$ for all $T_{n-1}\leqslant t<T_n$, which is the same as $H_{t-}=H_{T_{n-1}}$ for all $T_{n-1}<t\leqslant T_n$. Therefore, in view of (4) and (5) above, $A^{\mathrm{F}^N}(dt;x)$ can be expressed as (cf. 2.2.1.(9))

$$\begin{aligned} A^{\mathrm{F}^N}(dt;\,x)&=\sum_{n\leqslant 1} 1_{\{T_{n-1}<t\leqslant T_n\}}\cdot R^{(n)}(dt;\ x\,|\,H_{T_{n-1}}) \\ &=R(dt;x\,|\,H_{t-}), \end{aligned} \tag{8}$$

say. Here $R^{(n)}(dt;x\,|\,H_{T_{n-1}})$ is the conditional x-specific hazard associated with (T_n, X_n), given $H_{T_{n-1}}$.

To conclude this section, consider again the question of determining the distribution of the marked points from given F-hazards. For a single marked point with $\mathrm{F}=\mathrm{F}^N$, this was done at the end of the previous section. But it is easily seen that the same result holds for an entire marked point process as long as the hazards are based on the internal history F^N of the process. In fact, we can proceed iteratively, point by point: suppose that the hazards $R^{(n)}(dt;x\,|\,H_{T_{n-1}})$ are given for all history sets $H_{T_{n-1}}$. By convention $H_{T_0}=\varnothing$, so that the distribution of (T_1, X_1) can be determined directly from 2.2.1.(13). In general, to derive the distribution of (T_n, X_n), conditionally given $H_{T_{n-1}}$, we simply replace 2.2.1.(13) by

$$P(T_n\in dt,\, X_n=x\,|\,\mathscr{F}_{T_{n-1}})\doteq \prod_{T_{n-1}<s<t}\{1-R^{(n)}(ds\,|\,H_{T_{n-1}})\}\cdot R^{(n)}(dt;\ x\,|\,H_{T_{n-1}}) \tag{9}$$

with

$$R^{(n)}(dt|H)=\sum_{x\in E} R^{(n)}(dt;\, x|H).$$

The probability on the canonical space Ω of sample paths is then defined by extension. This construction is due to Jacod (1975). We think that in most applications of point process theory the natural way to set up the probability model is by specifying the conditional hazards. This corresponds closely to specifying the transition matrix for Markov chains,or the generator for Markov processes.

3. Prediction and state estimation

3.0. Introductory comments

The notion of hazard, which was the main theme above, can be viewed as an answer, in probability terms, to the question "What is going to happen next?" In a sense, therefore, the compensator, or the intensity process, forms a series of very short-term predictions. We also saw above that if the information process (history) has a simple structure, the hazards determine the distribution of an MPP.

First, it is natural to ask whether the general idea of "a dynamic prediction which is based on a history" is somehow peculiar and applies only to hazards. The answer is "No". In this section we outline the general concept of a *prediction process*. It is a rather abstract notion: a distribution-valued stochastic process. However, it provides a very natural framework for expressing interesting structural properties which are, in fact, often encountered in practical applications.

Second, as in the case of hazard, it is important to relate the predictions to the available information. This information may not contain everything one would like to know of the past, and this brings up the question "What has happened?" A solution, again in terms of probability distributions, is provided by a *state estimation* (or filtering) formula, which has its own dynamics and is connected to Bayes theorem.

Third, it is important to relate the problems of state estimation and prediction to each other. As often in real life, one's view about the future depends heavily on what one actually knows of the past.

We illustrate all these concepts by an example, discussing a situation where uncontrolled random factors introduce an element of heterogeneity into the population, and this then becomes a source of dependence between the observations.

3.1. The general prediction process

3.1.1. The basic concepts

In this section we introduce some concepts related to prediction and discuss their meaning. At first, in order to show the generality of the ideas and to allow for more flexibility, no reference is made to point processes.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathrm{F}=(\mathcal{F}_t)$ a history (again satisfying the usual conditions). Let $(E^*, \mathcal{E}^*)$ be a (Polish) space with $\mathcal{E}^*$ the Borel sets, $Y\colon \Omega\to E^*$ a random variable, and $\mathcal{P}(E^*)$ the probability distributions on $(E^*, \mathcal{E}^*)$, with the topology of weak convergence of measures.

The following result is originally due to Aldous (1981), and in this form it appears in Norros (1985).

Proposition

There exists a (right continuous, with left limits) process $\mu^{\mathrm{F}}=(\mu_t^{\mathrm{F}})$ *which is such that, for any* $t\geq 0$, μ_t^{F} *is a regular conditional probability given* $\mathcal{F}_t$*:*

$$\mu_t^{\mathrm{F}}(B)=P(Y\in B|\mathcal{F}_t), \qquad B\in\mathcal{E}^*. \tag{1}$$

μ^{F} is unique up to indistinguishability.

Alternatively, of course, we can consider conditional expected values of the form

$$M_t^{\mathrm{F}}(f)=E\{f(Y)|\mathcal{F}_t\}, \tag{2}$$

where f: $E^*\to\mathrm{R}^1$ is a (bounded) test function. We have $\mu_t^{\mathrm{F}}(B)=M_t^{\mathrm{F}}(1_B)$, and because of the regularity of μ_t^{F}, also

$$M_t^{\mathrm{F}}(f)=\int_{E^*} f\,d\mu_t^{\mathrm{F}}. \tag{3}$$

μ^{F} is called the F-*prediction process* of Y. The corresponding expected values $\{M_t^{\mathrm{F}}(f)\}$ are readily seen to form an F-martingale: for $0<s<t$,

$$E\{M_t^{\mathrm{F}}(f)|\mathcal{F}_s\}=E[E\{f(Y)|\mathcal{F}_t\}|\mathcal{F}_s]=E\{f(Y)|\mathcal{F}_s\}=M_s^{\mathrm{F}}(f). \tag{4}$$

If $M_t^{\mathrm{F}}(f)$ is viewed as the prediction of the value of $f(Y)$, based on the information $\mathcal{F}_t$, (4) is an expression of the following property:

> "Predicting the value of a future prediction concerning $f(Y)$ is the same as predicting $f(Y)$ directly."

In the case where Y is a marked point (T, X), or more generally, a marked point process (T_n, X_n), and the internal history $\mathrm{F}=\mathrm{F}^N$ is used, the martingales $\{M_t^{\mathrm{F}}(f)\}$ admit a well-known *integral representation* e.g. Boel *et al.* (1975), Brémaud (1981). This representation has many uses, some of which we consider below, and it also serves as an introduction into the concept of *innovation*.

Consider first a single marked point (T, X) with mark space E, and a (bounded) test function f: $\mathrm{R}_+^1\times E\to\mathrm{R}^1$, so that $M_t^{\mathrm{F}}(f)=E\{f(T, X)|\mathcal{F}_t\}$. Since the internal history $\mathrm{F}=\mathrm{F}^N$ is used, the only randomness in $\{M_t^{\mathrm{F}^N}(f)\}$ is in "when T comes and what X is". Consequently we can write

$$M_t^{\mathrm{F}^N}(f)=h(t)\cdot 1_{\{t<T\}}+f(T, X)\cdot 1_{\{t\geq T\}}, \tag{5}$$

where $h(t)$ is the deterministic function

$$h(t)=E\{f(T, X)|T>t\}. \tag{6}$$

By a straightforward calculation, and considering for simplicity the absolutely continuous case, we find that (5) can be written in the integral form

$$M_t^{\mathrm{F}^N}(f)=M_0^{\mathrm{F}^N}(f)+\sum_{x\in E}\int_0^t C_s^f(x)M(ds; x). \tag{7}$$

Here

$$M(dt; x)=N(dt; x)-\lambda_t^{\mathrm{F}^N}(x)\,dt \tag{8}$$

is the fundamental martingale used in defining the mark-specific F^N-hazards, with $\lambda_t^{\mathrm{F}^N}(x)\,dt=P(T\in dt, X=x|T\geq t)\cdot 1_{\{T\geq t\}}$ (see section 2.2.1). The integrand

$$C_t^f(x)=f(t, x)-h(t) \tag{9}$$

is called *the innovation gain* corresponding to $(T, X)=(t, x)$. In other words, for times $t<T$ the updating of the prediction $M_t^{\mathbf{F}^N}(f)$ is according to

$$M_t^{\mathbf{F}^N}(f)=M_0^{\mathbf{F}^N}(f)-\sum_{x\in E}\int_0^t C_s^f(x)\lambda_s^{\mathbf{F}^N}(x)\,ds, \tag{10}$$

and at time T there is a jump (innovation) of size $C_T^f(X)$:

$$M_T^{\mathbf{F}^N}(f)=M_{T-}^{\mathbf{F}^N}(f)+C_T^f(X). \tag{11}$$

(Note that $\Sigma_{x\in E}\int_0^t C_s^f(x)N(ds;x)=C_T^f(X)\cdot 1_{\{T\leq t\}}$.) Since there is only a single marked point, the updating stops after T and $M_t^{\mathbf{F}^N}(f)$ remains at the (then known) value $M_t^{\mathbf{F}^N}(f)=f(T, X)$ (cf. (5) above). Note that (5) and (6) are equivalent to considering the prediction process

$$\mu_t^{\mathbf{F}^N}(du\times\{x\})=\begin{cases}P(T\in du,\ X=x\,|\,T>t) & \text{on}\quad \{T>t\}\\ 1_{\{T\in du,\,X=x\}} & \text{on}\quad \{T\leq t\}.\end{cases} \tag{12}$$

The integral representation (7) extends again easily to marked point processes. Consider $Y=(T_n, X_n)_{n\geq 1}$ on the canonical space Ω of sample paths, i.e. $E^*=\Omega$, a test function $f\colon \Omega\to\mathbf{R}^1$, and the internal history $\mathbf{F}^N$. First note that, since $M_{T_n}^{\mathbf{F}^N}(f)$ is a function of H_{T_n}, it can be represented in the form

$$M_{T_n}^{\mathbf{F}^N}(f)=f^{(n-1)}(T_n, X_n), \tag{13}$$

where $f^{(n-1)}$ is determined by $H_{T_{n-1}}$. Then, considering again for simplicity the absolutely continuous case and defining $C_t^f(x)$ on $\{T_{n-1}<t\leq T_n\}$ by

$$C_t^f(x)=f^{(n-1)}(t, x)-\frac{\sum_{x\in E}\int_t^\infty P(T_n\in ds,\ X_n=x\,|\,\mathcal{F}_{T_{n-1}})f^{(n-1)}(s, x)}{P(T_n>t\,|\,\mathcal{F}_{T_{n-1}})}, \tag{14}$$

we have a situation which is completely analogous to considering (9) and (6) above. A straightforward calculation gives that

$$\begin{cases} M_t^{\mathbf{F}^N}(f)=M_{T_{n-1}}^{\mathbf{F}^N}(f)-\sum_{x\in E}\int_{T_{n-1}}^t C_s^f(x)\lambda_s^{\mathbf{F}^N}(x)\,ds, & T_{n-1}<t<T_n,\\ M_{T_n}^{\mathbf{F}^N}(f)=M_{T_n-}^{\mathbf{F}^N}(f)+C_{T_n}^f(X_n). \end{cases} \tag{15}$$

Thus $C_t^f(x)$ specifies how the prediction $M_t^{\mathbf{F}^N}(f)$ is updated. In particular, at each marked point (T_n, X_n) the sudden change, innovation gain, is exactly $C_{T_n}^f(X_n)$. But (15) is only an alternative way of writing the integral representation (7).

3.1.2. Applications of the prediction process

In this section we consider simple ways to apply the above-introduced ideas on dynamic prediction and innovation. The first of such concepts concerns dependence between the part life lengths in a device.

Example A (continuation from 2.2.2). Let **S** and (T_n, X_n) be as before, but consider the special

case $k=3$. In addition, we assume that the life lengths refer to the following simple three-part device:

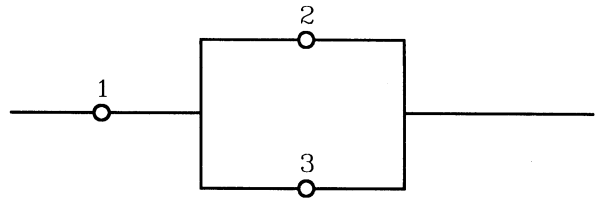


Fig. 2. The three-part device of example A.

If part 1 fails first, the device also fails. However, if part 2 fails while parts 1 and 3 are still in the working state, this is likely to increase the stress on part 3, also making it more prone to failure. A similar thing may happen to part 2 if part 3 fails first. We can now ask: can this intuitive idea be expressed in some mathematically convenient way?

There are clearly very many possibilities for such definitions of dependence. Since we cannot cover this area in any actual detail, we only give one definition which is directly based on stochastic ordering: the random vector **S** of part life lengths (or the corresponding device) is said to be *weakened by failures* with respect to F (abbreviated F^N-WBF) if, for all increasing test functions $f(\mathbf{S})$, the process $\{M_t^F(f)\}$ jumps downwards at the part failure times (Arjas & Norros, 1984; Norros, 1985). This property can also be expressed in terms of the F-prediction process: **S** is F^N-WBF if the prediction process $(\mu_t^{F^N})$ makes downward jumps, in the sense of multivariate stochastic ordering, at all part failure times.

It follows from the martingale representation result (7) that if the life lengths are F^N-WBF, then all processes $C_t^f(x)$ corresponding to increasing test functions must be negative. But this leads to the following simple observation: the process $\{M_t^{F^N}(f)\}$ must be increasing between all failure times. In a sense, therefore, the prediction concerning **S** can become worse only at failure times. For the three-part device as in example A, the sample paths $t \to M_t^{F^N}(f)$ must therefore have the following pattern:

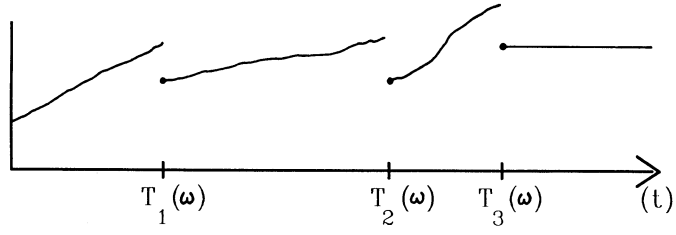


Fig. 3. A sample path of $\{M_t^{F^N}(f)\}_{t>0}$ when the three-part device is F^N-WBF.

In fact, the negativity of the innovation gains, together with some well-known elementary properties from martingale theory, can be shown to imply the following result Arjas & Norros (1984), Norros (1985).

Proposition

Suppose that the considered device is F^N-weakened by failures in the above sense. Then the part life lengths are associated, i.e. for all increasing test functions f^* and g^* of **S**,

$$Cov\{f^*(\mathbf{S}), g^*(\mathbf{S})\} \geq 0.$$

A concept slightly different from WBF, but based essentially on the same set of ideas, is the following: consider the random vector **S** of life lengths and denote by

$$a_t(\mathbf{S})=\{(S_i-t)\vee 0\}_{1\leqslant i\leqslant k} \tag{1}$$

the corresponding vector of *residual life lengths* at time t. Then we say that **S** is F-IFR if the distributions $P\{a_t(\mathbf{S})\in\cdot\,|\mathcal{F}_t\}$ are decreasing in t, in the same sense of stochastic ordering (Arjas, 1981a).

This definition has an obvious interpretation in terms of *ageing*: no matter what is observed according to F, the remaining life lengths become stochastically shorter as time advances. In the univariate case, and if $\mathrm{F}=\mathrm{F}^N$ is considered, the definition is equivalent to the classical notion of IFR.

