STATISTICS IN MEDICINE, VOL. 15, 1757–1770 (1996)

NON-PARAMETRIC BAYESIAN APPROACH TO HAZARD REGRESSION: A CASE STUDY WITH A LARGE NUMBER OF MISSING COVARIATE VALUES

ELJA ARJAS

Department of Mathematical Sciences, University of Oulu, Linnanmaa, SF-90570 Oulu, Finland

AND

LIPING LIU

Department of Probability and Statistics, Peking University, P. R. China

SUMMARY

A 'packaged' non-parametric multiplicative hazard regression model is proposed, and applied to a study of the effects of some genetic and viral factors in the development of spontaneous leukaemia in mice. Hierarchical modelling and data augmentation are used to deal with the large number of missing covariate values. A Bayesian procedure is adopted, and the Metropolis–Hastings algorithm is used in the numerical computation of the posterior distribution.

1. INTRODUCTION

During the past two decades, the Cox proportional hazards regression model has played a very important role in survival analysis, especially in the assessment of different risk factors (covariates) on survival. A characteristic feature of the Cox model is that the hazard rate is factored into a non-parametric baseline hazard and a parametric relative risk function, and these are then estimated separately. In the generalized form, which allows for time dependent covariates, the hazard rate, λ_i, of the ith individual at time t is written in the form

$$\lambda_i(t; Z_{1i}(t), \ldots, Z_{ki}(t)) = \lambda_0(t) \exp\left(\sum_{j=1}^{k} \beta_j Z_{ji}(t) \right)$$

where $Z_{ji}(t)$ is a (time dependent) covariate ($j = 1, \ldots, k$).

Sometimes the exponential relative risk functions $\exp(\beta_j Z_j)$ may not be suitable for modelling the covariate effects. Many methods have been proposed to generalize this aspect of the Cox model. Recently, some authors proposed using splines to achieve a more flexible functional form; this also makes the relative risk function non-parametric (see Hastie and Tibshirani[1]). However, the basic factorization, which in practice means that the relative risk functions are estimated first, from a partial likelihood expression not involving the baseline hazard, is still made in these papers. With the fast development of computing facilities and algorithms, particularly the so-called Markov chain Monte Carlo (MCMC) methods in Bayesian inference (see Besag *et al.*[2]), it seems to us that this division has become much less crucial than before. Arjas and Gasbarra[3]

CCC 0277–6715/96/161757–14

Received September 1994
Revised November 1995

treat the hazard rate arising from simple right censored survival data non-parametrically, using an MCMC method for the numerical calculations, and Arjas and Liu[4] extend this method to several covariates, applying the technique in a case study. In this article, as in Arjas and Liu,[4] time t itself (or the baseline hazard) does not need to have a special role, but is treated in the same way as the other continuous covariates.

In classical regression modelling, interactions between covariates are often described by adding some product terms. In the present paper, a 'packaging' method is introduced into the non-parametric Bayesian framework, so that covariates are divided into groups, with the idea that those within each group are more closely connected than covariates across different groups. In this way we can replace complete multiplicativity by a somewhat weaker structural assumption.

It often happens that some data have not been recorded on all individuals, or they have been lost. Even so, these observations may contain useful information. Broadly, missing data can be of two different kinds, depending on whether a missing value is a response (for example, a survival time is right-censored) or a covariate. The former problem has been discussed in the literature much more extensively than the latter. In this paper emphasis is on missing covariate values, although both problems arise in our case study. Different methods of imputation have been suggested to deal with this problem, but usually individuals with missing covariate values are simply ignored; for more detailed discussion see Little.[5] Tanner and Wong[6] and Kong *et al.*[7] proposed two methods of data augmentation in the Bayesian framework, by repeatedly imputing missing data from their predictive distributions, and then using the (weighted) average posterior distribution based on both observed and augmented data to get an approximate posterior distribution of the underlying parameters. There are many similarities between their methods and the Gibbs sampler. The former depend on the property that if there were no missing data, the model would correspond to simple random sampling from a known distribution. In the present Bayesian approach it is in natural to treat unobserved covariate values in the same way as unknown parameters. Unlike the imputation methods, where a missing value is replaced by its point estimate, we shall end up with an approximation to the joint posterior distribution of all 'unobservables'. One advantage of such an approach is that there is not much difference whether missing values arise from discrete or continuous covariates, whereas some other methods are based on a categorization of the continuous covariates (see, for example, Little[5]).

