Biometrika (1997), **84**, 3, *pp.* 505–522
Printed in Great Britain

On prequential model assessment in life history analysis

BY ELJA ARJAS AND DARIO GASBARRA
Rolf Nevanlinna Institute, University of Helsinki, P.O. Box 4, 00014 Helsinki, Finland
e-mail: elja.arjas@RNI.Helsinki.fi dario@solar.oulu.fi

SUMMARY

We extend the prequential approach introduced by A. P. Dawid to continuous time marked point processes. A prequential forecasting system is expressed by means of the stochastic intensities of the point process. The exact independence and 1-exponentiality of the derived compensators are used to assess the predictive performance of the model and the inferential procedure.

Some key words: Bayesian inference; Goodness of fit; Marked point process; Markov chain Monte Carlo; Prequential analysis.

1. INTRODUCTION

1·1. *General*

Goodness of fit or, more generally, model assessment is an important but also somewhat problematic aspect of the statistical analysis of life history data. With the development of increasingly sophisticated statistical models whose foundation is in the theory of counting processes and their associated conditional intensities, typically combined with non-parametric or semiparametric inferential procedures, classical goodness-of-fit tests have lost much of their appeal and practical value.

Instead, an entire new generation of methods has been developed for this task, which rest directly on the modern definition of the conditional intensity: a counting process minus the integrated intensity forms a martingale. Therefore, in a 'correctly specified' intensity model, the corresponding incremental differences should behave like white noise. For such techniques, see e.g. Wei (1984), Arjas (1988), Barlow & Prentice (1988), Therneau, Grambsch & Fleming (1990) and, for a review, Andersen et al. (1993, Ch. 4).

A problem with this approach, however, is that the exact martingale property holds, even if the statistical model is 'true', only if also the 'true' value of the parameter is used. However, when the maximum likelihood or partial maximum likelihood estimator is used as a plug-in in place of the true value, the corresponding intensity estimates are no longer predictable with respect to the observed filtration and the required martingale property is lost. This complicates the asymptotic theory leading to statistical tests considerably. For a full account of this, see Andersen et al. (1993).

Our inferential approach to survival analysis is Bayesian, and therefore plug-in values would in general be replaced by integrals with respect to the posterior. If the posterior is the one obtained from the entire dataset, the method has a similar defect as when plug-in values are used, however. In a sense, data are again used twice, once for the estimation and again in assessing the model. There is a way around this difficulty, however: one can use, at time t, the predictive intensity based on the pre-t data only. This, in fact, is nothing but the 'ordinary' martingale related notion of conditional intensity with respect to

observed histories. Therefore the observed differences between the counting processes and the corresponding integrated intensities, i.e. compensators, are exact martingales, by definition, with respect to the canonical probability which arises from the prior and the intensity model by integrating out the parameters. This idea is a direct extension into continuous time of the prequential analysis of a probabilistic model, put forward in a series of papers by A. P. Dawid and co-workers, e.g. Dawid (1984, 1992) and Seillier-Moiseiwitsch & Dawid (1993). In the case of a simple point process our results actually follow from those derived by Dawid if the interarrival times are viewed as data.

However, in general survival and life history models there is a considerable practical problem. Apart from some isolated special cases, where the updating of a posterior can be done analytically by applying conjugate distributions, the calculation of the data-based marginal intensities is very elaborate and has to be done numerically. This is even more problematic unless there is a way in which the numerical calculations could be done only once, and then considering the entire observation interval at the same time. Two approaches, both applying Markov chain Monte Carlo integration, are proposed to solve this problem.

This paper has much in common with Gelfand, Dey & Chang (1992), which also contains a large number of relevant references. In particular, both papers follow an exploratory approach to model assessment, using predictive distributions and, if necessary, Monte Carlo integration. However, there are also important differences: (i) our criteria for model assessment cannot be expressed in terms of expected values of 'checking functions'; and (ii) our approach is prequential, instead of building on cross-validation ideas. See, however, the comments on prequential model diagnostics by Raftery and Seillier-Moiseiwitsch in the discussion of Gelfand et al. (1992).

The contents of the paper are as follows. In the remainder of this section we first consider briefly the general notions and ideas behind prequential forecasting, and then apply them in the specific context of Bayesian inference from marked point process data. In § 2 we introduce some simple graphical methods and statistical tests for assessing the predictive performance of the postulated Bayesian model. In § 3 we consider in detail two examples, as illustrations of the methods proposed here. Then § 4 deals with computational aspects, and the paper concludes with some general remarks in § 5.

1·2. *Prequential analysis*

Here we recall briefly some ideas on discrete-time data sequences exposed in A. P. Dawid's papers. Let $\Omega = \{\omega = (X_n)_{n \in \mathbb{N}} : X_i \in E\}$, the space of sequences in some measurable space E. Suppose that at any time n a prequential forecaster has observed the values $x_1, \ldots, x_{n-1}$ of $X_1, \ldots, X_{n-1}$ and must issue a probability forecast P_n on the yet unobserved X_n. A prequential forecasting system is defined by a rule which for any n associates a choice of P_n under any possible set of outcomes $x_1, \ldots, x_{n-1}$ of $X_1, \ldots, X_{n-1}$. One obvious way to build a prequential forecasting system is to consider a joint probability distribution P on the space of sequences Ω and derive the conditional distributions P_n. On the other hand, any given prequential forecasting system is consistent with one and only one probability measure P on Ω. For example, a statistical forecasting system can be built by giving a statistical model and an inferential procedure. Once we have a realisation $\{x_n\}$ of the random sequence $\{X_n\}$ we would like to assess the adequacy of P as a probabilistic explanation of the observations. Such assessment should depend on P only through the sequence of forecasts $\{P_n\}$ that were in fact made by using the prequential forecasting system. This fundamental requirement was called by A. P. Dawid the Prequential Principle.

In the following we extend this approach to continuous-time marked point processes. Of course, a marked point process can be viewed as a discrete sequence of time epochs and marks imbedded in continuous time. Nevertheless, in the proper continuous-time setting it is natural to build the prequential forecasting system by using conditional intensities.

1·3. *Marked point processes*

We formulate the inferential problem in terms of a marked point process. Suppose that the data are of the form $\{(T_n, X_n)\}$, where $0 \equiv T_0 < T_1 < \ldots < T_N$ are the observed times of occurrence and X_n is a description of the event which occurred at T_n. For simplicity, the marks X_n are assumed to take values in a countable set E. We denote by H_t the pre-t history

$$H_t = \{(T_n, X_n) : T_n \leqslant t\}, \tag{1·1}$$

consisting of the events in the data which occurred before and including time t, and by H_{t-} the corresponding history when the inequality in (1·1) is strict. The internal filtration of the process is denoted by $(\mathscr{H}_t)$ with $\mathscr{H}_t = \sigma\{H_t\}$.

It is well known that there is a one-to-one correspondence between probabilities on the canonical path space Ω of the marked point process, consisting of the sequences $\omega = \{(t_n, x_n) : n \geqslant 0\}$ such that $x_n \in E$, $0 \equiv t_0 \leqslant t_1 \leqslant \ldots$, and $t_n < \infty$ implies $t_n < t_{n+1}$, on the one hand, and on the distribution of the initial mark X_0 and the $(\mathscr{H}_t)$-compensators of the counting processes

$$N_t(x) = \sum_{T_n \leqslant t} 1_{\{T_n \leqslant t, X_n = x\}}(t) \quad (t \geqslant 0, x \in E) \tag{1·2}$$

on the other. For simplicity, we consider here only absolutely continuous compensators, in which case we can express everything in terms of conditional intensities.

In order to tie this framework explicitly with statistical inference, we introduce into the notation an unknown parameter θ. It can be viewed as an initial unobserved mark X_0 of the marked point process at time $T_0 = 0$. The parameter space Θ may be finite dimensional real or abstract, but its role will always be the same: for any given value $\theta \in \Theta$ we assume that there is a corresponding P^θ defined on the path space Ω. In practice, as noted above, P^θ is specified most conveniently in terms of an initial distribution for X_0 and the corresponding conditional intensities, $\lambda_t^\theta(x)$ say, $x \in E$, $t \geqslant 0$. In classical statistical inference, the family $M = \{P^\theta : \theta \in \Theta\}$ is called a statistical model of the marked point process.