We mention but one rather obvious result concerning this ageing property: if **S** is F^N-IFR, then it is also F^N-WBF (Norros, 1985). For a comprehensive discussion of the relationships with other multivariate definitions of ageing and dependence see Shaked & Shanthikumar (1989).

3.2. *State estimation of history*

3.2.1. *Derived point processes*

In many applications it is natural to think that the observable responses result from an underlying evolution which may not be directly observable. We start by describing a framework where both the underlying and the observed process are MPPs, and where the evolution of the underlying process fully determines the observations. Both processes are first related to the underlying process history. In section 2 we then consider the more interesting problem of estimating the underlying process from the observations.

Example A (continuation from 3.1.2). Suppose that parts 2 and 3 which are connected in parallel in the originally three-part device, are viewed together as a subunit, say $\hat{2}$. Thinking of part 1 as sub-unit $\hat{1}$, we then consider the two subunit life lengths

$$\hat{S}_{\hat{1}}=S_1 \quad \text{and} \quad \hat{S}_{\hat{2}}=S_2\vee S_3. \tag{1a}$$

The corresponding two-point "derived" MPP is naturally defined as $(\hat{T}_n, \hat{X}_n)$, $n=1, 2$, where

$$\begin{aligned} &\hat{T}_1=\hat{S}_{\hat{1}}\wedge\hat{S}_{\hat{2}}, \quad \hat{X}_1=\hat{i} \quad \text{on} \quad \{\hat{T}_1=S_{\hat{i}}\} \\ &\hat{T}_2=\hat{S}_{\hat{1}}\vee\hat{S}_{\hat{2}}, \quad \hat{X}_2=\hat{i} \quad \text{on} \quad \{\hat{T}_2=S_{\hat{i}}\} \end{aligned} \tag{1b}$$

The general definition of a *derived MPP* is most conveniently given in terms of the corresponding counting process:

Let (T_n, X_n) be an MPP with internal history F^N, and let $\hat{E}$ be a countable set with points denoted by $\hat{x}$. Then we say that an MPP $(\hat{T}_n, \hat{X}_n)$, with marks in $\hat{E}$, is derived from (T_n, X_n) if the corresponding counting process $\{\hat{N}_t(\hat{x})\}$, $\hat{x}\in\hat{E}$, can be written in the form

$$\hat{N}_t(\hat{x})=\sum_{x\in E}\int_0^t I_s(x,\hat{x})N(ds;x), \qquad t\geqslant 0. \tag{2}$$

where $\{I_t(x,\hat{x})\}$, $x\in E$, $\hat{x}\in\hat{E}$, is F^N-adapted and left-continuous and takes only values 0 and 1.

It is well known that these properties of $\{I_t(x,\hat{x})\}$ imply its F^N-predictability. The intuitive idea behind the definition is that, based on what happened in the underlying process (T_n, X_n) strictly prior to time t, a new point in that process at time t will either add or not add a point to

the derived process. It is obvious from the definition that each pre-t development of (T_n, X_n) determines that of $(\hat{T}_n, \hat{X}_n)$. More formally, denoting by $\hat{\mathbf{F}}=(\hat{\mathcal{F}}_t)$ the internal history of $(\hat{T}_n, \hat{X}_n)$, we have that

$$\hat{\mathcal{F}}_t \subset \mathcal{F}_t^N \text{ for all } \quad t \geq 0. \tag{3}$$

Still considering the history $\mathbf{F}^N$, we now see how the hazards arising from the derived process are connected with the hazards in the underlying MPP. Denoting such $\hat{x}$-specific hazards by $\hat{A}^{\mathbf{F}^N}(dt, \hat{x})$, we can reason informally as follows:

$$\begin{aligned}
\hat{A}^{\mathbf{F}^N}(dt, \hat{x}) &= E\{\hat{N}(dt; \hat{x}) | \mathcal{F}_{t-}^N\} \\
&= E\left(\sum_{x \in E} I_t(x, \hat{x}) N(dt; x) | \mathcal{F}_{t-}^N\right) \quad \text{(by (2))} \\
&= \sum_{x \in E} I_t(x, \hat{x}) E\{N(dt; x) | \mathcal{F}_{t-}^N\} \quad \text{(since } I_t(x, \hat{x}) \text{ are left continuous)} \\
&= \sum_{x \in E} I_t(x, \hat{x}) A^{\mathbf{F}^N}(dt; x).
\end{aligned} \tag{4}$$

This can easily be replaced by a formally correct martingale argument, and so we have that the $\mathbf{F}^N$-hazards of the two processes are related exactly as the counting processes in (2).

Example A (continuation). Here we have

$$\hat{N}_t(\hat{1}) = \sum_{n=1}^{2} 1_{\{\hat{T}_n \leq t, \hat{X}_n = \hat{1}\}} = 1_{\{S_1 \leq t\}} = N_t(1),$$

and

$$\begin{aligned}
\hat{N}_t(\hat{2}) &= \sum_{n=1}^{2} 1_{\{\hat{T}_n \leq t, \hat{X}_n = \hat{2}\}} \\
&= 1_{\{S_3 < S_2 \leq t\}} + 1_{\{S_2 < S_3 \leq t\}} \\
&= \int_0^t 1_{\{S_3 < s\}} N(ds; 2) + \int_0^t 1_{\{S_2 < s\}} N(ds; 3).
\end{aligned}$$

Therefore we can choose

$$I_t(1, \hat{1}) = 1, \; I_t(2, \hat{2}) = 1_{\{S_3 < t\}} \quad \text{and} \quad I_t(3, \hat{2}) = 1_{\{S_2 < t\}}.$$

The corresponding $\mathbf{F}^N$-hazards are then

$$\hat{A}^{\mathbf{F}^N}(dt; \hat{1}) = A^{\mathbf{F}^N}(dt; 1)$$

and

$$\hat{A}^{\mathbf{F}^N}(dt; \hat{2}) = 1_{\{S_3 < t\}} A^{\mathbf{F}^N}(dt; 2) + 1_{\{S_2 < t\}} A^{\mathbf{F}^N}(dt; 3). \tag{5}$$

The indicators on the right in (5) show what parts are potentially critical to subunit $\hat{2}$ in the sense that their failure causes the subunit's failure. For example, if part 3 is already down, failure of part 2 has this effect.

It is easy to see that also more general, prediction process-related properties are often inherited from the original MPP to the derived process, as long as the history is $\mathbf{F}^N$. This is

because functions of the derived process sample paths are always functions of the original process's sample paths. Supposing an obvious monotonicity in this relationship, the properties F^N-WBF and F^N-IFR, defined in the obvious way for the derived MPP, hold if they hold for the original MPP.

3.2.2. Change of history

We now study the case where the history is the one arising from the derived process $(\hat{T}_n, \hat{X}_n)$, i.e. $\hat{F}$. This corresponds to the idea that (T_n, X_n) is not directly observable, but that $(\hat{T}_n, \hat{X}_n)$ can be observed. The problem is then: what can be said about the behaviour of the underlying process, given the information in the pre-t-observations. Problems of this kind are often called *state estimation* or *filtering*.

The natural answer is again in terms of time-dependent conditional probability distributions. It is therefore closely related to, and could be derived from, the prediction processes discussed in 3.1. However, here we prefer a direct approach, and will only indicate the connection below.

Denote by H the set of possible values H which the variables H_t defined in 2.2.2.(7) could assume, i.e. H consists of finite sets of the form $\{(t_i, x_i); 1 \leqslant i \leqslant k\}$, where no two times t_i are the same and $x_i \in E$. H is endowed with the natural topology and Borel σ-fields $\mathscr{H}$. The questions of interest, as formulated above, can now be expressed in terms of conditional probabilities of $\{H_t \in B\}$, with $B \in \mathscr{H}$. Conditioning on $\mathscr{F}_t^N$ determines of course H_t exactly, i.e.

$$\pi_t(B) \doteqdot P(H_t \in B | \mathscr{F}_t^N) = 1_B(H_t), \qquad B \in \mathscr{H},\ t \geqslant 0. \tag{1}$$

Our interest therefore focuses on the corresponding $\hat{F}$-conditional distribution

$$\hat{\pi}_t(B) \doteqdot P(H_t \in B | \hat{\mathscr{F}}_t) = E\{\pi_t(B) | \hat{\mathscr{F}}_t\}, \qquad B \in \mathscr{H},\ t \geqslant 0. \tag{2}$$

It follows from the fact that $\{\pi_t(B)\}$ is piecewise constant for fixed $B \in \mathscr{H}$ that $\hat{\pi}_t(B)$ admits a right continuous version with left limits, and this property can be extended to the probability measure valued processes $(\hat{\pi}_t)$ (with respect to weak convergence). Thus the situation is completely analogous to the one in 3.1 where the prediction process was discussed. In fact, we can obviously write

$$\hat{\pi}_t(B) = \mu_t^{\hat{F}} 0 H_t^{-1}(B), \qquad B \in \mathscr{H} \tag{3}$$

where $\mu^{\hat{F}}$ is the $\hat{F}$-prediction process corresponding to the sample paths $Y = (T_n, X_n)_{n \geqslant 1}$ of the underlying MPP.

We now show how the distributions $\hat{\pi}_t$ can be used as links between various F^N- and $\hat{F}$-related quantities. Starting with the hazards, we first proceed informally and write

$$\begin{aligned} \hat{A}^{\hat{F}}(dt, \hat{x}) &= E\{\hat{N}(dt; \hat{x}) | \hat{\mathscr{F}}_{t-}\} \\ &= E[E\{\hat{N}(dt; \hat{x}) | \mathscr{F}_{t-}^N\} \hat{\mathscr{F}}_{t-}] \\ &= E\{A^{F^N}(dt; \hat{x}) | \hat{\mathscr{F}}_{t-}\}. \end{aligned} \tag{4}$$

This suggests that the $\hat{F}$-hazards could actually be obtained as averages of the corresponding F^N-hazards, with respect to the distributions $(\hat{\pi}_{t-})$. This is in fact correct: with the notation as in 2.2.2.(8) and 3.2.1.(4), and writing $I_t(x, \hat{x}) = I_t^*(x, \hat{x} | H_{t-})$, we have that

$$\begin{aligned} \hat{A}^{F^N}(dt; \hat{x}) &= R(dt; \hat{x} | H_{t-}) \\ &= \sum_{x \in E} I_t^*(x, \hat{x} | H_{t-}) R(dt; x | H_{t-}). \end{aligned} \tag{5}$$

Then the exact statement corresponding to (4) is the equality

$$\hat{A}^{\hat{\mathrm{F}}}(dt, \hat{x})=\int_{\mathrm{H}} \hat{\pi}_{t-}(dH)R(dt, \hat{x}|H). \tag{6}$$

A similar relationship holds between the prediction processes μ^{F^N} and $\mu^{\hat{\mathrm{F}}}$. Considering $\mu_t^{\mathrm{F}^N}$ as a transition kernel P_t: $\mathscr{H} \to E^*$ from pre-t F^N-histories to values of Y,

$$\mu_t^{\mathrm{F}^N}(B)=P_t^*(B|H_t), \qquad B \in \mathscr{E}^*, \tag{7}$$

we have that

$$\mu_t^{\hat{\mathrm{F}}}(B)=\int_{\mathrm{H}} \hat{\pi}_t(dH)P_t^*(B|H), \qquad B \in \mathscr{E}^*. \tag{8}$$

This is easily verified by using Fubini's theorem.

There are essentially two approaches for determining the distributions $\hat{\pi}_t$ explicitly. The direct approach is to fix t and consider $\hat{\pi}_t$ as a conditional distribution on $(\mathrm{H}, \mathscr{H})$, with respect to $P_t=P|\mathscr{F}_t^N$, the restriction of P to $\mathscr{F}_t^N$. This is again best illustrated by an example.

Example A (continuation from 3.2.1). Assume that the part life lengths are independent and exponential with parameters ϱ_i, $i=1, 2, 3$. Then, for example, on $\{t<\hat{S}_2\}$

$$\begin{aligned} P(S_3 \leqslant t|\hat{\mathscr{F}}_t)&=\frac{\exp(-\varrho_2 t)\{1-\exp(-\varrho_3 t)\}\cdot 1_{\{t<\hat{S}_2\}}}{1-\{1-\exp(-\varrho_2 t)\}\{1-\exp(-\varrho_3 t)\}} \\ &=\hat{\pi}_t(B), \end{aligned}$$

where $B=\{H \in \mathrm{H}: (s,3) \in H \text{ for some } 0<s\leqslant t\}$.

The other approach for determining $\hat{\pi}_t$ is *dynamic*, showing how $(\hat{\pi}_t)$ is updated as the observation time changes. In fact, it can be shown by a straightforward application of the general filtering formula e.g. Brémaud & Jacod (1977), Brémaud (1981), Koch (1986) that the following holds:

Proposition

Assuming absolute continuity of the hazards, we have that for all $B \in \mathscr{H}$

$$\begin{aligned} \hat{\pi}_t(B)=&\hat{\pi}_0(B) \\ &+\int_0^t \left(\sum_{x \in E} \int \hat{\pi}_{s-}(dH)r_s(x|H)[1_B(H \cup \{(s, x)\})-1_B H)]\right) ds \\ &+\sum_{\hat{x} \in \hat{E}} \int_0^t \left(\frac{\sum_{x \in E} \hat{\pi}_{s-}(dH)I_s^*(x, \hat{x}|H)r_s(x|H)1_B(H \cup \{(s, x)\})}{\hat{\lambda}_s^{\hat{\mathrm{F}}}(\hat{x})}-\hat{\pi}_{s-}(B)\right) \\ &\times\{\hat{N}(ds, \hat{x})-\hat{\lambda}_s^{\hat{\mathrm{F}}}(\hat{x})\, ds\}, \end{aligned} \tag{9}$$

where the notation is as in (5) *and* (6) *with* $R(dt, x|H)=r_t(x|H)\, dt$ *and* $\hat{A}^{\hat{\mathrm{F}}}(dt, \hat{x})=\hat{\lambda}_t^{\hat{\mathrm{F}}}(\hat{x})\, dt$.

The proof is given in Arjas & Norros (1988). The second integral is an $\hat{\mathrm{F}}$-martingale with a similar structure as in 3.1.1.(7) but with the innovation gain corresponding to the observation of $\hat{\mathrm{F}}$. It can also be viewed as a continuous time version of Bayes' formula.

It was mentioned at the end of section 3.2.1 that many interesting qualitative properties of the underlying point process are inherited by the derived process, as long as the conditioning

history is that of the underlying process. If the latter is not observed, as is commonly the case, we can ask whether the corresponding properties still hold, but now with respect to the coarser history. Unfortunately, the averaging operation just discussed tends to ruin such useful properties as monotonicity. A classical example of this is the above two-part parallel subunit with independent exponential life lengths (e.g. Barlow & Proschan, 1975). $\hat{T}_2$ defined in 3.2.1.(1b) is not IFR with respect to its internal history $\sigma\{\hat{T}_2; \hat{T}_2 \leqslant t\}$, $t \geqslant 0$. The $\hat{\mathrm{F}}$-hazard rate of this subunit is given by

$$\hat{\lambda}_t(\hat{2}) = \frac{\varrho_2 \exp(-\varrho_2 t)\{1-\exp(-\varrho_3 t)\} + \varrho_3 \exp(-\varrho_3 t)\{1-\exp(-\varrho_2 t)\}}{1-\{1-\exp(-\varrho_2 t)\}\{1-\exp(-\varrho_3 t)\}} \cdot 1_{\{t \leqslant \hat{S}_2\}}. \tag{10}$$

This can be viewed as a weighted average of ϱ_2, ϱ_3 and 0, with the weight of 0 tending rapidly to zero but the weight of $\varrho_2 \wedge \varrho_3$, corresponding to the "safer" of the two parts, approaching one as $t \to \infty$ if there is no failure. Indeed, we see from 3.2.1.(5) that (10) is the same as

$$\hat{\lambda}_t(\hat{2}) = \varrho_2 P(S_3 < t \mid \hat{\mathscr{F}}_{t-}) + \varrho_3 P(S_2 < t \mid \hat{\mathscr{F}}_{t-}), \tag{11}$$

which also agrees with (4) in this special case.