We use a well known data set to illustrate the above ideas and methods. Section 2 contains a brief description of the example and the model, with an emphasis on the problem of missing covariate values. In Section 3 we describe the estimation algorithm. The paper concludes with a discussion of the results.

2. DATA SET AND STATISTICAL MODEL

Our empirical study concerns genetic and viral factors that may influence the development of spontaneous leukaemia in mice. The data set was presented in Kalbfleisch and Prentice,[8] and combined with an extensive and thoughtful statistical analysis, mainly using the proportional hazards regression estimates and significance tests. Hastie and Tibshirani[1] considered the same data, fitting a generalized Cox model and using cubic splines to obtain smooth estimates of the relative risk functions.

In this example, there are measurements on six covariates of which four are categorical (Z_1, MHC phenotype; Z_2, Gpd-1 phenotype; Z_3, sex; and Z_4, coat colour), taking only values 0 and 1, and the other two are continuous (Z_5, antibody level; and Z_6, virus level). None of these is time dependent. The sample size is 204. Z_5 is strongly (negatively) correlated with Z_6; a high antibody level is usually combined with a low virus level, and vice versa. In the primary data set several

different causes of death were recorded; those of main interest are thymic leukaemia (67 mice) and non-thymic leukaemia (12 mice). Following analyses 1–4 in Kalbfleisch and Prentice,[8] we treat causes of death other than thymic leukaemia as right-censorings and then study the hazards related to this particular type of cancer.

A large number of covariate values were missing: 1 for Z_1, 104 for Z_2, 4 for Z_5 and 29 for Z_6. There are two interesting additional features regarding these missing values. The first is that 91 recorded values of Z_6 exceeded 10,000, but were right censored at 10,000, and this value was then reported in the primary data set. The other is that the determination of Z_2 (Gpd-1 phenotype) began only midway through the experiment, with the result that measurements on Z_2 are available on only 100 mice. The covariates Z_2 are therefore not missing at random, but rather the opposite is true; if the animal survived the first year of the experiment, then Z_2 is known, otherwise not. The 'missingness' is therefore completely determined by the survival times. It would be a waste if Z_2 were simply deleted from the model, because then its determination on 100 mice had been in vain. On the other hand, deleting all such animals from the sample would mean that all short survival times were removed, and therefore hazard rate estimation could reasonably commence only at a point where full covariate information became available. Neither of these options seems satisfactory, as it would clearly be desirable to make the fullest possible use of the information in the data. Our solution, which is described later in this section, is to treat the missing covariates in the same way as unknown model parameters, applying Bayesian inferential techniques for all unobservables. Note, however, the practical difficulty arising from the scale of this problem for the missing values of Z_2 alone have $2^{104} > 10^{31}$ possible configurations!

To introduce a hazard rate model for these data, we first assume that all covariate values are known for the individual considered. The basic model is of the form

$$\lambda_i(t; Z_{1i}, \ldots, Z_{6i}) = f(t, Z_{1i}, \ldots, Z_{6i}) \tag{1}$$

where $\lambda_i(t)$ is the hazard rate of the ith individual at time (age) t, and f is a positive function. (In the following we only consider covariates which do not depend on time, because this is the case in the mice data. However, time dependent covariates do not present any extra difficulties and can be dealt with in exactly the same way.) Certainly, it would be unrealistic to try to estimate the non-parametric model (1) from data without making any structural assumptions regarding the function f. A common choice is to assume that all covariates and time act on the hazard rate multiplicatively, and then we get

$$\lambda_i(t; Z_{1i}, \ldots, Z_{6i}) = f_0(t) f_1(Z_{1i}) \ldots f_6(Z_{6i}) \tag{2}$$

where $f_0, \ldots, f_6$ are positive functions. Different models can be set up by choosing these functions in different ways. A full decomposition of (1) into a product of univariate functions means that the relative risk, corresponding to a change of any covariate from one level to another, remains constant in all other covariate values. This may be unrealistic, and therefore some compromise between (1) and (2) might be appropriate. By studying the biological background more carefully, we can perhaps find groups of two or more covariates (possibly including time t), such that there is potential interaction within each group, but not between the groups. In other words, the relative risks, corresponding to a change of the covariate values within a group, remain constant in covariate values outside it. 'Packaging' covariates in this way will not be technically complicated as long as there is at most a single continuous covariate within a group. For example, pairs (or triples) of categorical covariates can be viewed as a single categorical variable taking values in the corresponding product space, whereas functions of a categorical and a continuous covariate can be thought of as a set of univariate functions, indexed by the categories.