1·4. *A Bayesian forecasting system*

In Bayesian inference both the parameters θ and the process sample path H_∞ are viewed as random elements. The joint distribution of (θ, H_∞) is then determined by probabilities of the form

$$\mathrm{pr}(\theta \in A, H_\infty \in B) = P_{(\theta, H_\infty)}(A \times B) = \int_A P^\theta(B)\pi(d\theta), \tag{1·3}$$

where $\pi(\theta)$ is the prior distribution for θ.

It is characteristic of Bayesian inference that one is thinking of the dynamical prediction problem in time, where the prediction at time t always concerns the unobserved future sample path, say $H_{(t,\infty)}$, and the predictions are updated continuously on the basis of the observed H_t. At time $t = 0$, the predictive distribution is simply the marginal of H_∞,

obtained from (1·3) by letting $A = \Theta$. We denote this probability by P and the corresponding expectation by E. The updating of P corresponds to viewing H_∞ as a pair $(H_t, H_{(t,\infty)})$ and conditioning the joint distribution (1·3) on H_t, again integrating out the parameter θ. This continuous updating of predictions, which in practice is done by applying Bayes' formula on the corresponding posterior distributions for θ, is at the heart of our method for model assessment.

In more technical terms, this corresponds to focusing interest, not on θ and the $(P^\theta, \mathscr{H}_t)$-intensities $\lambda_t^\theta(x)$, but on the predictive $(P, \mathscr{H}_t)$-intensities $\hat{\lambda}_t(x)$. It is well known that these two types of intensity are linked by averaging with respect to the posterior $\pi(d\theta \mid H_{t-})$, that is,

$$\hat{\lambda}_t(x) = E\{\lambda_t^\theta(x) \mid H_{t-}\}. \tag{1·4}$$

However, the real advantage of the predictive intensities is that we can use the existing theory of point processes and martingales, without needing to condition on something unknown. In other words, with this change from $M = \{P^\theta;\ \theta \in \Theta\}$ to P we have eliminated all technical problems arising from an 'unknown parameter θ', and are actually mimicking the learning process of a Bayesian statistician working with the postulated model (1·3). Prequential model assessment means that one is considering these predictions sequentially and comparing them with the actual observed development of the marked point process.

There are several ways of doing this. An obvious one is to use the defining property of the conditional intensities, according to which the differences $N_t(x) - \int_0^t \hat{\lambda}_s(x)\, ds$ are $(P, \mathscr{H}_t)$-martingales. A more intuitive way of writing this is based on the well-known interpretation

$$\hat{\lambda}_t(x)\, dt = P\{dN_t(x) = 1 \mid H_{t-}\};$$

the differentials

$$dN_t(x) - P\{dN_t(x) = 1 \mid H_{t-}\}$$

are then expressing a direct continuous time analogue of the prequential method of Dawid (1984). However, there are also more refined results, which lead to exact reference distributions and statistical tests.

For the sake of simplicity, we start by considering a particular mark $x \in E$ and the corresponding sequence of precise times at which x occurs, say

$$\tau_0^x \equiv 0, \quad \tau_{k+1}^x = \inf\{T_n > \tau_k^x;\ X_n = x\}.$$

We have then the following well-known result.

PROPOSITION 1. *The spacings* $\hat{\Lambda}_{\tau_{k+1}^x}(x) - \hat{\Lambda}_{\tau_k^x}(x)$ $(k = 0, 1, 2, \ldots)$ *of the* $(P, \mathscr{H}_t)$*-compensator* $\hat{\Lambda}_t(x) = \int_0^t \hat{\lambda}_s(x)\, ds$ *form a sequence of independent* 1*-exponential random variables.*

An equivalent alternative formulation of this result is that the time-transformed counting process $\{\hat{N}_u(x);\ u \geqslant 0\}$, defined by

$$\hat{N}_u(x) = N_{\hat{\Lambda}_u^{-1}(x)}(x), \tag{1·5}$$

is a Poisson process with fixed intensity 1. Thus we can view the statistician as monitoring the occurrences of mark x in time and matching the integrated intensities, obtained from the P-distribution and updated from the previous observations, against the yardstick of spacings of the Poisson (1)-process. Yet another equivalent way is to consider the residuals

$$1 - \exp\{\hat{\Lambda}_{\tau_k^x}(x) - \hat{\Lambda}_{\tau_{k+1}^x}(x)\},$$

which, according to P, are independent and follow the $[0, 1]$-uniform distribution. These

latter random variables have an obvious interpretation in terms of observed residuals of the spacings $\tau_{k+1}^x - \tau_k^x$, obtained sequentially from P and the data. In fact, in this simple case of a univariate process we could just as well denote $\tau_{k+1}^x - \tau_k^x$ by η_{k+1} and consider the process $\{\eta_k : k \geqslant 1\}$ of a discrete time parameter. This would correspond exactly to the original formulation of the prequential method in Dawid (1984).

However, the 1-exponentiality and independence properties hold in much more generality than that contained in Proposition 1; see, for example Aalen & Hoem (1978) and Norros (1986). It is somewhat difficult to give a formulation that would both be concise and cover all cases which could arise in an assessment of life history models. The following statement, which is a direct corollary of Theorem 2.1 in Norros (1986), appears to be sufficiently general for most purposes.

PROPOSITION 2. *Consider an arbitrary finite collection* $\{\tau_i;\ 1 \leqslant i \leqslant k\}$ *of* $(\mathscr{H}_t)$*-stopping times such that*

(i) $P(0 < \tau_i < \infty) = 1$ *for all* $1 \leqslant i \leqslant k$,

(ii) $P(\tau_i = \tau_j) = 0$ *whenever* $i \neq j$,

(iii) *each* τ_i *is completely unpredictable in the sense that the corresponding* $(P, \mathscr{H}_t)$*-compensator* $(\hat{\Lambda}_t^i)$ *of* τ_i *is continuous.*

Then the random variables $\hat{\Lambda}_{\tau_i}^i$ $(1 \leqslant i \leqslant k)$ *are independent and* 1*-exponential.*

Again, as an alternative, we can say that the variables $1 - \exp(-\hat{\Lambda}_{\tau_i}^i)$ are independent and (0, 1)-uniform.

The crucial element of Proposition 2 is the independence statement, which holds in great generality even if the stopping times τ_i themselves are dependent and, for example, their order of occurrence is not specified in advance, as in Proposition 1.

Note that, as was mentioned at the beginning of this section, given a nonnegative and bounded $\{\mathscr{H}_t\}$-predictable process $\lambda(x, t, H_{t-})$, one can always construct a probability measure P, to be used as a prequential forecasting system, on the canonical space of marked point process histories Ω, such that, under P, $\{N_t(x)\}$ has $(P, \mathscr{H}_t)$-intensity which coincides with $\lambda(x, t, H_{t-})$ (Brémaud, 1981, IV T4, p. 168).

In particular, Propositions 1 and 2 hold for such a constructed P. Therefore we are not necessarily restricted to Bayesian estimation and prediction. For example, suppose that we have specified a parametric model for the marked point process through a family of $(\mathscr{H}_t)$-intensities $\lambda(x, t, \theta, H_{t-})$, for $\theta \in \Theta$. We could then use a point estimate, such as the maximum likelihood estimate, $\hat{\theta}(t, H_{t-})$, for $\theta \in \Theta$ at time t, and finally form another probability on $\mathscr{H}$ by means of the plug-in intensity function

$$\hat{\lambda}(t, x, H_{t-}) = \lambda\{t, x, \hat{\theta}(t, H_{t-}), H_{t-}\}.$$

A slightly different frequentist idea would be to use the confidence intervals around the maximum likelihood estimate to account for the uncertainty regarding the true value of θ. If the confidence intervals at time t were based on the normal distribution with mean $\hat{\theta}(t, H_{t-})$ and variance $\sigma^2(t, H_{t-})$, we could then build a prequential forecasting system on the intensity

$$\tilde{\lambda}(t, x, H_{t-}) = \frac{1}{\sqrt{\{2\pi\hat{\sigma}^2(t, H_{t-})\}}} \int_{\mathbb{R}} \lambda(t, x, H_{t-}, \theta) \exp\left[-\frac{\{\theta - \hat{\theta}(t, H_{t-})\}^2}{2\hat{\sigma}^2(t, H_{t-})} \right] d\theta.$$

Actually, a statistician may use whatever ingredient he or she likes to define λ, such as Bayes' formula, maximum likelihood, nonparametric methods, astrology and so on. In any case, when a predictable intensity process is given, a nice probability measure on the space of histories is automatically defined.