3.2.3. An example: accounting for heterogeneity

Consider two parts with respective life lengths S_1 and S_2. Suppose that there is some reason to believe that S_1 and S_2 are not independent. Perhaps the parts were manufactured under similar but randomly varying conditions, or they are tested in the same random environment. In either case, both parts tend to be either more, or less, prone to failure than "normal", depending on what the random conditions are. In the same way, other parts manufactured or tested under different conditions would be likely to have a different propensity to fail. This kind of variation is often called *heterogeneity*.

Under such circumstances, one can attempt to model the random conditions in an explicit manner. The simplest alternative is to use a random proportionality factor, say Z, which remains fixed over time and whose values reflect some particular conditions. However, Z will usually not be an actual physical quantity: it may be difficult to give it a meaning outside the model, and its values cannot be observed.

Thus we are led to consider the following simple model. Let (T_n, X_n) be the two-point MPP corresponding to S_1 and S_2 as in the previous examples, and consider the history $\mathrm{F}^* = (\mathscr{F}_t^*)$, where $\mathscr{F}_t^* = \sigma\{Z\} \vee \sigma\{(T_n, X_n);\ T_n \leqslant t\}$. (We could think of Z as a random mark at the origin, but this interpretation is not important.) We then set up the probability model by assuming that Z has given distribution Ψ, and that the two part life lengths have F^*-hazard rates

$$\lambda_t^{\mathrm{F}^*}(i) = Z \cdot a_t^*(i) \cdot 1_{\{S_i \geqslant t\}}, \qquad t \geqslant 0,\ i = 1, 2, \tag{1}$$

where $a_t^*(1)$ and $a_t^*(2)$ are given functions.

It is obvious from this definition that, since the functions $a_t^*(i)$ are fixed, S_1 and S_2 are conditionally independent given Z. However, this independence does not hold unconditionally. Let us now see how the distribution of Z changes in time when the two-point MPP is observed. It is clear intuitively that time spent without failures causes the estimate of Z to go down, whereas a failure gives a sudden increase (innovation). More exactly, consider $\mathrm{F}^N = \sigma\{(T_i, X_i):\ T_i \leqslant t,\ i = 1, 2\}$ and the conditional distributions

$$\pi_t^*(\cdot) = P(Z \in \cdot \mid \mathscr{F}_t^N).$$

A simple calculation shows, denoting $a_t^*=a_t^*(1)+a_t^*(2)$, that on $\{t<T_1\}$

$$\pi_t^*(dz)=\frac{\Psi(dz)\exp\left(-z\int_0^t a_s^*\,ds\right)}{\int_0^\infty \Psi(dz)\exp\left(-z\int_0^t a_s^*\,ds\right)}, \tag{3a}$$

on $\{T_1\leqslant t<T_2\}$

$$\pi_t^*(dz)=\frac{\Psi(dz)z\exp\left(-z\int_0^{T_1} a_s^*\,ds-z\int_{T_1}^t a_s^*(3-X_1)\,ds\right)}{\int_0^\infty \Psi(dz)z\exp\left(-z\int_0^{T_1} a_s^*\,ds-z\int_{T_1}^t a_s^*(3-X_1)\,ds\right)}, \tag{3b}$$

and on $\{T_2\leqslant t\}$

$$\pi_t^*(dz)=\frac{\Psi(dz)z^2\exp\left(-z\int_0^{T_1} a_s^*\,ds-z\int_{T_1}^{T_2} a_s^*(X_2)\,ds\right)}{\int_0^\infty \Psi(dz)z^2\exp\left(-z\int_0^{T_1} a_s^*\,ds-z\int_{T_1}^{T_2} a_s^*(X_2)\,ds\right)}. \tag{3c}$$

The corresponding F^N-intensities are then given by (cf. 3.2.2.(6))

$$\lambda_t^{\mathrm{F}^N}(i)=a_t^*(i)\int_0^\infty z\pi_{t-}^*(dz)\cdot 1_{\{S_i\geqslant t\}},\qquad t\geqslant 0. \tag{4}$$

Let us then consider the situation where only one of the two parts is observed, say part 1. Thus we consider the single-point derived point process

$$\hat{T}=S_1, \tag{5a}$$

which is the same as

$$\hat{T}=\begin{cases}T_1 & \text{if } X_1=1\\ T_2 & \text{otherwise.}\end{cases} \tag{5b}$$

Let also $\hat{\mathrm{F}}=(\hat{\mathscr{F}}_t)$ be defined by $\hat{\mathscr{F}}_t=\sigma\{\hat{T};\ \hat{T}\leqslant t\}$, and denote

$$\hat{\pi}_t(\cdot)=P(Z\in\cdot\,|\hat{\mathscr{F}}_t). \tag{6}$$

There are two ways to proceed in order to derive $(\hat{\pi}_t)$ and the $\hat{\mathrm{F}}$-intensity for part 1. One is to start from (4) and do one more averaging, now corresponding to the fact that part 2 is no longer observed. However, it is much simpler to go back to (1) and, using the conditional independence of the two-part life lengths given Z, simply ignore part 2. Let $\hat{\mathrm{F}}^*=(\hat{\mathscr{F}}_t^*)$ with $\hat{\mathscr{F}}_t^*=\sigma\{Z\}\vee\hat{\mathscr{F}}_t$. We then find that

$$\begin{aligned}\lambda_t^{\hat{\mathrm{F}}^*}(1)&=Z\cdot a_t^*(1)\cdot 1_{\{S_1\geqslant t\}}\\ &=\lambda_t^{\mathrm{F}^*}(1),\qquad t\geqslant 0,\end{aligned} \tag{7}$$

i.e. the $\hat{\mathrm{F}}^*$- and F^*-intensities for S_1 are the same. The conditional distribution of Z when only $\hat{\mathrm{F}}$

is observed is easily found to be

$$\hat{\pi}_t(dz)=\frac{\Psi(dz)\,\exp\left(-z\int_0^t a_s^*(1)\,ds\right)}{\int_0^\infty \Psi(dz)\,\exp\left(-z\int_0^t a_s^*(1)\,ds\right)}\cdot 1_{\{t<\hat{T}\}}$$

$$+\frac{\Psi(dz)z\,\exp\left(-z\int_0^{\hat{T}} a_s^*(1)\,ds\right)}{\int_0^\infty \Psi(dz)z\,\exp\left(-z\int_0^{\hat{T}} a_s^*(1)\,ds\right)}\cdot 1_{\{t\geq\hat{T}\}}. \tag{8}$$

The corresponding $\hat{\mathrm{F}}$-intensity is then

$$\lambda_t^{\hat{\mathrm{F}}}(1)=a_t^*(1)\int_0^\infty z\hat{\pi}_{t-}(dz)\cdot 1_{\{t\leq\hat{T}\}}. \tag{9}$$

Note that $\hat{\mathrm{F}}$ is the internal history of the "single-point derived process". Therefore the unconditional survival probability of part 1 can be obtained from the exponential formula

$$P(S_1>t)=\exp\left(-\int_0^t \lambda_t^{\hat{\mathrm{F}}}(1)\,ds\right), \tag{10}$$

where $\lambda_t^{\hat{\mathrm{F}}}(1)$ is determined on $\{t\leq\hat{T}\}$.

The important thing to observe after these explicit but somewhat tedious-looking calculations is that the F^N-intensity and the $\hat{\mathrm{F}}$-intensity of part 1, appearing in (4) and (9) respectively, are different. This is because the additional observation of part 2 in F^N gives indirect information about the unobserved variable Z, which in turn affects the "original" hazards (1). This effect is particularly important in cases where part 2 also acts as a *censoring mechanism*. Then the full observation of $\hat{\mathrm{F}}$ can be impossible because observation always terminates at $T_1=S_1\wedge S_2$.

In order to get a rough idea about the size of this effect, consider the *special case*:

$$\begin{aligned}&\Psi=\text{unit exponential distribution}\\ &a_t^*(i)=a_i,\qquad i=1,\,2\ \text{(constants)}.\end{aligned} \tag{11}$$

Then one finds easily from (8) and (9) that for $t<\hat{T}$

$$\hat{\pi}_t(dz)=(1+a_1t)\,\exp\{-(1+a_1t)z\}\,dz \tag{12}$$

and

$$\lambda_t^{\hat{\mathrm{F}}}(1)=\frac{a_1}{1+a_1t},$$

and therefore by (10)

$$P(\hat{T}>t)=P(S_1>t)=\frac{1}{1+a_1t}.$$

Similar calculations based on (3) and (4) show that

$$\lambda_t^{F^N}(1)=\frac{a_1\cdot 1_{\{t<T_1\}}}{1+(a_1+a_2)t}+\frac{2a_1\cdot 1_{\{T_1\le t<T_2,\ X_1=2\}}}{1+(a_1+a_2)T_1+a_1(t-T_1)}. \tag{13}$$

The difference between (12) and (13) is shown graphically in Fig. 4.

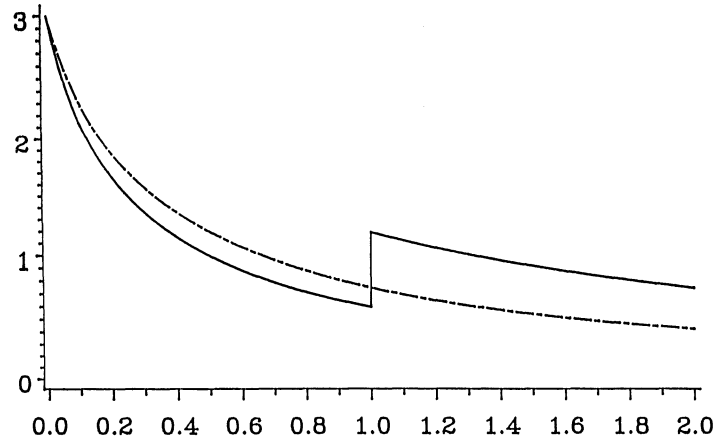


Fig. 4. The intensities $\{\lambda_t^{F^N}(1)\}$ (solid line) and $\{\lambda_t^{\hat{F}}(1)\}$ (broken line) when $a_1=3$, $a_2=1$ and $S_2=1$.

3.2.4. State estimation and unobservables: a discussion

We have seen in the above examples from reliability theory how the essentially Bayesian updating mechanism changes "a coherent observer's view" about what happened in the past, and thereby also modifies the prediction concerning the future in a way which corresponds to averaging with respect to time-dependent weights. This can be confusing, because often important qualitative properties of the model, which hold when there is more complete information of the past, are then lost.

This problem arises in practical statistical applications quite often (see, for example, Manton *et al.*, 1981; Aalen, 1987a, b, 1988; Heckman & Singer, 1984; Hougaard, 1984, 1986, 1987; Schumacher *et al.*, 1987). As an example, consider a human sample where the individuals have certain fixed characteristics which have not been observed. Again, during the course of the follow-up, the estimates of the true values of these characteristics change. In other words, although an individual under observation remains the same, the observer's estimate about his or her characteristics changes. This is not "wrong"; often there is actually a *selection mechanism* in the studied population which corresponds to the updating. For example, "frail" individuals tend to die earlier than "sturdy". Thus, at later age, the sturdy ones tend to be left. This does not mean that ageing would make an individual healthier. It is merely a reflection of what is known to the observer about the individual in question.

It can be very tempting to formulate unobservable variables explicitly. The reason is that often such variables can capture some of the variation between individuals which otherwise remains uncontrolled. Making such control explicit can be a considerable aid in interpreting the results from a statistical analysis, by "bringing back" qualitative properties, such as an individual's ageing, which were perhaps hidden by a selection mechanism.

The added control is particularly helpful if one tries to find support from observational data to a causality claim. For an interesting recent discussion of causality in dynamic statistical

models see Aalen (1987c) and its references, in particular Schweder (1970). Using the terminology in, for example, Suppes (1970), the control may help the investigator distinguish between a genuine and a spurious cause. (See also Holland (1986), and its discussion.) In epidemiology, it can help in identifying a confounding variable (see, for example, Miettinen & Cook, 1981).

Sometimes in causality reasoning the role of *cause* is given to an event which actually didn't occur when the data were collected. Future policy changes in the society, or new legislation, are examples of this. So is the introduction of a hypothetical, fully efficient treatment which is assumed to eliminate a particular cause of death. This is the classical problem in competing risks theory. Then, it is very natural to assume that some individuals are more "frail" than others, and that the frail ones, even if one potential cause of death was removed, are likely to die more easily from the remaining causes. Apparently the simplest way to model such frailty is to treat it as an unobservable proportionality factor Z which depends on the individual, exactly as in 3.2.3.(1). By interpreting part 1 failure as death from the considered cause and part 2 failure as death from all other possible causes, the example in 3.2.3 then applies without any further change. For a more elaborate model for frailty see Yashin *et al.* (1986).

These examples bring up, once more, the question about the existence of a "correct" level of information in the conditioning. We iterate our position that there is no such uniquely correct level. The ultimate, and certainly "correct", heterogeneity is that all individuals are different. For life lengths, one could say that the value of T itself, or the element ω in the underlying space Ω, represents heterogeneity. Although such claims cannot be denied as being untrue, they are not likely to lead to useful models and interpretation. Clearly, detailed information is always a good starting point. Depending on what goals one has, it is sometimes convenient to ignore some of this information, but sometimes one needs to "hypothesize" more in order to obtain results with meaningful and interesting interpretations. In the latter case, however, the variables introduced have rarely a precise meaning outside the postulated model framework.

4. Likelihood-based inference for MPPs

4.1. Introduction

So far we have discussed only probabilistic modelling aspects related to marked point processes. In the background there has been, of course, a need to set up a convenient and flexible framework for describing the complex evolutions which are often encountered in longitudinal observational studies. It is particularly important to be able to handle censoring, or more generally, the fact that individuals are followed for different periods of time.

Here we finally touch upon questions regarding statistical inference. We do not claim that the likelihood function is everything that is of interest in parametric statistical inference, but that is how far we go.

The time aspect is again crucial (cf. Arjas, 1985). In fact, we view the statistician as "living through the observation interval", accumulating data progressively as they become available through observation, and forming the corresponding likelihood expression essentially by applying the chain rule of conditional probabilities. In this way the observation interval is split into "infinitesimal experiments", and this makes it possible to handle very easily the individuals belonging, or not belonging, to the risk set.

The other aspect which we stress is modelling economy. It would often be an overwhelming task to set up a statistical model for an entire MPP sample path, describing all data one has. Thus we aim at what can be called *partial model specification*. The role of innovations which

was discussed in the previous section comes up here again; it is really the innovation gains that form the likelihood expression.

4.2. The product form of the likelihood

The common product form of the likelihood expression in statistics is the consequence of an independence assumption between the observations in the sample. In survival analysis and, more generally, in longitudinal observational studies, the observed individuals are commonly assumed to behave independently, and this makes the likelihood to be a product over the set of individuals. However, the likelihood associated with observations arising from an MPP model is also a product in a much more interesting sense: over time.