By using the hints given in Kalbfleisch and Prentice (reference 8, pp. 210–220) about the background of the covariates, and after discussion with a geneticist colleague, we decided to 'package' age and sex (t and Z_3) into one group, and Gpd-1 phenotype and virus level (Z_2 and Z_6) into another. Antibody level (Z_5), which has a strong correlation with virus level (see also Hastie and Tibshirani[1]), was deleted from the model. In this way we arrived at the form

$$\lambda_i(t; Z_{1i}, \ldots, Z_{6i}) = f_0(t, Z_{3i})\, f_1(Z_{1i})\, f_2(Z_{2i}, Z_{6i})\, f_3(Z_{4i}). \tag{2'}$$

(Following the above reasoning, Z_5 could only be added as a third (and then second continuous) argument into the function f_2. At present we do not have similar procedures for handling non-parametric functions of two or more continuous covariates as for univariate functions; such methods are currently being developed. A less satisfactory solution would be to discretize Z_5, leaving only Z_6 continuous.)

A non-parametric Bayesian approach is adopted for estimation. First note that the model contains four unknown (random) functions of a continuous variable, $f_0(t, 0), f_0(t, 1), f_2(0, Z_6)$ and $f_2(1, Z_6)$. We first specify the structure of model (2'), together with its prior distribution. In the sequel Greek letters will denote hyperparameters. Following Arjas and Gasbarra,[3] we make the simplifying structural assumption that these functions are piecewise constant. Thus we assume, for example, that the function $f_0(t, 0)$ can be written in the form of a jump process as

$$f_0(t, 0) = \sum_{j \geqslant 0} I\{S_j \leqslant t < S_{j+1}\} a_j \tag{3}$$

where I is the indicator function, and the jump times, S_j, and the levels a_j, $j \geqslant 1$, are random. Note, however, that when these processes are mixed in (2') according to the prior or the posterior distribution, to form a corresponding predictive hazard rate, the resulting functions will be continuous in t.

The prior distribution for the function (3) is specified by assuming that the jump times $0 = S_0 < S_1 < \ldots$ form a time-homogeneous Poisson sequence with a given hyperparameter μ, and that $\{a_0, a_1, \ldots\}$ is a sequence of positive random variables whose joint distribution is defined inductively as follows. Denoting by $\gamma(.; \alpha, \beta)$ the gamma density with scale parameter α and shape parameter β, (i) a_0 is assumed to follow a gamma prior distribution $\gamma(.; \alpha_0, \alpha_0')$, and (ii) given $\{a_0, \ldots, a_j\}$, a_{j+1} has gamma prior distribution $\gamma(.; \alpha, \alpha/a_j)$ $(j \geqslant 0)$. Here α_0, α_0' and α are hyperparameters specified by the analyst, controlling the initial level and the fluctuations of $f_0(t, 0)$. In particular, the initial level a_0 has prior mean $E_{\text{prior}}(a_0) = \alpha_0/\alpha_0'$ and precision $1/\text{var}_{\text{prior}}(a_0) = \alpha_0'^2/\alpha_0$. Moreover, $E_{\text{prior}}(a_{j+1} | a_0, \ldots, a_j) = a_j$, corresponding to a neutral prior assumption regarding trends in the function (3), whereas its fluctuations are controlled by the intensity μ and by the coefficients of variation $\sqrt{\{\text{var}_{\text{prior}}(a_{j+1} | a_0, \ldots, a_j)\}}/E_{\text{prior}}(a_{j+1} | a_0, \ldots, a_j) = 1/\sqrt{\alpha}$ of the conditional jump distribution. A small value of μ and a large value of α reflect the prior assumption that the function $f_0(t, 0)$ should remain fairly close to its initial value $f_0(0, 0)$. Such tight control can be eased by increasing μ, allowing for more jumps, and/or decreasing α, letting the jumps increase in size. For more details on the choice of the hyperparameters we refer the reader to Arjas and Gasbarra.[3]

The functions $f_0(t, 1), f_2(0, Z_6)$ and $f_2(1, Z_6)$ are assumed to have a similar structure and prior distribution, but with different parameters and hyperparameters. For definiteness we suppose that the initial level of $f_2(0, Z_6)$ is always equal to 1. We also make the simplifying prior assumption that these four functions are independent.