2. MODEL ASSESSMENT

So far, we have derived from a possibly highly structured marked point process a simple standard Poisson process. The assessment of the predictive performance of such a model is reduced to checking the fit of the corresponding derived process sample path to the standard Poisson process assumption. There is a huge literature about tests of the Poisson process hypothesis. Here we give some simple tools that we found to be sensible.

A good first stage is to do a graphical check, looking at the total-time-on-test plot, see e.g. Andersen et al. (1993, p. 450). The plot process, $n \to S_n = \sum_{i \leqslant n} \hat{\Lambda}^i_{\tau_i}$, under the Bayesian forecasting system, has independent 1-exponential increments.

Asymptotically, the Kolmogorov law of the iterated logarithm gives a sharp result on the behaviour of the random walk process $\{S_n - n\}$: with probability 1,

$$\limsup \frac{S_n - n}{\sqrt{(2n \log\log n)}} = +1, \quad \liminf \frac{S_n - n}{\sqrt{(2n \log\log n)}} = -1. \tag{2·1}$$

Note that this result is true for any random walk with independent and identically distributed increments of mean 0 and variance 1. In other words, with probability 1, for any $\varepsilon > 0$,

$$n - (1+\varepsilon)\sqrt{(2n \log\log n)} \leqslant S_n \leqslant n + (1+\varepsilon)\sqrt{(2n \log\log n)}$$

will hold eventually for all n, and $\{S_n\}$ will satisfy the inequalities

$$S_n \geqslant n + (1-\varepsilon)\sqrt{(2n \log\log n)}, \quad S_n \leqslant n - (1-\varepsilon)\sqrt{(2n \log\log n)} \tag{2·2}$$

infinitely often. Of course this result does not apply as such in finite samples. Nevertheless, if $\{S_n\}$ infringes the boundaries and does not return, this can be used as evidence against the model.

As an attempt to quantify the evidence for or against the model we consider some functionals of the whole sample path $\{S_n\}$. We observe that S_n with respect to P is gamma distributed with shape parameter n and scale parameter 1. Denoting the corresponding cumulative distribution by $F_\gamma(.; n, 1)$ we find that the random variables $G_n = F_\gamma(S_n; n, 1)$ follow the $[0, 1]$-uniform distribution. Note that the increments of $\{G_n\}$ are not independent, nor is $\{G_n\}$ a martingale. Obvious test statistics are $G_{1n} = \max_{k \leqslant n} G_k$ and $G_{2n} = \min_{k \leqslant n} G_k$. By computing the reference distributions of G_{1N} and G_{2N} we can assign p-values to the whole sample, that is $P(G_{1N} \geqslant g_{1N})$ and $P(G_{2N} \leqslant g_{2N})$ where $g_{1N} = \max_{k \leqslant N} g_k$, $g_{2N} = \min_{k \leqslant N} g_k$ are the observed values of the statistics. We call these statistics prequential p-values. As usual, p-values close to 0 would be used as evidence against the model P.

In principle, the distribution of G_{1n} could be computed recursively. Here approximate prequential p-values were determined by a simple Monte Carlo method, by generating independent identically distributed samples of the process $\{G_n\}$.

3. TWO EXAMPLES

3·1. *General*

We consider two simple illustrative examples. In the first example the data come from a point process, where there is a strong correlation between successive spacings. We show how our diagnostic method can be used to reveal the inadequacy of a statistical model in which such correlation has not been accounted for. In the second example we consider a sample of right-censored survival times from a mixture of two exponential distributions

with different parameters, and demonstrate how an attempt to fit a single exponential distribution to such data leads to an obvious rejection of the model.

3·2. *Example* 1: *Point process with serially correlated spacings*

Consider a simple point process $0 < T_1 < T_2, \ldots$, denoting the spacings by $\eta_n = T_n - T_{n-1}$, with $T_0 \equiv 0$. We look at two competing Bayesian models for such data.

Model M_0. Suppose that (i) the model parameter θ is a real-valued random variable with prior distribution $F_\gamma(.; \alpha, \beta)$, where α is the shape parameter and β is the scale parameter, and that (ii) conditionally on θ, the spacings $\{\eta_n; n \geqslant 1\}$ are independent and θ-exponentially distributed. In other words, $\{T_n\}$ is a doubly stochastic Poisson process, or Cox process, with conditional intensity given by $\lambda_t^\theta \equiv \theta$.

Model M_1. Suppose that (i) θ is as in model M_0 above, but that (ii), conditionally on $(\theta, \eta_1, \eta_2, \ldots, \eta_{n-1})$, the spacings η_n are distributed according to the exponential distribution with parameter θ/η_{n-1}. The conditional intensity is now given by $\lambda_t^\theta = \theta/\eta_{N_t -}$. According to model M_1, long (short) spacings are typically followed by long (short) spacings, and therefore the points T_n tend to be clustered.

Denote by P_i the probabilities induced by the respective models M_i $(i = 0, 1)$ on the space $\Theta \times \Omega$. Due to the simplicity of the models we can find analytic expressions for the $(P_0, \mathscr{H}_t)$-compensators. Under M_0, by a well-known conjugacy property of the gamma distribution, the predictive $(P_0, \mathscr{H}_t)$-intensity is given by

$$\tilde{\lambda}_t^0 = E_0(\theta \mid H_{t-}) = \frac{\alpha + N_{t-}}{\beta + t},$$

so that the corresponding cumulative intensity from the interval $(T_{n-1}, T_n]$ is

$$\int_{T_{n-1}}^{T_n} \tilde{\lambda}_s^0 \, ds = (\alpha + n - 1) \log\left(\frac{\beta + T_n}{\beta + T_{n-1}}\right).$$

Under M_1, an analogous computation gives the cumulative (P_1, H_t)-intensity

$$\int_{T_{n-1}}^{T_n} \tilde{\lambda}_s^1 \, ds = (\alpha + n - 1) \log\left\{\left(\beta + \sum_{i=1}^{n} \frac{\eta_i}{\eta_{i-1}}\right) \Big/ \left(\beta + \sum_{i=1}^{n-1} \frac{\eta_i}{\eta_{i-1}}\right)\right\}.$$

Mixture model $M_{0,1}$. A third model, which includes both M_0 and M_1 as special cases and which in principle permits consideration of model selection probabilities adapting to the data, would have the following mixture form. Let $q \in [0, 1]$ be given and let ξ be a $\{0, 1\}$-valued random variable with

$$P_{0,1}(\xi = 0) = 1 - q, \quad P_{0,1}(\xi = 1) = q.$$

Then define the probability $P_{0,1}$ on $\Theta \times \Omega \times \{0, 1\}$ by specifying the conditional probability $P_{0,1}(.\,|\,\xi)$ through $P_{0,1}(.\,|\,\xi) = \xi P_1(.) + (1 - \xi) P_0(.)$.