In order to provide a simple introduction to this idea, we first consider a discrete time MPP with $t \in \{1, 2, \ldots\}$. Let $\{N_t(x); t=1, 2, \ldots, x \in E\}$ be the corresponding counting processes and $\mathbb{F}^N$ the internal history. (Since only $\mathbb{F}^N$ is used in this section, we no longer indicate the history explicitly as a superscript in our notation.) Writing $N_t = \Sigma_{x \in E} N_t(x)$ and

$$p_t(x) = P\{\Delta N_t(x) = 1 | \mathcal{F}_{t-1}^N\}$$

$$p_t = P(\Delta N_t = 1 | \mathcal{F}_{t-1}^N) = \sum_{x \in E} p_t(x), \tag{1}$$

we see easily that the likelihood corresponding to observed pre-t history must have the form

$$L_t = \prod_{s \leq t} p_s^*, \tag{2}$$

where

$$p_s^* = \prod_{x \in E} p_s(x)^{\Delta N_s(x)} \cdot (1-p_s)^{1-\Delta N_s}. \tag{3}$$

The product form of (2) is a simple consequence of the chain rule of conditional probabilities, applied sequentially at times $s=1, 2, \ldots, t$, and the fact that the internal history $\mathbb{F}^N$ is used. Also note that always exactly one of the terms on the right of (3) differs from one: if $s=T_n$ for some n, i.e. there is a "point" at s, then this term is $p_{T_n}(X_n)$, and if $s \neq T_n$ for all n then the term is $1-p_s$. Thus (3) is really the familiar likelihood expression corresponding to a single multinomial trial, apart from the fact that the probabilities p_s depend on the pre-s history.

Let us then form the corresponding likelihood expression in continuous time. We do not aim at a rigorous derivation here, but recall from 2.2.1 that $A(dt; x) = P\{N(dt; x) = 1 | \mathcal{F}_{t-1}^N\}$. Thus $A(dt; x)$ has a meaning completely analogous to $p_t(x)$ in (1). This suggests that the general likelihood expression for pre-t observations should be a product of terms of the form

$$\prod_{x \in E} A(ds; x)^{N(ds;x)} \cdot \{1 - A(ds)\}^{1-N(ds)}. \tag{4}$$

On the other hand, since the product is over continuous time, there is the technical problem of how such an infinite product should be defined. We split this problem into two parts, corresponding to the two factors in (4).

The first factor is actually very easy to handle since, for any sample path, $\Pi_{x \in E} A(ds; x)^{N(ds;x)} \neq 1$ only if $s=T_n$ for some n. (Otherwise the exponent $N(ds; x) = \Delta N_s(x) = 0$ for all x.) Therefore, the obvious way to define the first factor is to write

$$\prod_{s \leq t} \prod_{x \in E} A_s(ds; x)^{N(ds;x)} = \prod_{T_n \leq t} A(dT_n, X_n). \tag{5}$$

For the second factor, we use as a guideline the discussion at the end of section 2.1.2. Expressing this factor as a product integral, we take the infinite product to mean

$$\prod_{s\leq t}\{1-A(ds)\}^{1-N(ds)}=\exp\,(-A_t^c)\cdot\prod_{s\leq t}(1-\Delta A_s)^{1-\Delta N_s} \tag{6}$$

(see 2.1.2.(8a, b), or, for more details, Gill & Johansen (1989) and Andersen *et al.* (1988)). Here ΔA_s is again a notation for the jumps ("instantaneous hazards") $\Delta A_s=A_s-A_{s-}$, and $A_t^c=A_t-\Sigma_{s\leq t}\Delta A_s$. Recall that there are always at most a countable number of jump points. The combination of (5) and (6) gives now the general likelihood expression (Jacod, 1975) corresponding to pre-t observations:

$$L_t=\prod_{T_n\leq t}A(dT_n;\,X_n)\cdot\prod_{s\leq t}(1-\Delta A_s)^{1-\Delta N_s}\cdot\exp\,(-A_t^c). \tag{7}$$

In most statistical applications the hazards are either purely discrete or purely continuous. In the former case the exponential part in (7) is equal to one, and the rest is exactly as (2) and (3) above. In the *absolutely continuous case* the middle factor in (7) is equal to one. Expressing the first and the third factor in a density form, and ignoring the time differentials, we then obtain the likelihood expression

$$\bar{L}_t=\prod_{T_n\leq t}\lambda_{T_n}(X_n)\cdot\exp\left(-\int_0^t\lambda_s\,ds\right). \tag{8}$$

We now make some comments regarding these expressions.

(i) Apart from the marks X_n, $\bar{L}_t$ in (8) has the well-known form of a Poisson-likelihood.

(ii) For a random sample $\{S_1, S_2, \ldots, S_n\}$ of i.i.d. life lengths, say, from a distribution with density f, the likelihood is clearly $\Pi_{1\leq i\leq k}f(S_i)$. That this is indeed the same as (8) for $t\geq\max(S_1,\ldots,S_k)$ is seen as follows: let the MPP (T_n, X_n) be defined as in 2.2.2.(2). Using the hazard rate notation and the exponential formula we have that $f(S_i)=r(S_i)\exp\{-\int_0^{S_i}r(s)\,ds\}$. On the other hand, the F^N-intensity for the ith individual is clearly

$$\lambda_t(i)=r(t)\cdot 1_{\{t\leq S_i\}}, \tag{9}$$

and $\lambda_t=\Sigma_{i=1}^k\lambda_t(i)=r(s)\Sigma_{i=1}^k 1_{\{t\leq S_i\}}$. Therefore,

$$\begin{aligned}\prod_{1\leq i\leq k}f(S_i)&=\prod_{1\leq i\leq k}\left\{r(S_i)\,\exp\left(-\int_0^{S_i}r(s)\,ds\right)\right\}\\&=\prod_{1\leq i\leq k}\lambda_{T_i}(X_i)\cdot\exp\left(-\int_0^{\infty}r(s)\sum_{i=1}^{k}1_{\{s\leq S_i\}}\,ds\right)\\&=\prod_{1\leq i\leq k}\lambda_{T_i}(X_i)\cdot\exp\left(-\int_0^{\infty}\lambda_s\,ds\right).\end{aligned} \tag{10}$$

(The number $\Sigma_{i=1}^k 1_{\{t\leq S_i\}}$ is usually called the size of the risk set at time t.)

(iii) The reasoning in (ii) extends easily to non-identically distributed and censored observations. Suppose that the data consist of $\{(S_i^*, \delta_i);\ 1\leq i\leq k\}$, where

S_i^*=the time (or age) at which the ith individual was last observed, and

$$\delta_i=\begin{cases}1 & \text{if the } i\text{th individual failed at } S_i^*\\ 0 & \text{if the } i\text{th individual was censored at } S_i^*.\end{cases} \tag{11}$$

In order to form a corresponding MPP, we let $\{T_n;\ 1\leqslant n\leqslant k\}$ be the order statistics of $\{S_i^*;\ 1\leqslant i\leqslant k\}$, and define the corresponding mark as

$$X_n=(i,\ \delta_i) \quad \text{on} \quad \{T_n=S_i^*\}. \tag{12}$$

In other words, X_n indexes the individual which was last observed at the nth smallest of the times $\{S_1^*, \ldots, S_k^*\}$, and also indicates whether the individual then failed or was censored. Considering for simplicity the case where all mark-specific hazards are absolutely continuous, we assume that the failure intensities are of the form

$$\lambda_t(i,\ 1)=r_i(t)\cdot 1_{\{t\leqslant S_i^*\}}, \qquad 1\leqslant i\leqslant k, \tag{13a}$$

and the censoring intensities similarly of the form

$$\lambda_t(i,\ 0)=c_i(t)\cdot 1_{\{t\leqslant S_i^*\}}, \qquad 1\leqslant i\leqslant k. \tag{13b}$$

Then the likelihood expression (8) corresponding to the observed data is seen to factor as follows:

$$\begin{aligned}\bar{L}_\infty &= \prod_n \lambda_{T_n}(X_n)\cdot\exp\left(-\int_0^\infty \lambda_s\, ds\right)\\ &= \prod_{\{i:\,\delta_i=1\}} \lambda_{S_i^*}(i,\ 1)\cdot \prod_{\{i:\,\delta_i=0\}} \lambda_{S_i^*}(i,\ 0)\cdot\exp\left(-\int_0^\infty \sum_{i=1}^k \{\lambda_s(i,\ 1)+\lambda_s(i,\ 0)\}\, ds\right)\\ &= \prod_{\{i:\,\delta_i=1\}} \left\{r_i(S_i^*)\cdot\exp\left(-\int_0^{S_i^*} r_i(s)\, ds\right)\right\}\\ &\quad \prod_{\{i:\,\delta_i=0\}} \exp\left(-\int_0^{S_i^*} r_i(s)\, ds\right)\cdot\prod_{i=1}^k \left\{c_i(S_i^*)^{1-\delta_i}\exp\left(-\int_0^{S_i^*} c_i(s)\, ds\right)\right\}.\end{aligned} \tag{14}$$

Denoting the distribution which corresponds to r_i by F_i, with density f_i and survival function $\bar{F}_i$, the first two factors on the right-hand side of (14) can be written as

$$\prod_{\{i:\,\delta_i=1\}} f_i(S_i^*)\cdot \prod_{\{i:\,\delta_i=0\}} \bar{F}_i(S_i^*). \tag{15}$$

This is the usual likelihood expression for censored survival data, which appears in the literature. The third factor in (14), which is then omitted, may or may not depend on the parameter of interest in the statistical model. If it does, (15) is called a *partial likelihood*. We return to this question in more detail in section 3 below.

(iv) The picture concerning the structure of the likelihood expression is complemented by briefly considering the likelihood ratio between two distributions. Let P and P' be such distributions, denote their restrictions to $\mathcal{F}_t^N$ respectively by P_t and P_t', and suppose again for simplicity that the hazards are absolutely continuous. It then follows from (8) that

$$L_t^*=\frac{dP_t'}{dP_t}=\prod_{T_n\leqslant t}\frac{\lambda'_{T_n}(X_n)}{\lambda_{T_n}(X_n)}\exp\left(-\int_0^t(\lambda_s'-\lambda_s)\, ds\right). \tag{16}$$

The evolution of the likelihood ratio, as a stochastic process, is therefore directly dependent on the comparison between $\lambda_{T_n}(X_n)$ and $\lambda'_{T_n}(X_n)$ at the marked points (T_n, X_n), and otherwise between the "crude" intensities λ_s and λ_s'.

It is a well-known analytic result that (16) is the solution of the following integral equation:

$$L_t^* = L_0^* + \sum_{x \in E} \int_0^t L_{s-}^* \left(\frac{\lambda_s'(x)}{\lambda_s(x)} - 1 \right) \{ N(ds;x) - \lambda_s(x)\, ds \} \tag{17}$$

(see, for example, Brémaud, 1981). This representation is, however, interesting in its own right, and it is quite analogous to formula 3.1.1.(7). In fact, considering left-continuous versions of the intensities $\lambda_t(x)$ and $\lambda_t'(x)$, the integrand is left-continuous, and the integration is with respect to the "fundamental" F^N-martingales $M_t(x) = N_t(x) - \int_0^t \lambda_s(x)\, ds$ (with respect to P). These facts are well known to imply that (L_t^*) is itself an F^N-martingale. (Remark: compare this with the classical result in statistics which says, using the obvious notation, that

$$E_{\theta_0}\left(\frac{f_\theta(X)}{f_{\theta_0}(X)} \right) = 1.$$

This identity is really the simple random sample analogue of the martingale property of (L_t^*). The representation (17) also illustrates nicely how the innovation gain process delivers the correct updates

$$L_{T_n-}^* \cdot \frac{\lambda_{T_n}'(X_n)}{\lambda_{T_n}(X_n)} - L_{T_n-}^*$$

to the likelihood ratio L^* at the marked points (T_n, X_n) (cf. 3.1.1.(15)).

4.3. Innovations, non-innovations and partial model specification

We now study how the general structure of the likelihood expressions, which was considered above, relates to questions arising from parametric statistical modelling of observational data. The presentation here is a somewhat simplified version of Arjas & Haara (1984). Andersen *et al.* (1988) provides a very readable account into these ideas, among other things.

Suppose that the statistical model of the considered MPP is parameterized by (θ, ψ), where

- θ is viewed as the *parameter of interest*, typically parametrizing the intensity of a "response" (such as a failure), and
- ψ is viewed as a *nuisance parameter* which is needed, to complement θ, in order to parameterize the intensities of all other marked points in the data.

Obviously this kind of division of roles depends very much on what is "of interest", and what level of information is used in the conditioning. Typically one would let ψ parameterize censoring and possible evolution of time-dependent covariates.

The calculation in (iii) of the previous section now serves as a guideline to what follows. Our goal is to separate from the likelihood (or likelihood ratio) those parts which do not depend on θ, the parameter of interest. In likelihood-based statistical inference this part can then be treated as a constant, and therefore ignored.

Considering again for simplicity only absolutely continuous hazards, we give the following first *definition*: marked points (T_n, X_n), for which the F^N-intensity $\lambda_{T_n}^{(\theta,\psi)}(X_n)$ does not depend on θ, are called *non-innovations*.

The motivation behind this definition is obvious from the previous section: the contribution to the likelihood from observing such a point is a factor which depends only on ψ. Note, however, that non-innovations usually cannot be eliminated from the data as observations which would be irrelevant. They will typically enter the likelihood expression implicitly, through their contribution to the history. Thus, for example, in (iii) of the previous section,

one should certainly not ignore the censored observations in 4.2.(11), even though the hazard rates $C_i(t)$ in 4.2.(13b) would only depend on ψ.

This idea of non-innovations immediately suggests a simplification to the often complicated task of specifying statistical models for an MPP: if the untestable assumption is accepted, that certain hazards do not depend on the parameter of interest, these hazards do not even need to be specified in likelihood-based inference. This idea, called *partial model specification*, can therefore lead to a considerable saving in the modelling efforts.

We can actually often go one step further and split the factors which arise from the remaining marked points, *innovations*, into two parts. One that depends on θ, the parameter of interest, and one that does not. As a motivating example, we could think of a progressive censoring scheme where, every time there is a failure, some other individuals are censored. Failure would then be considered as a response and parameterized by θ, whereas censoring would not depend on θ.

To formalize this idea, suppose that to the observed MPP (T_n, X_n) there corresponds a particularly simple derived MPP $(\hat{T}_n, \hat{X}_n)$ such that

$$\begin{cases} \hat{T}_n = T_n \\ \hat{X}_n = p(X_n) \end{cases} \tag{1}$$

where $p\colon E \to \hat{E}$ is some known function. We then call $\hat{X}_n$ a *pre-mark* of X_n. In view of the additivity of the hazards over the marks as discussed in 2.2.1, or using an elementary form of 3.2.1.(4), we have that

$$\hat{\lambda}_t^{(\theta,\psi)}(\hat{x}) = \sum_{x \in p^{-1}(\hat{x})} \lambda_t^{(\theta,\psi)}(x), \tag{2}$$

where $p^{-1}(\hat{x}) = \{x \in E : p(x) = \hat{x}\}$ is the set of x-marks corresponding to the same pre-mark $\hat{x}$. We then use exactly the same factoring as in 2.2.1.(12), writing the intensity $\lambda_t^{(\theta,\psi)}(x)$ as the product

$$\lambda_t^{(\theta,\psi)}(x) = \hat{\lambda}_t^{(\theta,\psi)}(\hat{x}) \cdot \varphi_t^{(\theta,\psi)}(x|\hat{x}). \tag{3}$$

Here $\varphi_t^{(\theta,\psi)}(x|\hat{x}) = \lambda_t^{(\theta,\psi)}(x)/\hat{\lambda}_t^{(\theta,\psi)}(\hat{x})$ has the interpretation of the conditional probability of obtaining mark x at t, given $\mathcal{F}_{t-}^N$ and that there is a marked point which has pre-mark $\hat{x}$.