As for f_1 and f_3, where the corresponding covariates are discrete, we can in principle assign to them any prior distribution supported by (a subset of) $[0, \infty)$. In this example, we simply use

a gamma prior so that, at least partly, conjugacy can be realized. To save the relative risk interpretation, and also for uniqueness, we let $f_1(0) \equiv 1$ and assign to $f_1(1)$ the gamma prior $\gamma(.; \beta_1, \beta_1')$, with β_1 and β_1' given. Thus, the prior guess of the relative risk associated with MHC level 1, compared with level 0, would be given by $E_{\text{prior}}(f_1(1)) = \beta_1/\beta_1'$, and the corresponding precision by $1/\text{var}_{\text{prior}}(f_1(1)) = \beta_1'^2/\beta_1$. Similarly, $f_3(0) \equiv 1$ and the prior of $f_3(1)$ is taken to be $\gamma(.; \beta_3, \beta_3')$, with β_3 and β_3' given. Again the prior assumption is made that $f_1(1)$ and $f_3(1)$ are independent, and also independent of the four univariate functions discussed above.

Under such a structure, we can easily write down the prior density of parameters of model (2') (see Arjas and Liu[4]).

Let us then return to the questions arising from missing covariate values. Suppose, generically, that for individual i the covariate value Z_{ui} is missing. For a full Bayesian approach, we consider first the prior assumption that the covariate value Z_{ui} is distributed according to a simple hierarchical submodel

$$Z_{ui} \sim \mathscr{D}_u(.; p_u); \qquad p_u \sim \pi_u(.; \xi_{u,0}). \tag{4}$$

Here $\mathscr{D}_u$ is a distribution with parameter p_u, and p_u follows the distribution π_u with given hyperparameter $\xi_{u,0}$. By inserting this submodel into the basic hazard regression model (making the natural conditional independence assumptions throughout), we get a hierarchical structure governing also the missing values. The computational burden is reduced significantly if $\mathscr{D}_u$ and π_u can be chosen from a conjugate family.

Each observed covariate value is now regarded as a realization from this submodel. If only this submodel had been considered, such values could be interpreted as observed responses arising from the submodel. When combined with the hazard regression model, however, the evaluation of the posterior distributions of missing values becomes an intermediate step, and in general will depend on information concerning both the observed covariates and the responses. Under submodel (4), the posterior distribution of a single missing covariate will depend on the other (observed) covariates of the same individual only indirectly, through their role as explanatory variables of the corresponding survival.

For a more direct description of the dependence between the covariates, submodel (4) can be replaced by

$$Z_{ui} \sim \mathscr{D}_u(.; p_{ui}); \qquad p_{ui} \sim \pi_u(.; \xi_{u,0}, Z_{-u,i}) \tag{4'}$$

where $Z_{-u,i}$ stands for all covariates of individual i except Z_{ui}.

In this present case study two procedures were tried. In the first one, (4) was assigned to all missing covariates, whereas in the second one (4') was used in modelling the missing Z_{2i}'s. So, to the missing values of the binary covariate Z_1 we simply assigned a Bernoulli prior, where the parameter p_1 follows a beta distribution $B(.; \tilde{\alpha}_1, \tilde{\beta}_1)$. For missing values of Z_2, when submodel (4) was adopted, we used the same structure, but with different hyperparameters $\tilde{\alpha}_2, \tilde{\beta}_2$ controlling the Bernoulli parameter p_2. In (4'), we let π_2 take the form of a logistic model, that is,