Define the filtration $\mathscr{G}_t = \mathscr{H}_t \vee \sigma(\xi)$. Then the $(P_{0,1}, \mathscr{G}_t)$-intensity of the counting process $\{N_t\}$ is $\tilde{\lambda}_t^\xi$. The $(P_{0,1}, \mathscr{H}_t)$-intensity is the mixture

$$\tilde{\lambda}_t^{01} = P_{0,1}(\xi = 1 \mid H_{t-}) \tilde{\lambda}_t^1 + P_{0,1}(\xi = 0 \mid H_{t-}) \tilde{\lambda}_t^0,$$

where, from Bayes' formula,

$$P_{0,1}(\xi = 1 \mid H_{t-}) = \frac{q Z_1(t)}{q Z_1(t) + (1 - q) Z_0(t)},$$

and

$$Z_\xi(t) = p_\xi(H_{t-}) = \int_{\Theta_\xi} p_\xi(H_{t-} \mid \theta_\xi)\, d\pi(\theta_\xi) \quad (\xi = 0, 1)$$

are the normalising constants. These quantities can be computed explicitly, leading to the expression

$$P_{0,1}(\xi = 1 \mid H_{t-}) = R/\{R + (1-q)(\beta + t)^{-\alpha - N_{t-}}\},$$

where

$$R = q\left(\prod_{i=1}^{N_{t-}-1} \eta_i^{-1}\right)\left(\beta + \sum_{i=1}^{N_{t-}} \frac{\eta_i}{\eta_{i-1}} + \frac{t - T_{N_{t-}}}{\eta N_{t-}}\right)^{-\alpha - N_{t-}}.$$

The $(P_{0,1}, \mathscr{H}_t)$-compensators can now be evaluated numerically.

A sample path segment of the process $\{T_n\}$ consisting of 500 points was generated by a computer from model M_1, with $\theta = 0{\cdot}6$. The hyperparameters were given the values $\alpha = 0{\cdot}1$ and $\beta = 0{\cdot}001$. This prior has a very large mean $\alpha/\beta = 100$, compared to the true value of 0·6, but it is also very flat, having variance $\alpha/\beta^2 = 10^5$. Our general experience with different priors is that, as long as they are reasonably uninformative, the diagnostics show little change. Typically, the differences were meaningful only for the predictions concerning the first few observations.

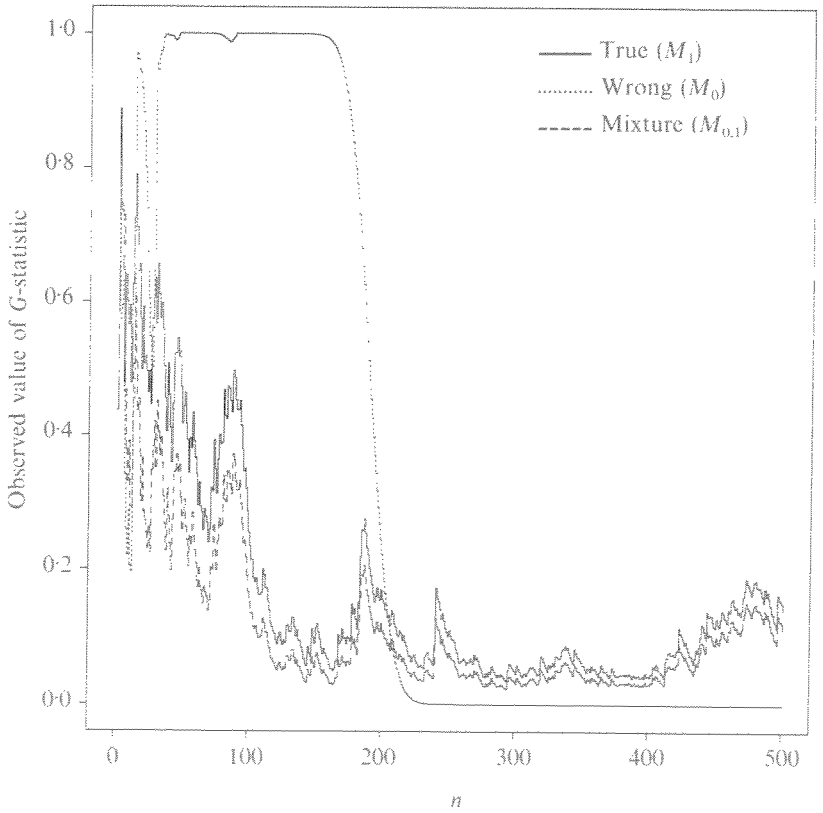

Fig. 1. Example 1: Observed sample path of the process $\{G_n\}$ under models M_0, $M_{0,1}$ and M_1; n, number of failures.

Figure 1 shows the corresponding sample paths of the process $\{G_n\}$ under the models M_0, M_1 and $M_{0,1}$. We can say straightaway that M_0 does not fit the data; $\{G_n\}$ appears to be oscillating between 0 and 1. Under M_1 and $M_{0,1}$, $\{G_n\}$ behaves nicely, and the sample paths are close to each other. This is predictable because, for data from model M_1, $P_{0,1}(\xi = 1 \mid H_t)$ converges to 1 almost surely according to P_1. The prequential p-values of the generated data under these three different models are shown in Table 1. The values of these statistics clearly reinforce the graphical conclusions. The same message is conveyed by the total-time-on-test plots displayed in Fig. 2, where we plotted the compensators $\tilde{\Lambda}^k_{T_k}$ against k ($k = 1, \ldots, 500$). Again one can see that, while the process $\{S_n - n\}$ has a nice random walk behaviour under M_1 and $M_{0,1}$, graphically coinciding in Fig. 2, staying mostly in the region prescribed by the law of the iterated logarithm, under model M_0 it leaves this region abruptly.

Table 1. *Prequential p-values for Example* 1

	M_1	M_0	$M_{0,1}$
$P(G_{1,500} \geqslant g_{1,500})$	0·594	≃0	0·629
$P(G_{2,500} \leqslant g_{2,500})$	0·337	≃0	0·253

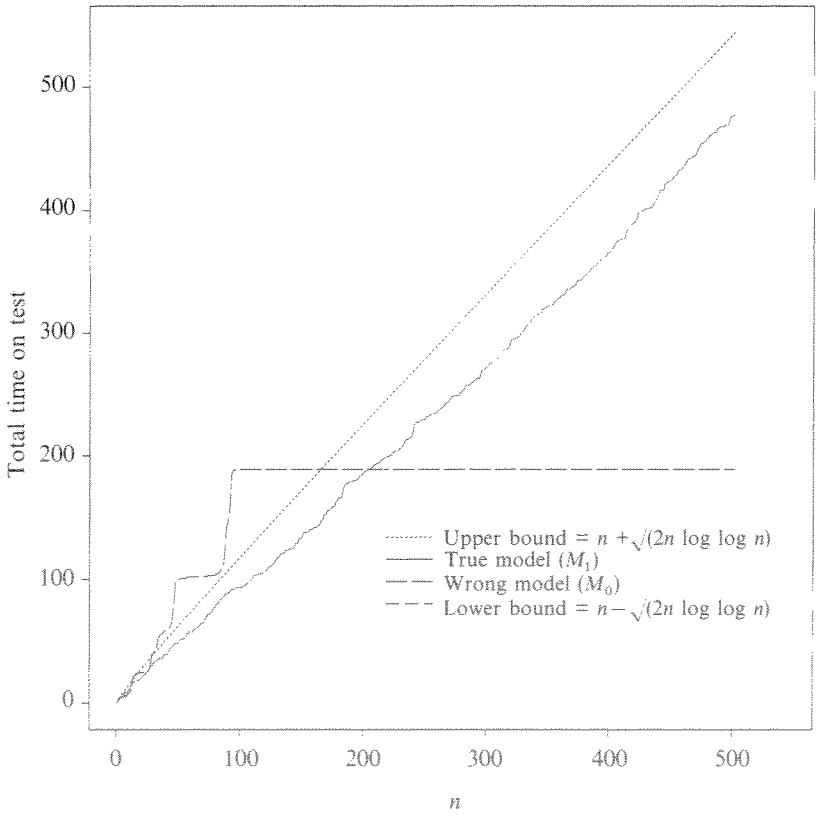

Fig. 2. Example 1: Total-time-on-test plot for the compensators $\tilde{\Lambda}^k_{T_k}$. The bounds $n \pm \sqrt{(2n \log \log n)}$ arising from the law of the iterated logarithm are shown; n, number of failures.