There are interesting special cases in which only the first of the two local characteristics, appearing on the right-hand side of (3), depends on the parameter of interest. In view of what was said above concerning non-innovations, we then have a situation where only the pre-mark intensities $\hat{\lambda}_t^{(\theta,\psi)}(\hat{x})$ give innovation contributions to the likelihood expression. This remains true for the likelihood contributions coming from "time intervals between points", represented by the exponential part in 4.2.(8), since obviously

$$\sum_{\hat{x} \in \hat{E}} \hat{\lambda}_t^{(\theta,\psi)}(\hat{x}) = \sum_{x \in E} \lambda_t^{(\theta,\psi)}(x). \tag{4}$$

Thus the method of partial model specification can be applied again. The paper of Andersen *et al.* (1988) provides interesting examples of this, with full details.

A similar reduction in the likelihood expression applies if only the second factor on the right-hand side of (3) depends on the parameter of interest. It is easy to check that then each observed marked point (T_n, X_n) with pre-mark $\hat{X}_n$ contributes the factor $\varphi_{T_n}^{(\theta,\psi)}(X_n|\hat{X}_n)$ to the likelihood, and that the exponential part in the likelihood does not depend on the parameter of interest at all.

Effectively, partial specification of a statistical model means that one only specifies those

hazards which depend on the parameter of interest. Sometimes statisticians, however, deliberately ignore factors in the likelihood expression although they depend on θ. The remaining part is then called *partial likelihood*.

The oldest, and certainly the most important, case of partial likelihood arises from Cox's proportional hazards model (Cox, 1972). There the partial likelihood can be viewed as a product of terms of the form $\varphi_{T_n}^{(\theta,\psi)}(X_n|\hat{X}_n)$, with notation as in (3). The motivation to use such a partial expression is again the modelling economy it provides: in Cox's model, $\varphi_t^{(\theta,\psi)}(x|\hat{x})$ does not depend on the nuisance parameter ψ.

To show this in some detail, let the data $\{(S_i^*, \delta_i);\ 1\leq i\leq k\}$ be as in 4.2.(11), and let (T_n, X_n) be defined as in 4.2.(12). We let the pre-mark $\hat{X}$ be simply

$$\hat{X}_n=p(X_n)=\delta_i \quad \text{on} \quad \{X_n=(i,\delta_i)\}, \tag{5}$$

i.e. $\hat{X}_n=1$ if there is a failure at T_n, and $\hat{X}_n=0$ if there is a censoring. Thus $\varphi_t^{(\theta,\psi)}\{(i,1)|1\}$ is the probability that individual i fails given that there is an observed failure at t. The quantity $\varphi_t^{(\theta,\psi)}\{(i,0)|0\}$ would have a similar meaning for censoring, but in practice it is never specified on the grounds that it is directly postulated not to depend on θ.

We now show how to arrive at the familiar form of the Cox partial likelihood. Suppose that the failure intensity of individual i, with respect to F^N, can be written in the form

$$\lambda_t(i,1)=\lambda_0(t)\cdot\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\cdot 1_{\{t\leq S_i^*\}}, \tag{6}$$

where $\lambda_0(\cdot)$ is an "unknown but fixed baseline hazard", $\boldsymbol{\beta}$ is a vector of model parameters, and $\mathbf{Z}_i$ is a vector of known fixed characteristics describing individual i. The baseline hazard is then taken to correspond to ψ, as an infinite dimensional nuisance parameter. (This kind of terminology is perhaps not quite fair since good estimates of $\lambda_0(\cdot)$ are needed for predicting survival probabilities.) Similarly, $\boldsymbol{\beta}$ is viewed as the parameter of interest. Then, from (3), we have that on the event $\{X_n=(i,1)\}$

$$\begin{aligned}\varphi_{T_n}(X_n|\hat{X}_n)&=\frac{\lambda_{T_n}(X_n)}{\hat{\lambda}_{T_n}(\hat{X}_n)}\\ &=\frac{\lambda_0(T_n)\exp(\boldsymbol{\beta}_i'\mathbf{Z}_i)}{\sum_i \lambda_0(T_n)\exp(\boldsymbol{\beta}_i'\mathbf{Z}_i)\cdot 1_{\{T_n\leq S_i^*\}}}\\ &=\frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_i)}{\sum_{\{i:\,S_i^*\geq T_n\}}\exp(\boldsymbol{\beta}_i'\mathbf{Z}_i)}\end{aligned} \tag{7}$$

The product over such values of n where $\hat{X}_n=1$ is now the familiar partial likelihood expression for Cox's model.

It is worth noting that the crude failure rates, which are sums of the intensities over index i, clearly depend on $\boldsymbol{\beta}$. Thus the partial likelihood is formed by ignoring two factors which both depend on "the parameter of interest". Nevertheless, the estimators based on partial likelihood are often known to have "good statistical properties" (see, for example, Gill, 1985; Jacod, 1987).

We close with two comments regarding statistical inference and unobservables.

If the hazards are originally conditioned on some unobservables, it is of course always possible to derive corresponding consistent hazards which do not. In the example of section 3.2.3, this meant that the "original" F^*-hazards were replaced by F^N-hazards or $\hat{\mathrm{F}}$-hazards. The likelihood expression corresponding to the observed data can then be constructed

according to the general principles discussed above, and in certain cases partial model specification may be possible. An alternative approach is to first form the likelihood expression with respect to the "original" hazards, and then eliminate the unobservables by averaging with respect to their distribution. While this latter approach is possible in some cases, thus leading to the same likelihood expression as the former, this is not always so. If the unobservables do not remain fixed over the observation interval, or if the model is specified only partially, the likelihood expressions obtained by the two methods can actually differ by a factor which depends on the parameter of interest. In the example of 3.2.3, if Z is known and the scope is only to estimate $a_i^*(1)$, with part 2 failures acting as a censoring mechanism, it is clearly possible to eliminate $a_i^*(2)$ completely from the likelihood expression (cf. the reasoning in 4.2.(iii)). This corresponds to observing the F^*-history up to time $T_1 = S_1 \wedge S_2$. However, if Z is unknown and the corresponding F^N-hazards are used to derive the likelihood expression, factors involving $a_i^*(2)$ also contain $a_i^*(1)$, the parameter of interest, and vice versa.

Our second remark concerns the methods of parameter estimation in the presence of unobservables. It was established in section 3.2 that the estimation of unobservables was essentially by an application of Bayes' rule. This then provided the necessary link between the two levels of information, and the corresponding statistical models. When the task is to estimate model parameters, most statisticians would routinely maximize the likelihood corresponding to observed data. However, it is very tempting to treat unknown model parameters together with possible unobservables, and simply apply the Bayesian paradigm to both. Conceivably, a drawback of such a method would be the high dimension of the vector estimate.

Acknowledgements

As is obvious from the references, much of the material presented here is based on my long-time collaboration with Ilkka Norros and Pentti Haara. It is a pleasure to acknowledge their part here.

These notes were written into the form of a paper during my six-week visit to the University of Western Australia. I am grateful to the Department of Mathematics, and to my host, Tim Brown, for having provided such an inspiring atmosphere and good facilities for the work.

References

Aalen, O. O. (1987a). Two examples of modelling heterogeneity in survival analysis. *Scand. J. Statist.* **14,** 19–25.

Aalen, O. O. (1987b). Mixing distributions on a Markov chain. *Scand. J. Statist.* **14,** 281–289.

Aalen, O. O. (1987c). Dynamic modelling and causality. *Scand. Actuar. J.* **1987,** 177–190.

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statist. Med.* **7,** 1121–1137.

Aldous, D. (1981). Weak convergence and the general theory of processes (incomplete draft of a monograph, unpublished).

Allison, P. D. (1984). *Event history analysis. Regression for longitudinal event data.* Sage University Paper 46. Sage Publications, Beverly Hills, London, New Delhi.

Andersen, P. K. & Borgan, Ø. (1985). Counting process models for life history data: a review (with discussion). *Scand. J. Statist.* **12,** 97–158.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1988). Censoring, truncation and filtering in statistical models based on counting processes. *Contemporary mathematics* (ed. N. U. Prabhu), **80,** 19–60. American Mathematical Society.

Arjas, E. (1981a). A stochastic process approach to multivariate reliability systems; notions based on conditional stochastic order. *Math. Oper. Res.* **6,** 263–276.

Arjas, E. (1981b). The failure and hazard processes in multivariate reliability systems. *Math. Oper. Res.* **6,** 551–562.

Arjas, E. (1985). Contribution to the discussion on the paper by P. K. Andersen & Ø. Borgan. *Scand. J. Statist.* **12,** 150–153.

Arjas, E. & Greenwood, P. (1981). Competing risks and independent minima, a marked point process approach. *Adv. Appl. Probab.* **13,** 669–680.

Arjas, E. & Haara, P. (1984). A marked point process approach to censored failure time data with complicated covariates. *Scand. J. Statist.* **11,** 193–209.

Arjas, E. & Haara, P. (1988). A note on the exponentiality of total hazards before failure. *J. Multivariate Anal.* **26,** 207–218.

Arjas, E. & Norros, I. (1984). Life lengths and association: a dynamic approach. *Math. Oper. Res.* **9,** 151–158.

Arjas, E. & Norros, I. (1988). On the filtering of histories of marked point processes (incomplete manuscript).

Barlow, R. & Proschan, F. (1975). *Statistical theory of reliability and life testing*. Holt, Rinehart and Winston, New York.

Boel, R., Varaiya, P. & Wong, E. (1975). Martingales on jump processes, I: Representation results. *SIAM J. Control Optim.* **13,** 997–1021.

Brémaud, P. (1981). *Point processes and queues. Martingale dynamics*. Springer-Verlag, New York, Heidelberg, Berlin.

Brémaud, P. & Jacod, J. (1977). Processus ponctuels et martingales: resultats recent sur la modelisation et le filtrage. *Adv. Appl. Probab.* **9,** 362–416.

Brown, T. C. (1988). The Doob–Meyer decomposition. Research Report. Department of Mathematics, University of Western Australia.

Brown, T. C. & Nair, M. G. (1988). A simple proof of the multivariate random time change theorem. *Stochastic Process. Appl.* **29,** 247–256.

Clayton, D. (1988). The analysis of event history data: a review of progress and outstanding problems. *Statist. Med.* **7,** 819–841.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *J. Roy. Statist. Soc. Ser. B* **21,** 411–421.

Cox, D. R. (1962). *Renewal theory*. Methuen, London.

Cox, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B* **34,** 187–220.

Gill, R. D. (1985). Note on product integration, likelihood and partial likelihood for counting processes (unpublished manuscript).

Gill, R. D. & Johansen, S. (1989). Product integrals and counting processes. *Ann. Statist.* (to appear).

Heckman, J. & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52,** 271–320.

Holland, P. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81,** 945–970.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity *Biometrika* **71,** 75–83.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73,** 387–396.

Hougaard, P. (1987). Modelling multivariate survival. *Scand. J. Statist.* **14,** 291–304.

Jacobsen, M. (1982). Statistical analysis of counting processes. *Lecture notes in statistics,* **12.** Springer-Verlag, New York, Heidelberg, Berlin.

Jacod, J. (1975). Multivariate point processes: predictable projection, Radon–Nikodym derivatives, representation of martingales. *Z. Wahrsch. Verw. Gebiete* **31,** 235–253.

Jacod, J. (1987). Partial likelihood process and asymptotic normality. *Stochastic Process. Appl.* **26,** 47–71.

Kalbfleisch, J. D. & Prentice, R. (1980). *The statistical analysis of failure time data.* Wiley, New York.

Karr, A. F. (1986). *Point processes and their statistical inference*. Marcel Dekker, New York, Basel.

Koch, G. (1986). A dynamical approach to reliability theory. *Proc. Int. School of Phys. "Enrico Fermi", XCIV,* 215–240. North-Holland, Amsterdam, New York.

Liptser, R. S. & Shiryayev, A. W. (1978). *Statistics of random processes,* **2.** Springer-Verlag, Berlin.

Manton, K. G. & Stallard, E. (1988). *Chronic disease modelling*. Griffin, London.

Manton, K. G., Stallard, E. & Vaupel, J. W. (1981). Methods for comparing the mortality experience of heterogeneous populations. *Demography* **18,** 389–410.

Miettinen, O. S. & Cook, E. F. (1981). Confounding: essence and detection. *Amer. J. Epidem.* **114,** 593–603.

Norros, I. (1985). Systems weakened by failures. *Stochastic Process. Appl.* **20,** 181–196.

Norros, I. (1986). A compensator representation of multivariate life length distributions, with applications. *Scand. J. Statist.* **13,** 99–112.

Schumacher, M., Olschewski, M. & Schmoor, C. (1987). The impact of heterogeneity on the comparison of survival times. *Statist. Med.* **6,** 773–784.

Schweder, T. (1970). Composable Markov processes. *J. Appl. Probab.* **7,** 400–410.

Shaked, M. & Shanthikumar, J. G. (1987). The multivariate hazard construction. *Stochastic Process. Appl.* **24,** 241–258.

Shaked, M. & Shanthikumar, J. G. (1989). Multivariate stochastic orderings and positive dependence in reliability theory. *Math. Oper. Res.* (to appear).

Suppes, P. C. (1970). *A probabilistic theory of causality*. North-Holland, Amsterdam.

Thygesen, L. (1988). While questions should be answered by socio-demographic risk research in the future? *Proceedings of the Nordic Seminar on Empirical Life History Analysis and Panel Studies,* Stockholm, 1987. Technical Report 46. Nordisk Statistisk Sekretariat.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks., *Proc. Nat. Acad. Sci. USA* **72,** 20–22.

Yashin, A. & Arjas, E. (1988). A note on random intensities and conditional survival functions. *J. Appl. Probab.* **25,** 630–635.

Yashin, A., Manton, K. G. & Stallard, E. (1986). Dependent competing risks: a stochastic process model. *J. Math. Biol.* **24,** 119–140.

Elja Arjas, Department of Applied Mathematics and Statistics, University of Oulu, SF-90570, Oulu, Finland

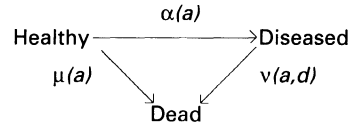
DISCUSSION OF ELJA ARJAS' LECTURE

NIELS KEIDING

University of Copenhagen

Elja Arjas has given us a strong and very useful survey of the application of the French "general theory of processes", including martingale theory, to current problems in reliability and survival analysis. I am certainly not the best choice for a critical discussant, having used the important paper in this journal by Arjas & Haara (1984) to attempt to clean up the general formulation of censoring and truncation and also discuss what we term filtering (cf. Andersen *et al.*, 1988).

A principal point of the martingale approach is its strong emphasis on *accumulation of information as time goes on*. Arjas has—here and earlier (cf. Arjas, 1985, 1986; Arjas & Haara, 1984, 1987)—advocated a radical version, his so-called *real-time approach*, where the time in question is always calendar time. To put this approach in perspective, I want below to quote some typical problems in survival analysis where several time-scales are of interest and where calendar time is not always of prime importance. A convenient starting point is the simple illness–death process specified by transition intensities between three states as follows:



where a is age, d duration since disease onset. Of course the *incidence* $\alpha(a)$, the *mortality of the healthy* $\mu(a)$ and the *mortality of the diseased* $\nu(a, d)$ may all depend on calendar time t in addition to a and d. In practice it is often useful to consider the mortality of the diseased primarily a function of age a if the disease is "mild" (such as diabetes) and primarily a function of disease duration d for aggressive diseases like most cancers.