$$p_{2i} = \frac{\exp(\tilde{p}_{21}Z_{1i} + \tilde{p}_{23}Z_{3i} + \tilde{p}_{24}Z_{4i} + \tilde{p}_{26}Z_{6i})}{1 + \exp(\tilde{p}_{21}Z_{1i} + \tilde{p}_{23}Z_{3i} + \tilde{p}_{24}Z_{4i} + \tilde{p}_{26}Z_{6i})} \tag{5}$$

where each $\tilde{p}_{2k}$ follows a uniform distribution in $[-\xi_{2k}, \xi_{2k}]$, with fixed hyperparameter ξ_{2k}. For the continuous covariate Z_6 we postulated an exponential submodel, with parameter p_6 having a gamma distribution $\gamma(.; \tilde{\alpha}_6, \tilde{\beta}_6)$. An advantage of this choice is that if Z_6 is beyond 10,000 (censored), then the conditional prior given p_6 and $Z_6 > 10{,}000$ is still p_6-exponential in the region $[10{,}000, \infty)$. This reduces the computational burden somewhat.

Now we have three hierarchical submodels for missing Z_1, Z_2 and Z_6, respectively. We make the prior assumption that they are independent. So, under (4), the prior density of Z_1, Z_2 and Z_6 is given by the product

$$\frac{p_1^{\tilde{\alpha}_1-1}(1-p_1)^{\tilde{\beta}_1-1}}{\mathrm{B}(\tilde{\alpha}_1,\tilde{\beta}_1)}\times\frac{p_2^{\tilde{\alpha}_2-1}(1-p_2)^{\tilde{\beta}_2-1}}{\mathrm{B}(\tilde{\alpha}_2,\tilde{\beta}_2)}\times\frac{\tilde{\alpha}_6^{\tilde{\beta}_6}}{\Gamma(\tilde{\alpha}_6)}p_6^{\tilde{\alpha}_6-1}\exp[-\tilde{\beta}_6 p_6] \tag{6}$$

where B(. , .) denotes the beta function. Similarly, the product of their likelihoods is

$$\prod_i\left[p_1^{Z_{1i}}(1-p_1)^{1-Z_{1i}}p_2^{Z_{2i}}(1-p_2)^{1-Z_{2i}}\right]\times\prod_{\{i:Z_{6i}<10000\}}\left[p_6\mathrm{e}^{-Z_{6i}p_6}\right]\prod_{\{i:Z_{6i}\geq 10000\}}\left[p_6\mathrm{e}^{-(Z_{6i}-10000)p_6}\right]. \tag{7}$$

Under (4'), (6) is changed into

$$\frac{p_1^{\tilde{\alpha}_1-1}(1-p_1)^{\tilde{\beta}_1-1}}{\mathrm{B}(\tilde{\alpha}_1,\tilde{\beta}_1)}\times\prod_{\{k=1,3,4,6\}}\frac{1\{-\xi_{2k}\leq\tilde{p}_{2k}\leq\xi_{2k}\}}{2\xi_{2k}}\times\frac{\tilde{\alpha}_6^{\tilde{\beta}_6}}{\Gamma(\tilde{\alpha}_6)}p_6^{\tilde{\alpha}_6-1}\exp[-\tilde{\beta}_6 p_6] \tag{6'}$$

and the only change in (7) is that p_2 is replaced by p_{2i} in (5).

From Figure 1 we can easily see the chain structure of these two hierarchical models. Unobservables (parameters and missing covariates) are surrounded by circles, and given values (observed data and hyperparameters) by squares. There are altogether four and five levels, respectively, with arrows indicating the dependence relationships between them.

In the present data where some survival times were right-censored, the response concerning individual i takes the form (Y_i, δ_i). Here Y_i is the time when this individual was last seen, and $\delta_i = 1$ or 0 indicates death or censoring. Under such a structure, and either assuming that censoring is non-informative or disregarding its contribution, the likelihood arising from the survival data and the hazard model (2') can be written as

$$\text{Lik(data; parameters)} = \prod_i [\, f_0(Y_i, Z_{3i}) f_1(Z_{1i}) f_2(Z_{2i}, Z_{6i}) f_3(Z_{4i})]^{\delta_i}$$
$$\times\exp\left[-\sum_i\int_0^{Y_i} f_0(s, Z_{3i}) f_1(Z_{1i}) f_2(Z_{2i}, Z_{6i}) f_3(Z_{4i})\,\mathrm{d}s\right]. \tag{8}$$

In a hierarchical model such as this one, it is not obvious which densities should be thought of as belonging to the prior and which to the likelihood, but the joint distribution of all unobservables (parameters and missing covariates) and observables (observed covariates and responses) is simply the product of (6) (or (6')), (7), (8) and the prior density of the parameters concerning the f's.