3·3. *Example 2: One-sample versus two-sample model*

Consider a right-censored sample of lifetimes

$$\{(X_i^*, \delta_i);\ X_i^* \geqslant 0,\ \delta_i = 0, 1,\ i = 1, \ldots, n\}.$$

Here X_i^* is the time the ith individual was last seen, and δ_i is the censoring indicator, that is $\delta_i = 1$ if the ith individual was dead at X_i^*, and $\delta_i = 0$ if it was censored. Denote the corresponding 'complete' lifetimes by X_i, that is $X_i = X_i^*$ if $\delta_i = 1$, and $X_i > X_i^*$ if $\delta_i = 0$.

We consider two following competing parametric models.

Model M_0. (i) The parameter of the model θ is a real-valued random variable with prior distribution $F_\gamma(.;\alpha,\beta)$ with given shape and scale parameters α and β, and (ii) conditionally on θ, the X_i are independent and identically distributed as θ-exponential. The model assumes also that the censoring mechanism is noninformative with respect to θ.

Model M_1. The individuals are divided into two subsamples, say into 'men' and 'women'. This leads to a marked point process formulation, with two observable marks m and w, say, and corresponding mark-specific hazard rates $\lambda_t^\theta(m)$ and $\lambda_t^\theta(w)$. These hazard rates are now modelled exactly as in model M_0, assuming independence with respect to the prior. We let $\theta = (\theta', \theta'')$, $\lambda_t^\theta(m) = \theta'$ and $\lambda_t^\theta(w) = \theta''$, with θ' and θ'' independent and distributed according to $F_\gamma(.;\alpha,\beta)$.

Let

$$N_t = \sum_{i=1}^{n} 1_{\{X_i^* \leqslant t, \delta_i = 1\}}(t) = \#\{\text{recorded deaths by time } t\},$$

$$R_t = n - \sum_{i=1}^{n} 1_{\{X_i^* < t\}}(t) = \#\{\text{individuals at risk at time } t\},$$

and define analogously N_t', R_t' and N_t'', R_t'' for the subsamples consisting of men and women. Denote by P_0 and P_1 the probabilities on the space $\Theta \times \Omega$ corresponding to the models M_0 and M_1.

Now, under P_0, the intensities corresponding to the 'individual counting processes' $N_t(i) = 1_{\{X_i^* \leqslant t, \delta_i = 1\}}$ are given by

$$\tilde{\lambda}_t(i) = E_0(\theta \mid H_{t-})\, 1_{[0, X_i^*]}(t) = \frac{\alpha + N_{t-}}{\alpha + \int_0^t R_s\, ds}\, 1_{[0, X_i^*]}(t).$$

The variables $\tilde{\Lambda}_{X_i^*}(i) = \int_0^{X_i^*} \tilde{\lambda}_s(i)\, ds$ form a, possibly right-censored, simple random sample of 1-exponential random variables. Similarly, the counting process $\{N_t\}$ is compensated by $\tilde{\Lambda}_t = \sum_{i=1}^n \tilde{\Lambda}_t(i)$. Under P_1, similar results will hold for both subgroups. As a result of the assumed prior independence of θ' and θ'', the two groups can be treated separately.

A sample of 500 lifetimes, 250 men and 250 women, was generated from model M_1. The true values of the intensity were $\theta' = 1{\cdot}2$ for men, and $\theta'' = 0{\cdot}8$ for women. These lifetimes were censored from the right by an independent random censoring mechanism, resulting in 46 censored observations for men and 75 for women. The Kaplan–Meier curves for males and females, although not shown here, clearly reflect the poorer survival prospects for males.

We computed the prequential forecasts for this two-sample dataset under both models, the true M_1 and the misspecified M_0. In both models the hyperparameters of the priors

were given the values $\alpha = \beta = 2$. We considered the stopping times

$$\tau'_h = \inf\{t : N'_t = h\} \quad (h = 1, \ldots, 250), \quad \tau''_k = \inf\{t : N''_t = k\} \quad (k = 1, \ldots, 250),$$
$$\tau_l = \inf\{t : N_t = l\} \quad (l = 1, \ldots, 500),$$

with the convention that $\inf\{\varnothing\} = +\infty$. Thus, for instance, τ'_h is the time it takes until h men have died. For the particular censored sample considered here, the largest finite values in these sequences were τ'_{204}, τ''_{175} and τ_{379}.

The compensators of these stopping times admit simple analytic forms. For example, under P_0 the intensity of τ_l is given by

$$R_t(\alpha + N_{t-})\left(\beta + \int_0^t R_s\, ds\right)^{-1} 1_{(\tau_{l-1}, \tau_l]}(t);$$

note that this intensity vanishes outside the interval $(\tau_{l-1}, \tau_l]$. Under P_1, by Proposition 2, the random variables $\Lambda_{\tau'_h}$ and $\Lambda_{\tau''_k}$ corresponding to the stopping times of τ'_h and τ''_k are independent and 1-exponential. Also the variables Λ_{τ_l} corresponding to τ_l are independent and 1-exponential under P_1, although dependent on $\Lambda_{\tau'_h}$, $\Lambda_{\tau''_k}$.

The advantage of considering these stopping times, rather than simply the individual survival times X_i^*, is that, as the individual observations are censored their cumulative intensities will be censored too, whereas the stopping times τ_k are not censored as long as there are at least k recorded failure times.

In Fig. 3 are shown the process-statistics $G'_n = F_\gamma(S'_n;\, n,\, 1)$, where $S'_n = \sum_{k=1}^n \Lambda_{\tau'_k}$, and the

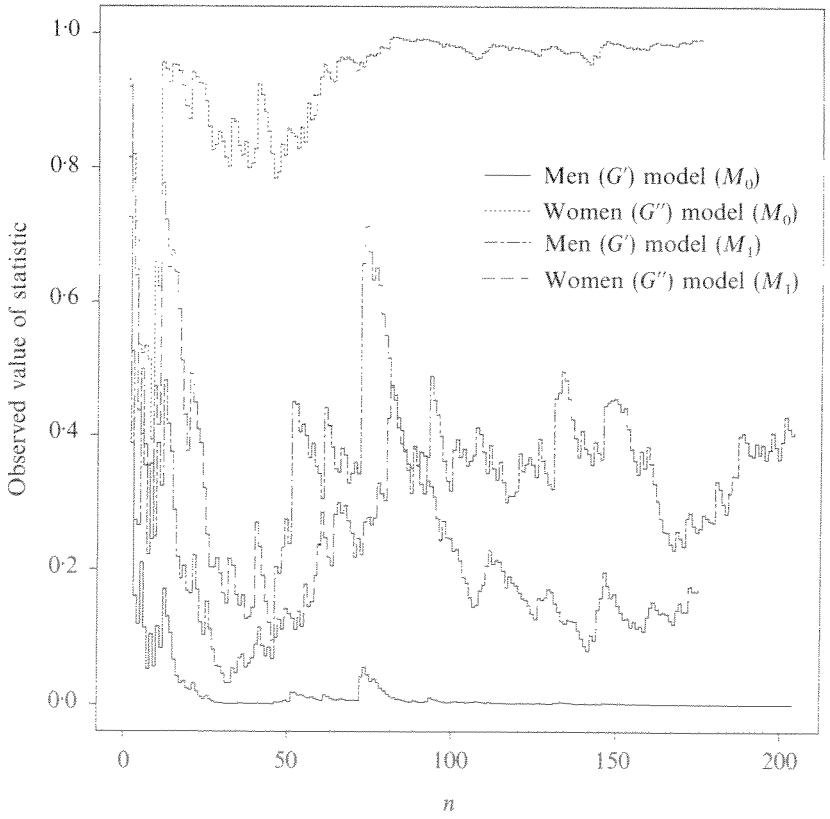

Fig. 3. Example 2: G'_n and G''_n statistics for the men's and women's subsamples under models M_0 and M_1; n, number of deaths.