The first example is *clinical trials with staggered entry* where we shall use duration d as the basic time variable. Patients arrive for the trial (contract the disease) sequentially in calendar time and stay on the trial until death (in the simplest situation we may disregard loss to follow-up). At a certain date, called "end of study", the survival/death status of all patients is assessed. Using duration as basic time-scale we then have innovative marks (deaths) and non-innovative marks (censorings). The situation is represented in the (calendar time, duration) plot in Fig. 1: this is called a *Lexis diagram* in demography. A very interesting analysis of the martingale structure in the bivariate (calendar time, duration) scale was given by Sellke & Siegmund (1983), cf. Slud (1984). This analysis emphasizes that problems quickly arise: as an example, while censoring may in general depend on the past development of the process (the "history") this is only true in the duration time-scale, not in calendar time without compromising the simple martingale structure (see Andersen *et al.* (1988) for further discussion).

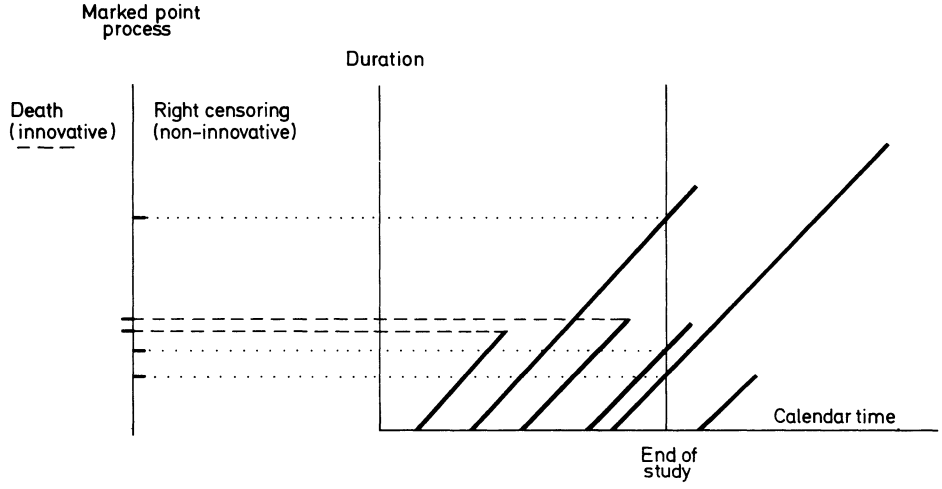


Fig. 1. Lexis diagram of a clinical trial with staggered entry and a single final analysis. Individuals enter the trial at diagnosis, sequentially in time, and survival/death status is assessed at end of study. On the extra vertical axis to the left the innovative and non-innovative marks are indicated.

Assume now that the analysis of the trial was not the final one but rather an *interim analysis* where perhaps some interesting but not statistically significant effect was noted. A (conditionally) independent confirmatory analysis may then be performed including only that part of the time at risk spent in the calendar period between interim and final analysis (Keiding *et al.*, 1987). The corresponding marked point process now also includes non-innovative marks due to patient entry (left truncation) (cf. Fig. 2). Keiding & Gill (1988) provided a general study of left truncation.

For our second example we take *age* as basic time variable and wish to study disease onset (incidence) as well as mortality from the disease. For *disease onset* the situation is almost identical to the above clinical trial: persons are born sequentially in time, always at age 0, and followed until disease onset, but at most until a certain final date. However, two complications arise: some persons were already alive at the initial date and have to be *left truncated* at their age then, and some persons die (or are lost to follow-up for other reasons) within the study period, without having contracted the disease. This means that two new non-innovative marks become possible: patient entry and death as "healthy". Fig. 3 outlines the situation. A variant of this incidence estimation problem occurs when onset is known only for survivors until the

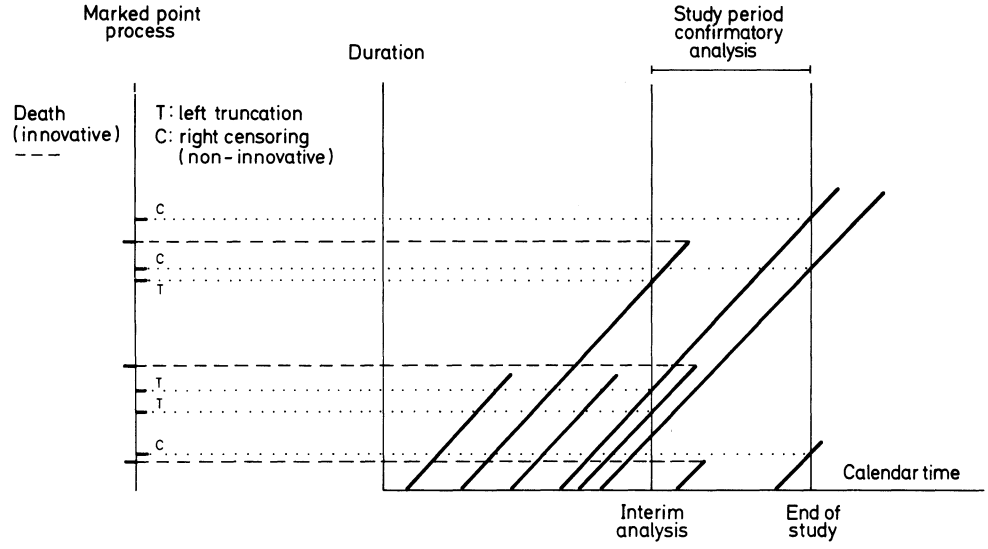


Fig. 2. Lexis diagram of a clinical trial with staggered entry and one interim and one final analysis. On the extra vertical axis are indicated the innovative and non-innovative marks of the marked point process corresponding to a confirmatory analysis based only on the information from calendar time since the interim analysis.

end of study. Keiding *et al.* (1989) showed how it is possible to obtain incidence estimates under such retrospective sampling, if information on current total population size and mortality of the diseases is available. However, the likelihood changes significantly under this sampling scheme in contrast to the simple truncation—censoring frameworks exemplified in the rest of this contribution. The retrospective estimation problem motivates a general study of non-parametric inference on intensities that vary with both age (or duration) *and* calendar time (see Keiding *et al.*, 1989; Capasso, 1987).

To study *mortality of the diseased* patients enter as they become diseased in calendar time at various ages, or they were already diseased at the start of the study period: both situations

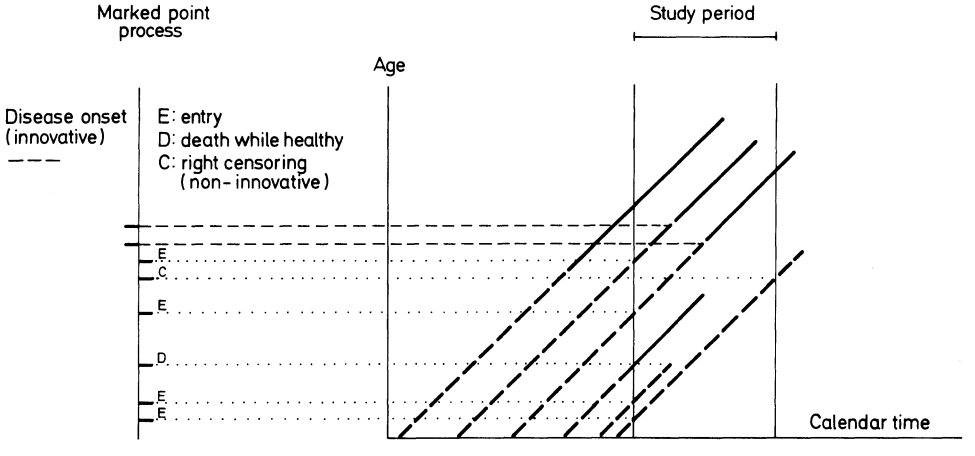


Fig. 3. Lexis diagram of a follow-up study of disease incidence, conducted on a dynamic cohort in a specific calendar time period. On the extra vertical axis are indicated the innovative marks (disease onset) and the three types of non-innovative marks (E, entry at initial date for persons already alive (and still not having contracted this disease) at that date; D, death while "healthy"; C, right censored due to study conclusion).

count as (non-innovative) left truncation. They are followed until death (innovative) or end of study period (non-innovative) (cf. Fig. 4). Observe that if disease incidence as well as death from disease are studied, the point where a patient gets the disease has an innovative mark in the context of disease incidence but a non-innovative mark in the context of mortality (patient entry).

In the above examples the times of disease onset and death are always assumed to be known precisely. In particular for disease onset this may often be an unrealistic assumption, indeed

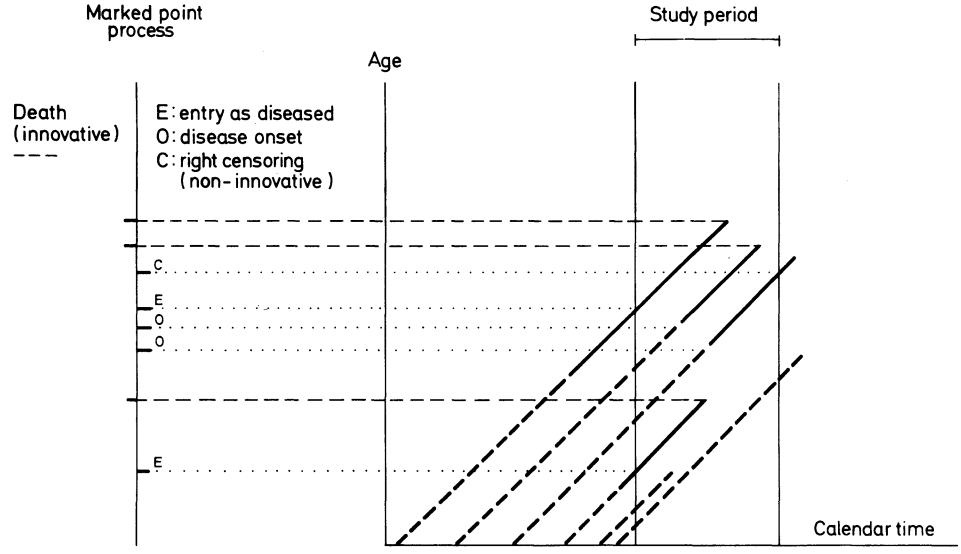


Fig. 4. Lexis diagram of a follow-up study of mortality among diseased, conducted on a dynamic cohort in a specific calendar time period. On the extra vertical axis are indicated the innovative marks (death); as well as the three types of non-innovative marks (E, entry (as diseased) at initial date; O, entry because of disease onset in study; C, right censored at study conclusion).

the onset may only be known up to a rather wide time interval. Methodology for interval-censored data was developed by Turnbull (1976) (cf. Dempster *et al.*, 1977), and recent important contributions include Finkelstein (1986), Grüger (1986) (cf. Grüger *et al.*, 1988) and Groeneboom (1987). It would be interesting to know whether Arjas's results on prediction and state estimation might be integrated into such a methodology: certainly one needs to predict where in the interval disease onset took place.

Finally a brief comment to Arjas's thoughts on heterogeneity. "The ultimate ... heterogeneity is that all individuals are different" (end of section 3). Although this is almost a truism, it is of interest to confront this and other statements with Westergaard's paradigm as developed in his textbooks on morbidity and mortality and general elementary statistics, here quoted from his survey paper (Westergaard, 1916, p. 246 ff.):

> No two persons are alike; one will be attacked by an epidemic disease, another will escape; one will marry, another remain single; etc., but on the whole in a certain class a certain number of deaths or marriages will take place. The question now is, What are the limits of deviation from the average? Can we find a conformity to the binomial law of error similar to that found in playing at dice or cards, or do other laws of frequency apply to these phenomena?
>
> In vital or economic statistics most numbers have a much wider margin of deviation than is experienced in games. Thus the death rate, the birth rate, the marriage rate, or the relative

frequency of suicide fluctuates within wide limits. But it can be proved that, by dividing the observations, sooner or later a marked tendency to the binomial law is revealed in some part of the observations.

* * *

If no conformity to the binomial law can be found in a series of observations, the first task is the classification of the data. The most obvious classifications are by sex, age, profession, residence, etc.

If we do not at once reach the binomial law, we must try further subdivisions. Since each of these classes or subclasses has its peculiar conditions, and is under the influence of different systems of causes, we can say that the more quickly we reach the point where the binomial law holds good, the fewer causes have been acting, whereas the greater the deviation from the binomial law, the greater the number of active causes.

This belief of a general "canonical" underlying probability structure may seem over-optimistic or even naive today, but the idea has served us well for decades. Is Arjas ready for a more general critique of this paradigm?

In conclusion I want to thank the former hardcore applied probabilist Elja Arjas once again for a most valuable survey. He has achieved very significant results in survival analysis in a short period of time and it is a great gain for the field that he has directed his attention towards it.

Additional references

Arjas, E. (1986). Stanford heart transplantation data revisited: a real time approach. In *Modern statistical methods in chronic disease epidemiology* (ed. S. H. Moolgavkar & R. L. Prentice), 65–81. Wiley, New York.

Arjas, E. & Haara, P. (1987). A logistic regression model for hazard: asymptotic results. *Scand. J. Statist.* **14,** 1–18.

Capasso, V. (1987). Stochastic age-dependent population dynamics. A counting process approach for parameter estimation. Preprint, Department of Mathematics, University of Bari, Italy.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B.* **39,** 1–38.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845–854.

Groeneboom, P. (1987). Asymptotics for incomplete censored observations. Report 87–18, Department of Mathematics, University of Amsterdam.

Grüger, J. (1986). Nichtparametrische Analyse sporadisch beobachtbarer Krankheitsverlaufsdaten. Dissertation, Universität Dortmund.

Grüger, J., Kay, R. & Schumacher, M. (1988). The validity of inferences based on incomplete observations in disease state models. Technical Report, Institute of Medical Biometry and Informatics, University of Freiburg.

Keiding, N., Bayer, T. & Watt-Boolsen, S. (1987). Confirmatory analysis of survival data using left truncation of the life times of primary survivors. *Statist. Med.* **6,** 939–944.

Keiding, N. & Gill, R. D. (1988). Random truncation models and Markov processes. Report MS-R 8817, Centre for Mathematics and Computer Science, Amsterdam. *Ann. Statist.* (tentatively accepted).

Keiding, N., Holst, C. & Green, A. (1989). Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality. *Amer. J. Epidemiol* (to appear).

Sellke, T. & Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70,** 315–326.

Slud,. E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Ann. Statistic.* **12,** 551–571.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38,** 290–295.

Westergaard, H. (1916). Scope and methods of statistics (with discussion). *Quart. Publ. Amer. Statist. Assoc.* **15,** 225–291.

ØRNULF BORGAN
University of Oslo

First of all I want to congratulate Elja Arjas on a very nice and interesting paper. In particular I found that he was able to present a difficult topic in a very inspiring and understandable manner.

Without doubt the marked point process approach presented by Arjas will be found increasingly important in the statistical modelling and analysis of life history data in the years to come. Here I will concentrate on one particular area where this approach may turn out to be very useful, namely the modelling and analysis of life history data where the exact time of occurrence for some of the events are not recorded.

Such a situation may arise in the study of non-reversible complications of chronic diseases, such as diabetes or cancer, and their effect on mortality. Here a patient is examined at visits to a hospital and observation of the exact time of occurrence of the disease complication is typically not possible. It is only possible to ascertain that the complication has occurred sometime between two such visits.

Another situation of a similar nature is long-term animal carcinogenicity trials, where it can only be assessed after death whether an animal has cancer or not. Again no exact time of the occurrence of the event of interest is available. However, often supplementary data are obtained by serial sacrifice, that is, animals are killed at certain times and it is assessed whether or not they already had the tumour. The modelling and analysis of such carcinogenicity trials have been discussed by many authors (see McKnight & Crowley (1984) for a review and discussion of a number of the proposals which have been made in the literature). In spite of all the research effort which has been put into this problem, a systematic and theoretically fully satisfactory approach is still missing.