3. ESTIMATION PROCEDURE

In principle, the joint posterior distribution of the model parameters and of the missing data is determined by the prior and the likelihood. In practice, however, in problems involving large hierarchical models, we cannot hope to get the posterior distribution (or even some characteristics of it) analytically. Fortunately, Markov chain Monte Carlo (MCMC) methods can often be used successfully to obtain approximate numerical solutions (see for example Besag *et al.*[2]). The main idea is to construct an ergodic Markov chain on the parameter space, with the posterior as the stationary distribution.

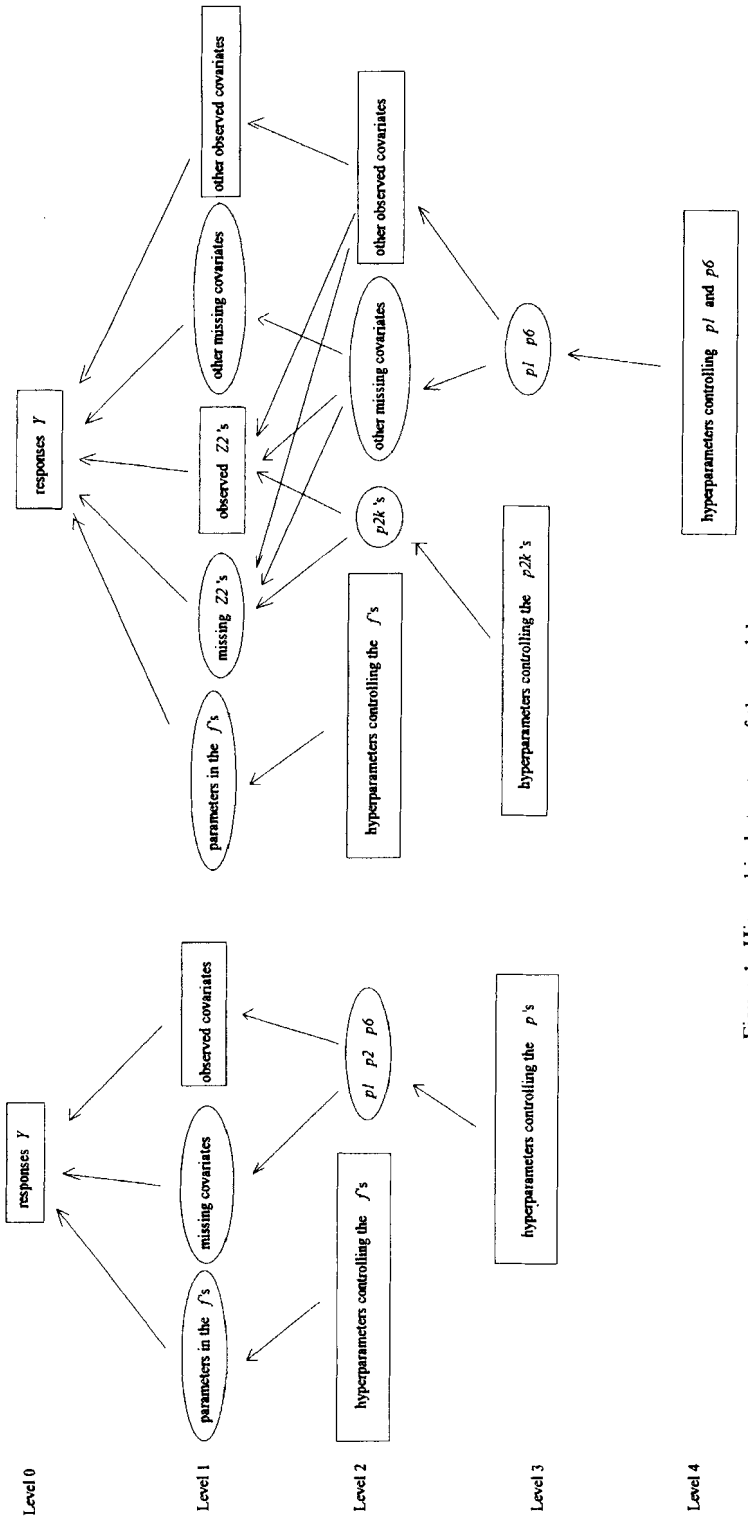

Figure 1. Hierarchical structure of the models

In the present case study, the set of parameters consists of $f_1(1), f_3(1)$, all four positive functions f_u of the form (3), all missing covariates, and the 'first level parameters' p_1, p_2 (or $\tilde{p}_{21}$, $\tilde{p}_{23}$, $\tilde{p}_{24}$ and $\tilde{p}_{26}$) and p_6 in the submodel. In the MCMC algorithm we generate first an initial point from the prior distribution, and then update all unobservables co-ordinate by co-ordinate. In each step of the updating, other co-ordinates are fixed at their current values. Both the Gibbs sampler and the Metropolis–Hastings algorithm are used. For $f_u(1)$ corresponding to a discrete covariate, we draw a new value from its full conditional distribution (which is just a gamma distribution) to replace the old one (Gibbs sampler); for a jump time or jump level of a function of a continuous variable (packaged case), we draw a candidate from a proper transition density (uniform for jump time, and gamma distribution for jump level, respectively), calculate the corresponding Hastings ratio (see for example Besag *et al.*[2], Roberts and Smith[9]), then decide whether to accept this candidate or to stay at the previous value, and go on to the next parameter. After this the algorithm proceeds to deal with the missing value. First we use the Gibbs sampler to update the parameters p_1 and p_2 (respectively p_6) whose full conditional distributions are still beta (gamma). Then we go one level higher in the hierarchy of the submodels. New candidates for missing covariates are drawn from Bernoulli or exponential distributions, depending on whether they are binary valued or continuous, and therefore the corresponding Hastings ratios are calculated so that some of the candidates are accepted as the new augmented missing data, and so on. Notice that in this substep, not only the submodel of missing values