Table 2. *Prequential p-values for Example 2*

(a) Under model M_1

Men	Women
$P_1(G'_{1,204} \geq g'_{1,204}) = 0{\cdot}700$	$P_1(G''_{1,175} \geq g''_{1,175}) = 0{\cdot}434$
$P_1(G'_{2,204} \leq g'_{2,204}) = 0{\cdot}281$	$P_1(G''_{2,175} \geq g''_{2,175}) = 0{\cdot}484$

(b) Under model M_0

Men	Women
$P_0(G'_{1,204} \geq g'_{1,204}) = 0{\cdot}821$	$P_0(G''_{1,175} \geq g''_{1,175}) = 0{\cdot}062$
$P_0(G'_{2,204} \leq g'_{2,204}) = 0{\cdot}003$	$P_0(G''_{2,175} \geq g''_{2,175}) = 0{\cdot}920$

(c) For the combined sample

	M_1	M_0
$P(G_{1,379} \geq g_{1,379})$	0·279	0·460
$P(G_{2,379} \leq g_{2,379})$	0·558	0·457

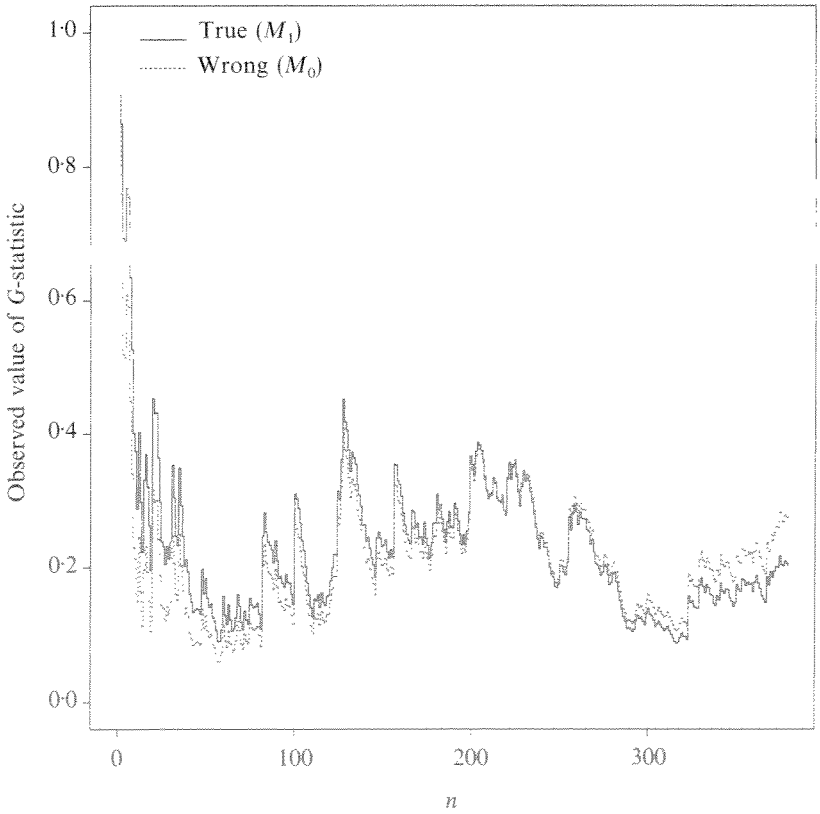

Fig. 4. Example 2: G_n statistics for the combined sample under models M_0 and M_1; n, number of deaths.

analogously defined $\{G_n''\}$, under both models M_1 and M_0, keeping the same observations previously generated under P_1. The interpretation is that we are using model M_1 to give separate forecasts for the next failure time within the male and the female subgroups. The prequential p-values under model P_1 are given in Table 2(a). From Fig. 3 it is clear also that the predictive performance of model M_0 is poor: the sample paths of the processes $\{G_n'\}$ and $\{G_n''\}$ appear to be converging respectively to 0 and to 1, whereas, if the model were correct, G_n' and G_n'' would be $[0, 1]$-uniformly distributed for each n. The intuition behind these tests is that in the data men's lifetimes tend to be shorter than expected from model M_0, and women's lifetimes longer. A statistician who has postulated model M_0 and is monitoring these statistics will soon become puzzled. This 'surprise' may be quantified by the prequential p-values under model P_0. The numerical values are in Table 2(b).

In Fig. 4 the sample paths of the process $\{G_n\}$ are shown, corresponding to the stopping times τ_k, under both models M_1 and M_0. Here we are forecasting the next failure time in the combined sample, i.e. regardless of sex. At the beginning the two models give almost the same forecasts for τ_k, and both seem to predict successfully. The reason for this is that, as long as the proportions of men and women among the survivors are similar, the opposite effects of the incorrect model specification cancel out. Only later, when the survivors' group is dominated by women, do the two models give different forecasts for the next failure time. The prequential p-values are displayed in Table 2(c), and show

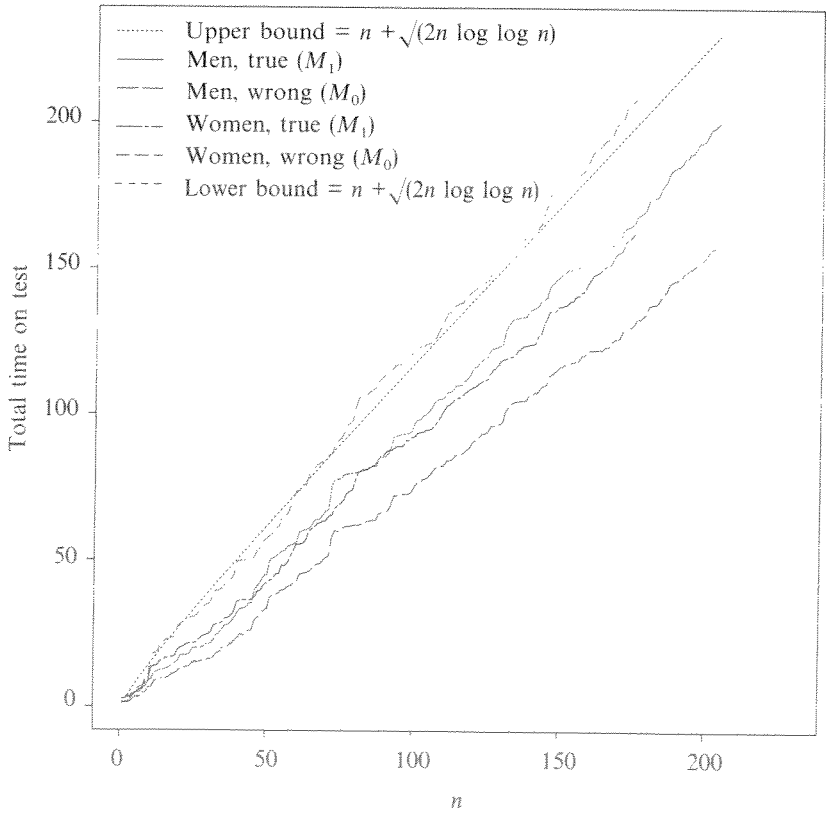

Fig. 5. Total-time-on-test plots of the compensators of the stopping times τ_k' and τ_k'' respectively for the men's and women's subsamples under models M_0 and M_1. The bounds $n \pm \sqrt{(2n \log \log n)}$ arising from the law of the iterated logarithm are shown: n, number of deaths.

that, unless we divide the sample into the two subsamples, our diagnostics do not have enough power to distinguish between the true and the misspecified models. Therefore the choice of the stopping time sequence is crucial for this model checking procedure; a model may give good predictions for some particular sequence of stopping times, but fail completely for another one.

Figure 5 corresponds to the same situations as Fig. 3, but now the total-time-on-test plots are shown, i.e. we plot (n, S_n), the cumulative intensity process, $S_n = \sum \Lambda_{\tau_k}$ against the number of observed deaths n. Under the incorrect model, M_0, the processes $\{S'_n\}$ and $\{S''_n\}$ arising respectively from the men's and women's samples exit from the upper and lower boundaries. This corresponds to the convergence of the processes $\{G'_n\}$ and $\{G''_n\}$, respectively, to 0 and 1 as shown in Fig. 3.