Using the marked point process approach, a natural and systematic way to address the above problems may be the following. We first model the occurrences of deaths and onsets of the disease/disease complication, as well as other events of interest like visits to a hospital or sacrifice of an animal, as one huge marked point process $\{(T_n, \mathrm{X_n}); n=1, 2, \ldots\}$ with internal history $\mathscr{F}_t^N$. From this marked point process (which is not observable) we may then derive a marked point process $\{(\hat{T}_n, \hat{X}_n); n=1, 2, \ldots\}$ with internal history $\hat{\mathscr{F}}_t$ consisting of all the observable events, as described in section 3.2 of the paper.

If we now, for the concreteness of the discussion, concentrate on the estimation of the integrated intensity for the occurrence of the disease/disease complication, we may think about estimating this in the following way. We start out by some initial guess for the unknown parameters in the full marked point process model. Then using the results of section 3.2 of the paper it ought to be possible to compute the expected number of individuals alive without the disease/disease complication at any time t as well as the expected number of occurrences of this event up to time t, given what we have observed, i.e. given $\hat{\mathscr{F}}_\infty$. Using these expected values we then update our estimates for the parameters in the original model using Nelson–Aalen type estimators. These steps may then be repeated in an iterative manner, cf. the EM-algorithm.

To work out carefully the details of the outline presented above is of course not at all simple. In this connection one would also have to consider conditions on the temporal pattern of the visits to the hospital or on the sacrifice design in order to make the parameters of the model identifiable. Also will remain the difficult problem of deriving the large sample properties of the resulting estimators. Nevertheless, I believe that this approach may turn out to be a promising one, and I am very much looking forward to hearing Elja Arjas' opinion on these matters.

Additional reference

McKnight, B. & Crowley, J. (1984). Tests for differences in tumor incidence based on animal carcinogenesis experiments. *J. Amer. Statist. Assoc.* **79,** 639–648.

PER KRAGH ANDERSEN
University of Copenhagen

First of all I wish to join the other discussants in congratulating Dr Arjas on a very clearly presented paper containing a lot of material which, I am sure, will serve as inspiration for many researchers in the area in years to come.

One such topic is the *unit-exponentiality of total hazards* (section 2.1.4) and in what follows I shall sketch some possible developments emerging from this theme. In one of its simplest forms this proposition states that if $T_1, \ldots, T_n$ are i.i.d. life times with hazard rate $r(\cdot)$ then the compensators $A_t^{\mathrm{F}}(i)$, $i=1, \ldots, n$, of the counting processes

$$N_t(i)=1_{\{T_i \leqslant t\}}, \qquad i=1, \ldots, n,$$

w.r.t. some history F have the property that $A_{T_1}^{\mathrm{F}}(1), \ldots, A_{T_n}^{\mathrm{F}}(n)$ are independent and identically exp(1) distributed.

For the history $\mathrm{F}=(F_t)$ with

$$F_t=\sigma\{N_u(i);\ i=1, \ldots, n;\ u \leqslant t\}$$

(Example (ii) of section 2.1.3) the compensators are simply

$$A_t^{\mathrm{F}}(i)=\int_0^t 1_{\{s \leqslant T_i\}}\, r(s)\, ds=R(t \wedge T_i)$$

but more interesting applications are possible for more general life-time models. One such example is the Cox regression model with random and possibly time-dependent covariates where the hazard rates are

$$r_i(t)=\lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_i(t)\}, \qquad i=1, \ldots, n$$

(cf. section 4.3). Here the proposition of section 2.1.4 still applies and the "residuals"

$$A_{T_i}^{\mathrm{F}}(i)=\int_0^{T_i} \lambda_0(s) \exp\{\boldsymbol{\beta}' \mathbf{Z}_i(s)\}\, ds, \qquad i=1, \ldots, n,$$

are independent and exp(1) distributed. Right censored life times T_i can be handled by treating the corresponding residual $A_{T_i}^{\mathrm{F}}(i)$ as right censored and finally one might argue that inserting consistent estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(\cdot)$ (Andersen & Gill, 1982) "should not destroy too much" the fact that $\{\hat{A}_{T_i}^{\mathrm{F}}(i),\ i=1, \ldots, n\}$ is approximately a sample of independent (possibly right censored) exp(1)-distributed variates (Kay, 1977; Arjas, 1988).

And in fact the plotting procedure suggested by Arjas (1988) (a *total time on test type plot* based on the residuals) seemed to work quite well and certain deviations from linearity of the plots could be interpreted as certain violations of the basic assumptions of the Cox regression model.

The small sample properties of the residuals have, however, been questioned by, for example, Lagakos (1981) and deviations from linearity of the *cumulative hazard plot* for the

residuals (Kay, 1977) have been demonstrated by Crowley & Storer (1983) to be a poor indication of lack of fit of the Cox regression model. The latter statement has in fact led the distributors of the statistical computer package BMDP to leave out the possibility of performing this goodness of fit plot in the latest versions of their P2L program.

If an explanation can be found why some procedures based on these residuals may work while others may not then the way may be paved to go into a deeper discussion of properties of other kinds of residuals and diagnostics for the Cox regression model. Barlow & Prentice (1988) suggested a whole class of such residuals the simplest one being $D_i - A^{\mathrm{F}}_{T_i}(i)$, where $D_i = N_\infty(i)$ is the indicator for individual i failing. Their general residual which for some vector $\mathbf{f}_i(\cdot)$ of weighting factors is

$$\int_0^\infty \mathbf{f}_i(t)\, dN_i(i) - \int_0^\infty \mathbf{f}_i(t) 1_{\{t \le T_i\}} \lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_i(t)\}\, dt$$

can then be written as

$$\mathbf{f}_i(T_i) D_i - \int_0^{T_i} \mathbf{f}_i(t)\, dA^{\mathrm{F}}_i(i).$$

In particular, the residual corresponding to the choice

$$\mathbf{f}_i(t) = \mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}, t)$$

where

$$\mathbf{E}(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^{n} \exp\{\boldsymbol{\beta}' \mathbf{Z}_i(t)\} \mathbf{Z}_i(t) 1_{\{t < T_i\}}}{\sum_{i=1}^{n} \exp\{\boldsymbol{\beta}' \mathbf{Z}_i(t)\} 1_{\{t \le T_i\}}}$$

is closely related to regression diagnostics for the model (Cain & Lange, 1984; Reid & Crépeau, 1984).

I wonder whether Dr Arjas has had any thoughts about such possibilities and their extension to more general life history models like those discussed by Andersen & Borgan (1985).

Additional references

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10,** 1100–1120.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Amer. Statist. Assoc.* **83,** 204–212.

Barlow, W. E. & Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* **75,** 65–74.

Cain, K. C. & Lange, N. T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* **40,** 493–499.

Crowley, J. & Storer, B. E. (1983). Contribution to the discussion of the paper by M. Aitkin, N. Laird & B. Francis. *J. Amer. Statist. Assoc,*. **78,** 264–292.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Appl. Statist.* **26,** 227–237.

Lagakos, S. W. (1981). The graphical evaluation of explanatory variables in proportional hazards regression models. *Biometrika* **68,** 93–98.

Reid, N. & Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika* **72,** 1–9.

BENT NATVIG
University of Oslo

A couple of years ago Elja Arjas asked me whether it would be a good idea that he wrote a book in the area of survival models and martingale dynamics. My answer was YES! The reason was that I found the basic literature in this area very close to being unreadable, at least for people in the mainstream of my own field, reliability theory. After having read Elja's manuscript it is just to conclude that he has already made a major step towards the goal of writing an important and very readable book, and I hope that it will appear as soon as possible. Moreover, Elja has given an excellent series of lectures at this conference, really being able to bridge the gap between mathematical rigour and human intuition in this area.

The problem with Elja's lecture is that it does not leave much space for a paid discussant. I have handed over to him a list of misprints to convince at least him that I have read his manuscript most carefully. This leaves me with three direct comments as a start.

The first two comments are concerned with section 3.1.2—Applications of the prediction process—and more specifically with the F^N-WBF (weakened by failures) property. F^N refers to the internal history. First a piece of good advice if one wants to understand this concept—forget Arjas & Norros (1984) and read Norros (1985). The concept seems very useful when modelling causal effects between components when these are either subjected to the same external stress, as for instance bad weather, fires and earthquakes, or are sharing a common load. A key proposition mentioned by Arjas is that the F^N-WBF property implies that the life lengths of the components are associated random variables. This latter property is very central in reliability theory, for instance when establishing bounds for the availabilities and unavailabilities in a fixed time interval for multistate monotone systems as in Funnemark & Natvig (1985). Associated random variables have very nice properties, but the verification of association from its definition is hopeless. Hence there is a great need to have sufficient conditions for random variables to be associated. The proposition referred to above gives a nice such condition.

When listening to Elja one can get the impression that some of the results are easy to arrive at, as for instance that the F^N-IFR (increasing failure rate) property implies the F^N-WBF property. However, the proof by Norros (1985) is based on properties of the residual prediction process, the latter being wiped under the carpet by Elja in his lecture. More specifically the result is immediate from the facts that the F^N-IFR property holds iff the residual prediction process is a decreasing process, and that the F^N-WBF property holds iff the residual prediction process jumps downwards at failure times.

My final direct comment refers to section 4—Likelihood based inference for marked point processes. I start by repeating what I said at the Bolkesjø meeting four years ago (see Natvig, 1985): "In a way I hope more statisticians will start to look into Bayesian methods as unprejudiced as people in reliability. Specifically, I will challenge the people gathered at this conference working on life history data. In some applications at least the database does not cry after asymptotic theory based on the martingale central limit theorem. On the other hand there is a great need to develop more and better Bayesian methods." Looking at Elja's lecture as a whole he seems to be on the right track. Especially I noted his statement in the introduction to section 4: "We do not claim that the likelihood function is everything that is of interest in parametric statistical inference, but that is how far we go".

I think Elja's deductions in this section clarify a series of obscure results in the book by Kalbfleisch & Prentice (1980), which led to the censoring of my interests in this area some years ago. Especially I liked the analysis of the information going down the drain when

considering the partial likelihood only. However, I did not like the paper's final sentence: "Nevertheless, the partial likelihood is known to have 'good statistical properties'." This is of course just ASYMPTOTICALLY!

Elja Arjas is concluding his introductory section by stating that: "The role of information will be one of the central themes in this talk". Both Elja and I have recently been working on the role of information in minimal repair models (see the hopefully forthcoming paper Natvig, 1988). Here I will review a part of our efforts.

An important discussion of minimal repair models is given in the review paper Bergman (1985, p. 24). Here the time S to failure of a device under study is assumed to have an absolutely continuous cumulative distribution function F and failure rate function r. In the terminology of Bergman (1985) a *statistical* minimal repair of the device means that if a failure occurs at time t then, after the repair, the survival probability to time $t+s$ equals $\{1-F(t+s)\}/\{1-F(t)\}$ and the failure rate function equals $r(t+s)$, $s\geq 0$. However, the author points out that we have to distinguish between *statistical* minimal repair, and *physical* minimal repair in which case the failed unit is restored to the exact physical condition as it had just before the failure.

This difference is made clearer by Arjas & Norros (1989) who simply use the term "black box" minimal repair for statistical minimal repair. To them (and to me) this seems to be a rather abstract notion for a device consisting of several components by simply asking: How does one repair a black box without knowing what is inside? A main point of these authors is that the notion of minimal repair must be related to the information at hand. This is in the best Bayesian spirit!

Consider the change of distributions which arises from exactly one minimal repair taking place at the first failure. Under what is called the G-history, where this corresponds to the "black box" minimal repair, the transformed survival function is given by

$$\begin{aligned} Q^{G}(S>t) &= \bar{F}(t) - \int_0^t \{\bar{F}(t)/\bar{F}(s)\}\, d\bar{F}(s) \\ &= \bar{F}(t)\{1+R(t)\}, \end{aligned} \tag{1}$$

where $\bar{F}(t)=1-F(t)$ and $R(t)=-\ln \bar{F}(t)$ is the cumulative hazard function of the device. The G-compensator of the single point counting process $N_t=1_{\{t\geq S\}}$ is given by

$$A_t^G = R(t\wedge S). \qquad t\geq 0. \tag{2}$$

Correspondingly, the G-compensator linked to (1) is given by

$$\begin{aligned} B_t^G &= -\ln\,[\bar{F}(t\wedge S)\{1+R(t\wedge S)\}] \\ &= A_t^G - \ln\,(1+A_t^G). \end{aligned} \tag{3}$$

For a general history F the F-minimal repair, defined by Arjas & Norros (1989) through the F-compensator, B^F, is given by

$$B_t^F = A_t^F - \ln\,(1+A_t^F). \tag{4}$$

As presented in the latter paper (4) seems to be a pure mathematical generalization of (3) just replacing the G-compensator of N_t, A_t^G, by the corresponding F-compensator A_t^F. However, a proper argument in favour of (4) has later been given by Arjas & Norros (1988). This argument is again based on theorem 2.1 of Norros (1987).

A somewhat surprising result in Arjas & Norros (1989) is that the transformed life length

corresponding to the F-minimal repair is stochastically shorter than the one corresponding to the "black box" minimal repair. The significance of this result for applications is still unclear, at least to me, and led to Natvig (1988). Here a system of n components is considered. Let $(i=1, \ldots, n)$

$$X_i(t)=\begin{cases}1 & \text{if the } i\text{th component functions at time } t\\ 0 & \text{if the } i\text{th component is failed at time } t.\end{cases}$$

Assume also that the stochastic processes $\{X_i(t), t\geq 0\}$ $i=1, \ldots, n$, are mutually independent. Introduce

$$\mathbf{X}(t)=\{X_1(t), \ldots, X_n(t)\}$$

and let

$$\phi\{\mathbf{X}(t)\}=\begin{cases}1 & \text{if the system functions at time } t\\ 0 & \text{if the system is failed at time } t.\end{cases}$$

We also assume the structure function ϕ to be coherent. For an excellent introduction to coherent structure theory, we refer to Barlow & Proschan (1975).

The following random variables are of key interest when concentrating on system behaviour *after* a minimal repair.

X=remaining system lifetime just *after* the failure of the system, which, however, is immediately "black box" minimally repaired.

Y=remaining system lifetime just *after* the simultaneous failure of a component and the system. The *component* is, however, immediately "black box" minimally repaired.

Introduce the event:

$D_i=\{$the ith component and the system fail simultaneously$\}$

and define the random variable Y_i by

$$Y_i=\begin{cases}\text{remaining system lifetime just } \textit{after} \text{ the simultaneous failure of the } i\text{th component and the system. } \textit{This component} \text{ is, however, immediately "black box" minimally repaired} & \text{if } D_i \text{ occurs}\\ 0 & \text{otherwise.}\end{cases}$$

Assuming that the components have absolutely continuous life distributions, note that

$$P(Y_i=0)=\sum_{j\neq i} P[j\text{th component and the system fail simultaneously}]$$

A result in the Arjas & Norros (1989) *spirit* would be

$$P(Y>u)\leq P(X>u) \qquad \text{for all } u\geq 0 \tag{5}$$

For the time being we have neither been able to prove (5) nor to come up with a counter-example. However, our rather extensive computer searches for such a counter-example, after this meeting, make us conjecture that (5) does in fact hold. For a series system we have equality in (5). However, we do not always have

$$P(Y_i>u \mid Y_i>0)\leq P(X>u) \qquad \text{for all } u\geq 0. \tag{6}$$

Here $P(X>u)$ is the survival distribution of the system after a system failure and an immediate "black box" minimal repair of the system. $P(Y_i>u \mid Y_i>0)$ on the other hand is the conditional

survival distribution of the system based on the information that the ith component has "caused" system failure followed by an immediate, natural "black box" minimal repair of this component. For a parallel system of two independent components having exponentially distributed life lengths with failure rates λ_1 and λ_2,

$$P(X>u)>P(Y_1>u|Y_1>0)=\exp(-\lambda_1 u)$$

for all $u>0$ if $\lambda_1=\lambda_2=1$, whereas the strict inequality is reversed for all $u>0$ if $\lambda_1=1$ and $\lambda_2=3$. At least this shows that the somewhat surprising result of Arjas & Norros (1989) must be interpreted with care.