but the entire model must be considered, that is, the full conditional distribution must arise from the joint density described at the end of Section 2. If (4′) is used for the missing Z_{2i}'s, then the updating of p_2 is replaced by updating the $\tilde{p}_{2k}$'s one by one, using the Metropolis–Hastings algorithm, with uniform distributions to generate proposals. In the substeps where Gibbs sampling is possible, we have made full use of the conjugate properties.

Note that the positions of the jump points of the four piecewise constant functions, as well as the number of such points in the considered range, do not remain fixed and will be updated as a part of the sampling algorithm. After a large number of iterations, the corresponding empirical mixtures behave approximately as continuous functions (see Figures 3 and 5). For more details concerning the algorithm we refer to Arjas and Gasbarra.[3]

By repeating the procedure described above, we get a sequence of parameter estimates. It is easy to see that this sequence forms a Markov chain, with the posterior distribution of the unobservables as its stationary distribution (see Roberts and Smith[9]). Because of the ergodic theorem, various (approximate) statistical inferences about this target distribution (posterior) can be drawn by using this sequence.

Note that this simple sweep strategy, in which the unobservables are always updated in a prespecified order, is by no means the only one to give the desired result. For example, a random order could be followed, and/or several co-ordinates could be updated in a single step of the Metropolis–Hastings algorithm.

In essence, the method we have described above involves sampling from the joint distribution of all unobservables, given the data and the hyperparameters. Some of them, such as unknown covariate values of some particular mouse, are not of much inferential interest as such and will therefore in practice be 'integrated out' in the course of the sampling. This leads to the desired posterior marginals of the model parameters. From an inferential point of view, it would clearly be equivalent to consider mixture models, where the integration over 'less interesting' unobservables is done analytically. For example, considering model (4) and the likelihood contribution of a mouse indexed by i, with an unknown covariate value Z_{2i}, we could replace the product

$$p_2^{Z_{2i}}(1-p_2)^{1-Z_{2i}} \times [\lambda_i(Y_i; Z_{1i}, Z_{2i}, \ldots, Z_{6i})]^{\delta_i} \times \exp\left[-\int_0^{Y_i} \lambda_i(s; Z_{1i}, \ldots, Z_{6i})\,ds \right]$$