4. Further developments: Implementation of Markov chain Monte Carlo integration

In the above examples it was possible to handle the intensities analytically. In other cases this may be impossible. The main problem is that the $(P, \mathscr{H}_t)$-compensator differences are expressed by double integrals of the form

$$\Lambda_b^{\mathscr{H}}(x) - \Lambda_a^{\mathscr{H}}(x) = \int_a^b \left\{ \int_\Theta \lambda_s(x, \theta, H_{s-}) \pi(d\theta \mid H_{s-}) \right\} ds \quad (0 \leqslant a < b), \tag{4·1}$$

where the probability measure on Θ is changing with time s. If we want to integrate over Θ by applying a Markov chain Monte Carlo method, for a recent survey see Tierney (1994), at first sight it seems that we should run a different Markov chain for every $s \in [a, b]$.

One way out of this difficulty would be to find an importance sampling probability measure $\mu(d\theta)$ such that $\pi(d\theta \mid H_{s-}) \ll \mu(d\theta)$ for all $s \in [a, b]$: in other words $\mu(d\theta)$ has to dominate $\pi(d\theta \mid H_a)$. For example, take the prior $\pi(d\theta)$ and apply the Radon–Nikodym theorem to write the integral (4·1) as

$$\int_a^b \left\{ \int_\Theta \lambda_s(x, \theta, H_{s-}) \frac{d\pi \mid H_{s-}}{d\mu}(\theta) \mu(d\theta) \right\} ds. \tag{4·2}$$

Then, if $\{\theta_i\}_{i \in \mathbb{N}}$ is a random sample from the distribution μ or, more generally, forms an ergodic Markov chain with equilibrium distribution μ, by an application of the law of large numbers the integral (4·2) can be approximated by

$$\int_a^b \sum_{i=1}^N \lambda_s(x, \theta_i, h_{s-}) \frac{d\pi}{d\mu}(\theta_i) p(H_{s-} \mid \theta_i) \left\{ \sum_{i=1}^N \frac{d\pi}{d\mu}(\theta_i) p(H_{s-} \mid \theta_i) \right\}^{-1} ds. \tag{4·3}$$

Therefore, by a change of measure, we are able to use a single Monte Carlo run. However, when the information carried by the data grows significantly in time, a single probability measure cannot be a good importance-sampling distribution over the entire interval $[a, b]$; typically there will be a few importance weights which are much larger than the rest, and this makes the approximation (4·3) very unstable. What one should do then is to split the time interval $[a, b]$ into sufficiently small subintervals and sample, for each one, a different Markov chain with appropriate importance-sampling distribution. Even then, the integration over time in formula (4·3) has to be done numerically.

To gain efficiency we considered a continuous version of a new Markov chain Monte

Carlo sampling scheme, the simulated tempering method introduced by Marinari & Parisi (1992) and discussed also by Geyer & Thompson (1995).

In principle, the idea is very simple: Monte Carlo integration is used not only over the parameter space Θ, but also over the time interval $[a, b]$ with respect to Lebesgue measure. Thus, given the observed history, H_T, our target is a measure on the product space $[a, b] \times \Theta$ given by

$$\rho_1(ds, d\theta) = (b-a)^{-1}\pi(d\theta \mid H_{s-})\, ds = p(H_{s-} \mid \theta)(b-a)^{-1}Z(s)^{-1}\pi(d\theta)\, ds, \tag{4·4}$$

where the normalising factor

$$Z(s) = p(H_{s-}) = \int_\Theta p(H_{s-} \mid \theta)\pi(d\theta)$$

is called the partition function. Typically, $Z(s)$ cannot be evaluated explicitly. Instead, for a measure $\mu(ds)$ on the time interval $[a, b]$, which we will call the pseudoprior, we define the importance-sampling probability measure

$$\rho_\mu(ds, d\theta) = \mu(ds)\pi(d\theta \mid H_{s-})$$

on $[a, b] \times \Theta$.

For two different pseudopriors, $\mu_0(ds)$ and $\mu_1(ds)$, such that $\mu_0 \ll \mu_1$ and $f(s, \theta, H_{s-})$ is a predictable and ρ_{μ_0}-integrable function, we have

$$\int\int_{[a,b]\times\Theta} f(s, \theta, H_{s-})\rho_{\mu_0}(ds, d\theta) = \int\int_{[a,b]\times\Theta} f(s, \theta, H_{s-})\frac{d\rho_{\mu_0}}{d\rho_{\mu_1}}(s, \theta)\rho_{\mu_1}(ds, d\theta). \tag{4·5}$$

Note that the Radon–Nikodym derivative here is simply given by

$$\frac{d\rho_{\mu_0}}{d\rho_{\mu_1}}(s, \theta) = \frac{d\mu_0}{d\mu_1}(s). \tag{4·6}$$

We now define on $[a, b]$ the pseudoprior density $\mu(ds) = \{CZ(s)/\hat{Z}(s)\}\, ds$, where $\hat{Z}(s)$ is a guess for $Z(s)$ and C is a normalising constant. Denote by $\phi(s)$ the density $CZ(s)/\hat{Z}(s)$ of μ with respect to Lebesgue measure; $\hat{Z}(s)$ should be explicitly computable.

Since

$$\rho_\mu(ds, d\theta) \propto \frac{Z(s)}{\hat{Z}(s)}\pi(d\theta \mid H_{s-})\, ds = \frac{p(H_{s-} \mid \theta)}{\hat{Z}(s)}\pi(d\theta)\, ds, \tag{4·7}$$

the Hastings-ratio in the Metropolis algorithm can be computed without knowing $Z(s)$ or the constant C normalising the right-hand side of (4·7). It is therefore possible to generate, on a computer, a Markov chain $\{(s_j, \theta_j)\}$ with equilibrium distribution $\rho_\mu(ds, d\theta)$.

From (4·5) and (4·6) it follows that

$$\begin{aligned}\int\int_{[a,b]\times\Theta} f(s, \theta, H_{s-})\pi(d\theta \mid H_{s-})\, ds &= (b-a)\int\int_{[a,b]\times\Theta} f(s, \theta, H_{s-})\rho_1(ds, d\theta) \\ &= \int\int_{[a,b]\times\Theta} \frac{f(s, \theta, H_{s-})}{\phi(s)}\rho_\mu(ds, d\theta).\end{aligned} \tag{4·8}$$

As a consequence, almost surely

$$(b-a)\lim_{n\to\infty} \frac{\sum_{i=1}^n f(s_i, \theta_i, H_{s-})\phi(s_i)^{-1}}{\sum_{i=1}^n \phi(s_i)^{-1}} = \int_a^b \left\{\int_\Theta f(s, \theta, H_{s-})\pi(d\theta \mid H_{s-})\right\} ds. \tag{4·9}$$

When $Z(s)$ is unknown, the weights $\phi(s_i)^{-1}$ cannot be computed explicitly. Nevertheless, since $\phi(s)$ is the s-marginal density of $\rho_\mu(ds, d\theta)$ with respect to Lebesgue measure, the samples $\{s_i\}$ can be used to estimate the weights nonparametrically. A good choice of $\hat{Z}(s)$ is crucial for the stability of the estimates, because the largest empirical weights will be in an area where we have few sample points s_i. This suggests an adaptive strategy for the choice of the pseudoprior: if the variation of the empirical density estimate $\hat{\phi}(s)$ is too large, we can choose a new pseudoprior adaptively by moving mass to the area where $\hat{\phi}(s)$ is smaller. The best choice for a new pseudoprior would be $\mu'(ds) \propto \mu(ds)/\hat{\phi}(s)$.

These ideas were tested on Example 2, as an alternative way of doing the numerical computations.