As a conclusion to my discussion I am very happy with the Finnish entrance on the reliability arena, looking forward to further work and further discussions.

Additional references

Arjas, E. & Norros, I. (1988). A martingale approach to reliability theory: on the role of filtration in the model specification. *Ann. Acad. Sci. Fenn., Ser. A, Mathematica* **13,** 183–189.

Arjas, E. & Norros, I. (1989). Change of life distribution via a hazard transformation: an inequality with application to minimal repair. *Math. Oper. Res.* **14,** 355–361.

Barlow, R. E. & Proschan, F. (1975). *Statistical theory of reliability and life testing. Probability models.* Holt, Rinehart and Winston, New York.

Bergman, B. (1985). On reliability theory and its applications. *Scand. J. Statist.* **12,** 1–41.

Funnemark, E. & Natvig, B. (1985). Bounds for the availabilities in a fixed time interval for multistate monotone systems. *Adv. Appl. Probab.* **17,** 638–665.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data.* Wiley, New York.

Natvig, B. (1985). Discussion of a paper by Bo Bergman. *Scand. J. Statist.* **12,** 33–37.

Natvig, B. (1988). On information based minimal repair and the reduction in remaining system lifetime due to the failure of a specific module. *J. Appl. Probab.* (submitted).

Norros, I. (1987). The "minimal repair", elimination and externalization of a totally inaccessible stopping time. Reports of the Department of Mathematics, University of Helsinki.

REPLY TO THE DISCUSSION

On the whole, I feel that the discussants have taken up questions which complement the paper very well. Being already quite long the way it stands, the paper has the major weakness that it discusses few concrete examples. Therefore, the reader might easily find himself lost and lacking motivation for careful study. With the aid of the discussants the gap to interesting applications has been made considerably narrower. My only regret is that there is no "Mr Mustard" among the discussants, saying that this is all rubbish. Some readers are bound to think so anyway, and they, too, would need support.

Niels Keiding

In Keiding's contribution the main points raised concern the role of time, which is beautifully illustrated by four Lexis diagrams, and heterogeneity. It is striking that four years ago, at the 10th Nordic Conference in Bolkesjø, as a discussant of Andersen & Borgan (1985), I was equally concerned about how time is used in the modelling. Although my views about time have not changed much in four years, I try not to repeat too much of what I wrote then (Arjas, 1985).

I tend to think that pictures with boxes and arrows, referring to states and transitions, are often more confusing than helpful. The reason is that, implicitly, it is implied that nothing happens to an individual as long as it remains in a box ("sojourn in a state"), and then there is a sudden transition into another box, a new state. The historical origins of such presentation are in Markov and Markov renewal models, in which the idea of states and sojourns is the starting point. I admit that simple point process models are not much different from these, and by employing two or more time-scales together one can extend from such traditional framework. However, complications can easily arise. For example, if a "Diseased" individual could be cured, should one draw an arrow back into the "Healthy" box? Doing this would suggest that any individual of a certain age, no matter *how many times* and *when* it had previously had the disease, would always have the same hazard of acquiring it again. If this does not sound a reasonable approximation, one would have to draw more boxes and arrows to the right, each of them adding potentially a new time-scale. How many boxes should one draw?

I think it is much more natural to use calendar time t, together with a pre-t history description H_t. This is not a restriction since all other time-scales are measurements corresponding to "how long ago some event (=marked point) occurred". Technically, this means that all time-scales can be expressed as simple differences, determined by the pair (t, H_t). Note also that all such time-scales advance at the same speed as calendar time. This corresponds to the 45° angle in the Lexis diagrams. In a very fundamental way, time is always "univariate".

Saying that all time-scales can be expressed in terms of the pair (t, H_t) is of course not the same thing as saying that only calendar time is relevant for statistical modelling. While there will always be martingales associated with the counting processes, both of which are indexed by calendar time (the martingale property is a part of the definition of the hazard process!), many important techniques for statistical *estimation* lead to expressions where time is indexed differently. The Nelson–Aalen estimators, which are based on duration (=time since entry), are in this category if the entry times differ from each other. Then it can be that there is no non-trivial way to define a history $\mathbf{H}$ with respect to which the considered expression would be an $\mathbf{H}$-martingale.

In the first of Keiding's examples (Fig. 2), I would proceed in the same manner as he does, considering the marked points due to patient entry (into the interim analysis) as non-innovative. This is because it is important that results from a confirmatory analysis are (conditionally) independent of what has been obtained previously. However, when considering *disease onset* in the second example (Fig. 3), unless all exposure and incidences are specifically restricted to the study period, I would be inclined to view patient entry as innovative: if a patient, who had contracted the disease before the beginning of the study, enters, the investigator will often know retrospectively the age at onset. A simple case to consider is where the individuals are independent and onset rate depends only on age. Then, and if there are no covariates, the contribution to the likelihood from a patient entry would simply be the density of the "age at onset" distribution.

In the case where *mortality of the diseased* is considered, patient entry could be innovative as well. If the time of onset becomes known at entry, each patient carries information about "the non-occurrence of death between onset and entry". However, this type of sampling is in most cases length biased: individuals who die quickly after onset are less likely to be included in the sample. Under fairly general stationarity conditions the contribution to the likelihood from a patient entry would be a factor of the form $\{1-F(d)\}/\mu_F$, where F is the distribution of the duration (from onset to death), μ_F is the corresponding mean, and d is the time from onset to entry.

The question of interval censoring relates closely to those put forward by Borgan, and I will try to answer them together, in the reply to Borgan.

Finally an attempt to understand, and to challenge, Westergaard's paradigm. My interpretation of what Westergaard means by "binomial law of error" is that he is writing about the probabilities for different individuals being equal, not so much of the outcomes being independent. This is because, in a large population, independence of the demographic events appears to be a relatively safe approximation.

Adopting this point of view, Westergaard seems to be implying that the probabilities for different individuals to experience the demographic event in question are not the same initially, but that they become so, or nearly so, after enough stratification. On the surface, this makes sense: individuals within a stratum tend to be more alike than individuals in different strata. From this I get the impression that Westergaard interprets probabilities, more or less, as physical characteristics of the individuals, or their "propensities towards the demographic event in question". Such an idea can be tenable in experiments like coin tossing where it is difficult to think what actual measurements on the coin, before it is tossed, could make *heads* more, or less, likely than *tails*. Therefore, the value 0.5 seems to capture something rather fundamental about every coin. However, I do not think that anything like this holds in demography: stratification with respect to demographic characteristics will not saturate itself. If the statistician would only briefly see the individuals in any one stratum, let alone interview them, he or she would quickly reject the idea that they have a common propensity, whatever this is taken to mean, to marry, to have a child, or to die. To me, speaking about "the number of active causes" as a well-defined concept seems more like a dream than real.

Another attempt to interpret Westergaard is to say that the probabilities approaching the binomial law are conditional and that the conditioning corresponds exactly to the degree of stratification. For example, we could speak of the probability that a male between ages 51 and 59 dies during a given year. But such stratum-specific probabilities must then apply for all individuals as long as they are characterized precisely by their membership in the stratum. In other words, whatever the degee of stratification, the binomial law holds *exactly* within each stratum, but with probabilities varying from stratum to stratum.

I am therefore bound to conclude that I cannot find a consistent but non-trivial way to interpret what Westergaard writes. To me, the issue is once again one between information and probabilities: to what extent the statistician wants to, or is able to, characterize the individuals, and to what extent he or she then uses probability statements to describe the remaining uncertainty about their behaviour. In demography, results of scientific and practical value can often be obtained if characteristics such as sex, age, and social class, are used, leaving the rest to be "random". But I fail to see anything canonical in such practice.

Ørnulf Borgan

I do indeed think that the marked point process framework is general enough to cover the two situations described by Borgan. This contrasts with the belief expressed, e.g. by Finkelstein (1986): "... the counting process framework cannot readily be adapted to the problem of interval-censored data, because it is difficult to define appropriately the increasing sequence of sigma-algebras. Thus, the development of statistical methodology which is appropriate for arbitrarily censored survival data requires a fundamentally different approach..." But, as Borgan correctly emphasizes "to work out the details ... is not at all simple".

In the first problem Borgan considers, the visits to a hospital could produce observed marked points of the form ("e" refers to examination)

$$\begin{cases} \{t, (e, 1)\}, & \text{corresponding to "examination made at time } t\text{, complication found"} \\ \{t, (e, 0)\}, & \text{corresponding to "examination made at time } t\text{, } \textit{no} \text{ complication found".} \end{cases}$$

We could then factor the corresponding mark-specific hazards as

$$\begin{cases} dA_t(e, 1)=dA_t(e)\cdot\varphi_t(1|e) \\ dA_t(e, 0)=dA_t(e)\cdot\varphi_t(0|e)=\mathrm{dA_t}(e)\{1-\varphi_t(1|e)\}, \end{cases}$$

where $\varphi_t(1|e)$ is the prevalence probability of the complication at time t, conditionally on the observed pre-t history. It is perhaps reasonable to assume that the timing of the hospital visits is non-innovative, in the sense that $dA_t(e)$ does not depend on the parameter of interest, but that the prevalence $\varphi_t(1|e)$ clearly should depend on such a parameter. Unfortunately $\varphi_t(1|e)$ will often have a complicated functional form; in general it will depend on the incidence rate of the complication and the mortality rates before and after the complication (cf. McKnight & Crowley, 1984).

In the second situation, which deals with induced tumorgenicity, sacrifice has a similar role as hospital visits in the first: the presence or absence of a complication (tumor) can be determined. The difference is, of course, that hospital visits of an individual can be repeated, but sacrifices cannot.

I found Borgan's suggestion concerning the estimation of the integrated incidence rate very interesting. But there may be an identifiability problem because the incidence rate is a necessary ingredient in estimating "the expected number of individuals with (or without) the complication". Instead of an essentially non-parametric Nelson–Aalen type estimator for the integrated rate, I would be inclined to try a parametric model for all hazards, and likelihood-based inference.

Per Kragh Andersen

Finding proper diagnostic methods for survival models is a very important, although somewhat problematic, task. Many methods in classical statistics are based on normal errors about a mean, and this additive structure has influenced strongly the ideas about what residuals are and how statistical models could be checked against data. An additive structure is sometimes considered in survival models as well: for example, in exponential and Weibull regression models the logarithm of the survival time is the sum of a linear expression and an error term which follows an extreme value distribution. However, I find it much more natural that the "randomness" in survival times is expressed in terms of hazards, and then the 1-exponentiality of the total hazards (generalized residuals) forms an invariance property on which the assessment of the model can be based.

But model checking is not straightforward; there are both technical and conceptual problems. A first technical problem in this context is caused by the censoring: a sample of "censored independent 1-exponential variables" does not have a uniquely defined distribution unless the censoring mechanism is specified. Even then, it would be awkward to consider statistics which depend on the censoring mechanism. The natural alternative is to employ the notion "total time on test" and to convert the censored 1-exponential variables into a stopped 1-Poisson process.

Another, and much more difficult technical problem is the behaviour of the plug-in estimators. For example, in the generalized residuals for Cox's regression models the "true" parameter $\boldsymbol{\beta}$ must be replaced by the partial MLE $\hat{\boldsymbol{\beta}}$. (In the case considered by Andersen, an estimate of the baseline hazard $\lambda_0(t)$ is also needed.) Although, disregarding censoring, each of the corresponding generalized residuals converges in distribution to a 1-exponential variable as the sample size increases, the corresponding large sample properties may not hold for

statistics which themselves depend on the sample size. An obvious and important example of this is the score, the logarithmic derivative of the (partial) likelihood: for the "true" $\boldsymbol{\beta}$ the score corresponding to the entire sample is asymptotically normal, with variance tending to infinity, whereas for $\hat{\boldsymbol{\beta}}$ it is identically equal to zero (by the definition of $\hat{\boldsymbol{\beta}}$). Recent results by Khmaladze (1988) may open up new ways to consider such problems asymptotically (Gill, 1988), but for small samples the prospects of finding distributionally invariant statistics with good diagnostic potential seem to me gloomy.

Finally a comment on a more general level. I see diagnostic methods primarily as a way for the statistician to check, and to convince the reader, that what he or she is inferring from the available data is reasonable in the problem at hand. I have therefore much more sympathy with explorative diagnostic methods than with rigid significance testing given a model hypothesis. Models are never correct, and accepting or rejecting them as true or false should not be the issue.

Bent Natvig

Bent Natvig's comments are so supportive that it would be sheer arrogance on my part to disagree with what he writes. I happily accept his judgement that I am "on the right [Bayesian] track".

I will only briefly comment on the notion of minimal repair, aiming to relate it to the more general problem of predicting a life length from the available information.

Let $\mu^{\mathbf{G}}=(\mu_t^{\mathbf{G}})$ be the prediction process concerning system lifetime S and based on $\mathbf{G}$, i.e. $\mu_t^{\mathbf{G}}(B)=P(S\in B|G_t)$ (see 3.1.1). Then, for $t>0$ and on $\{S\geqslant t\}$, the left limit $\mu_{t-}^{\mathbf{G}}$ a conditional probability distribution of S with support in $[t,\infty)$. Considering this distribution for $t=S$ gives $\mu_{S-}^{\mathbf{G}}$. This is the $\mathbf{G}$-prediction on S which was made immediately before the failure actually occurred. (Note that $\mu_s^{\mathbf{G}}$ is the degenerate distribution at S. Therefore the prediction process shrinks suddenly at $t=S$ to a point mass.)

Considering this distribution, with S as the new origin, shows how close the $\mathbf{G}$-prediction actually was at failure. This shifted distribution is (the left limit of) the $\mathbf{G}$-*residual prediction* process, which Natvig mentions, at S. It is also the distribution of "the extra lifetime gained in a black box minimal repair" (Norros, 1987), and can therefore replace the distribution of random variable X which was introduced by Natvig.

Similarly, one can consider the prediction process $\mu^{\mathbf{F}^N}=(\mu_t^{\mathbf{F}^N})$ *concerning* S, but based on $\mathbf{F}^N$, the component history. Then $\mu_{S-}^{\mathbf{F}^N}$, viewed from S onwards, is the $\mathbf{F}^N$-prediction at S concerning the remaining lifetime. It therefore corresponds to the distribution of Natvig's variable Y. Note that these prediction-based definitions are valid without any reference to actual repair being done on the system or on its parts, and without having assumed that the components are independent.

A reformulation of Natvig's conjecture (5) would now be that the $\mathbf{F}^N$-residual prediction at S is less, in the sense of stochastic order, than the corresponding $\mathbf{G}$-residual prediction. An intuitive way of expressing this is that component level monitoring gives a more accurate estimate of the remaining life than system level monitoring. This sounds quite reasonable, and, having got this far, it would be nice to provide a proof. Sadly, and in spite of having spent many hours on this, I have none. In order to finish the discussion with a positive suggestion, I formulate the following.

Exercise. Prove or disprove the above reformulation of Natvig's conjecture.

If successful, the reader has demonstrated that he or she has a more complete understanding of this material than the author or the discussant.

Additional references

Gill, R. D. (1988). Personal communication.

Khmaladze, E. V. (1988). An innovation approach to goodness-of-fit tests in R^n. *Ann. Statist.* **16,** 1503–1516.