Since the posterior densities $\pi(\theta | H_s)$ are unimodal, a simple guess for $Z(s)$ would be

$$\hat{Z}(s) = \max_{\theta \in \Theta} \pi(\theta) p(H_s | \theta),$$

or

$$\hat{Z}(s) = \pi\{\hat{\theta}_{MP}(T)\} p(H_s | \hat{\theta}_{MP}(T)),$$

where $\hat{\theta}_{MP}(T) = \text{argmax}\, \pi(\theta) p(H_T | \theta)$, possibly refined by using the Laplace approximation. Of course, in this example $Z(s)$ is known explicitly; nevertheless we proceeded without using this knowledge.

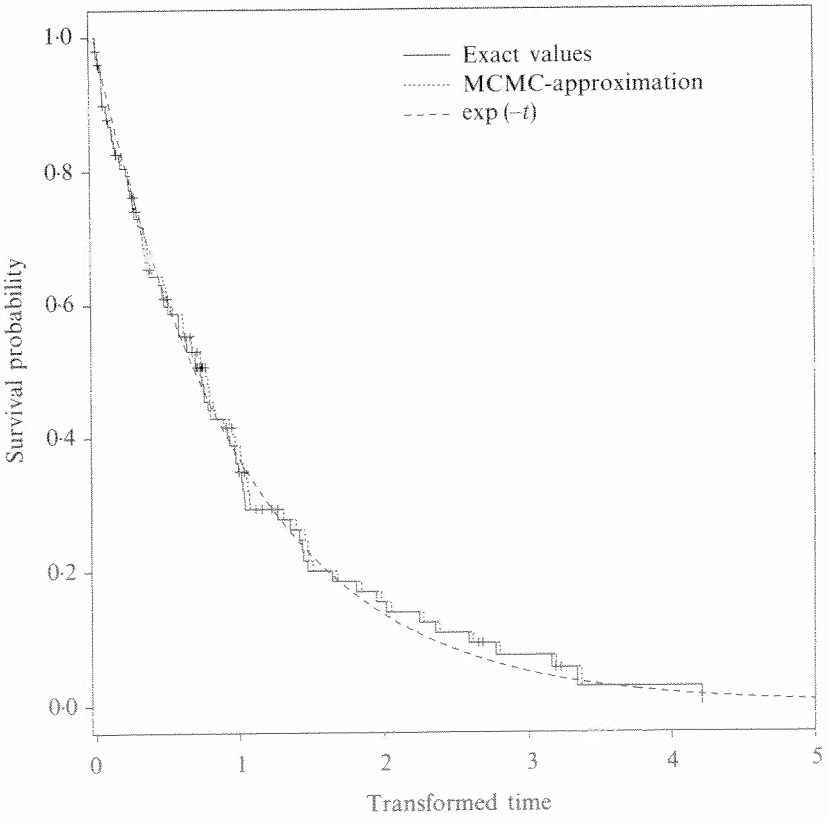

Fig. 6. Example 2: Time-transformed Kaplan–Meier survival curves of the, possibly right-censored, individual compensators for men. Values obtained by simulated tempering are compared with the exact ones.

In a simulation study, we used these rough approximations to construct the importance sampling distributions. The weights $\phi(s_i)^{-1}$ were computed both analytically using our knowledge of Z and empirically from the Monte Carlo samples by using a nonparametric kernel-smoothing estimate $\hat{\phi}(s)$ of the density of μ. Differences between the corresponding estimates were quite small. We compared exact values and Monte Carlo approximations for the men's cumulative intensities under model M_1. Note that these random variables are right censored. The corresponding time-transformed Kaplan–Meier survival curves are shown in Fig. 6. The survival probability of the 1-exponential distribution is shown for a comparison. On the basis of Fig. 6 we would be justified in saying that Monte Carlo error did not mislead our model assessment.

5. Discussion

As we have emphasised throughout, the 1-exponentiality of the cumulative intensities, which is here used as a reference in model assessment, holds with respect to the probability P when this is updated continuously from the data. The interpretation is therefore not in terms of 'correct statistical model and true parameter value'. In fact, many alternative models, large or small, may be found to be empirically consistent with such 1-exponentiality. However, a discrepancy between the 1-exponentiality, which corresponds to what a statistician expects to see when he or she is updating the predictions, staying within the postulated model, and the empirically observed behaviour of the cumulative intensities is an indication that such updating is not sufficient to account for the information in the data. In other words, the model is not sufficiently deep that a deductive learning according to Bayes' formula would form a satisfactory basis for sequential predictions.

This procedure appears to be a somewhat strange mixture of Bayesian and frequentist ingredients. The model itself and the updating procedure are Bayesian, but calculation of a statistic which would be sensitive to inadequacy of the model, and assessment of it against a reference distribution, are definitely frequentist ideas.

This combination of opposite inferential traditions is emphasised further by the fact that the computed statistics, the cumulative intensities up to the stopping times and the corresponding prequential p-values, are not functions of the data alone but depend also on the model probabilities, including the prior distribution, specified by the statistician. In addition, data are not considered as being fixed, as in Bayesian inference, nor are we thinking of 'drawing repeated samples under similar conditions' as in classical statistics. Instead, we are considering a nested sequence of predictions, and each data point (T_n, X_n) is treated as either fixed or random depending on whether it is, at the time of prediction, in the past or in the future.

The effect of the prior on the diagnostics appears in practice to be confined to the first few observations. So far as the asymptotic properties are concerned, such as the law of the iterated logarithm, two mutually absolutely continuous prior distributions will yield mutually absolutely continuous prequential models, and thus have exactly the same behaviour (Dawid, Seillier-Moiseiwitsch & Sweeting, 1992).

Finally, we wish to emphasise that, in spite of being able to use formal testing with exact reference distributions, our view of model assessment is that it is an inductive procedure where there is no definitive conclusion to be drawn. Prequential model assessment is done separately for each postulated model, without a need to consider alternatives. These aspects can be contrasted with model selection, which is a decision problem and

where the task is to choose one amongst a prespecified set of competing models. For asymptotic consistency of model-selection procedures, and their connection with prequential ideas, see Dawid (1992).

ACKNOWLEDGEMENT

We are grateful to Professors A. P. Dawid and D. M. Titterington and two anonymous referees for their comments. This research was supported by a research grant from the Academy of Finland.

REFERENCES

AALEN, O. O. & HOEM, J. M. (1978). Random time changes for multivariate counting processes. *Scand. Actuar. J.*, 81–101.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

ARJAS, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Am. Statist. Assoc.* **83**, 204–12.

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

BRÉMAUD, P. (1981). *Point Processes and Queues*. New York: Springer.

DAWID, A. P. (1984). Statistical theory: The prequential approach. *J. R. Statist. Soc.* A **147**, 278–92.

DAWID, A.P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 109–25. Oxford University Press.

DAWID, A. P., SEILLIER-MOISEIWITSCH, F. & SWEETING, T. J. (1992). Prequential tests of model fit. *Scand. J. Statist.* **19**, 45–60.

GELFAND, A. E., DEY, D. K. & CHANG, H. (1992). Model Determination using predictive distributions with implementation via sampling-based methods (with Discussion). In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 147–67. Oxford University Press.

GEYER, C. J. & THOMPSON, E. A. (1995). Annealing Markov Chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Assoc.* **90**, 909–20.

MARINARI, E. & PARISI, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19**, 451–8.

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

RAFTERY, A. E. (1992). Discussion on paper by A. E. Gelfand, D. K. Dey and H. Chang. In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 160–3. Oxford University Press.

SEILLIER-MOISEIWITSCH, F. (1992). Discussion on paper by A. E. Gelfand, D. K. Dey and H. Chang. In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, p. 165. Oxford University Press.

SEILLIER-MOISEIWITSCH, F. & DAWID, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Am. Statist. Assoc.* **88**, 355–9.

THERNEAU, T. M., GRAMBSCH, P. M. & FLEMING, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–60.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–62.

WEI, L. J. (1984). Testing goodness-of-fit for the proportional hazards model with censored observations. *J. Am. Statist. Assoc.* **79**, 642–52.

[*Received December* 1995. *Revised November* 1996]