Table I. Prior means and coefficients of variation (CV) of some parameters

		$f_0(t,0)$ and $f_0(t,1)$		$f_2(t,0)$ and $f_2(t,1)$		$f_1(1)$	$f_3(1)$
		a_0	$a_j\ (j>1)$	a_0	$a_j\ (j>1)$		
Experiments 1 and 3	Mean	0·0016	a_{j-1}	1	a_{j-1}	1	1
	CV	0·4472	0·4472	0·4472	0·4472	0·4472	0·4472
Experiments 2 and 4	Mean	0·0016	a_{j-1}	1	a_{j-1}	1	1
	CV	0·4472	0·3162	0·4472	0·3162	0·4472	0·4472

coming from expressions (7) and (8) by the mixture

$$p_2 \times \lambda_i(Y_i; Z_{1i}, 1, Z_{3i}, \ldots, Z_{6i})]^{\delta_i} \times \exp\left[-\int_0^{Y_i} \lambda_i(s; Z_{1i}, 1, Z_{3i}, \ldots, Z_{6i})\mathrm{d}s \right]$$

$$+ (1-p_2) \times \lambda_i(Y_i; Z_{1i}, 0, Z_{3i}, \ldots, Z_{6i})]^{\delta_i} \times \exp\left[-\int_0^{Y_i} \lambda_i(s; Z_{1i}, 0, Z_{3i}, \ldots, Z_{6i})\mathrm{d}s \right].$$

Here, however, because of the simplicity of its implementation on a computer, we have preferred to work with the product form likelihood corresponding to the hierarchical model structure (Figure 1).

4. DISCUSSION

The possibility of drawing random samples from the posterior distribution gives rise to a large number of interesting inferential techniques. Here we just describe some aspects.

We made four simulation experiments, using a Sun Sparc workstation and S-plus programming language. Computation was time-consuming, essentially because of the large number of unobservables (and possibly also because of the language we used), and therefore only 1500 iterations were done in experiment 1, 1000 in experiments 2 and 3, and 500 in experiment 4. From follow-up, and from a number of additional trials, we concluded that the numerical stability of the algorithm was quite good. Also the acceptance rates of the proposals in the Metropolis–Hastings procedures were high enough, typically around 80 per cent for continuous parameters, and over 25 per cent for binary missing covariate values. We therefore felt that the results thus obtained were sufficiently accurate for a practical assessment of the method. The four experiments consisted of using two sets of hyperparameters, and two submodels to deal with the missing covariates; in experiments 1 and 2 we used submodel (4) for all missing covariates, whereas in experiments 3 and 4 we used (4′) to treat the missing Z_2's. The first set of hyperparameters was $\mu_0 = 0{\cdot}01$ (for jump times of f_0), $\alpha_0 = 5$ (shape parameter for jump levels of f_0), $\mu_2 = 0{\cdot}0005$ (for jump times of f_2), $\alpha_2 = 5$ (shape parameter for jump levels of f_2), $\beta_1 = \beta_3 = 5$ (shape parameters for $f_1(1)$ and $f_3(1)$), $\tilde{\alpha}_1 = \tilde{\beta}_1 = \tilde{\alpha}_2 = \tilde{\beta}_2 = 1$ (hyperparameters for missing Z_1 and Z_2's), and $\tilde{\alpha}_6 = 1$ and $\tilde{\beta}_6 = 4000$ (hyperparameters for the missing Z_6's). In the second set we used hyperparameter values $\alpha_0 = \alpha_2 = 10$ and $\tilde{\alpha}_2 = \tilde{\beta}_2 = 2$, keeping the others unchanged. In submodel (4′) we chose $\xi_{21} = \xi_{23} = \xi_{24} = 5$ and $\xi_{26} = 0{\cdot}001$. In this way the priors we specified are proper probability densities, but do not impose tight prior constraints on the unobservables. The main difference between these two sets of hyperparameters is that in the first group, the control on the fluctuation of the sample paths of the four piecewise constant functions was less stringent. This difference is reflected in the behaviour of the posterior estimates (see below). We summarize the prior means and coefficients of variation of some parameters in Table I.

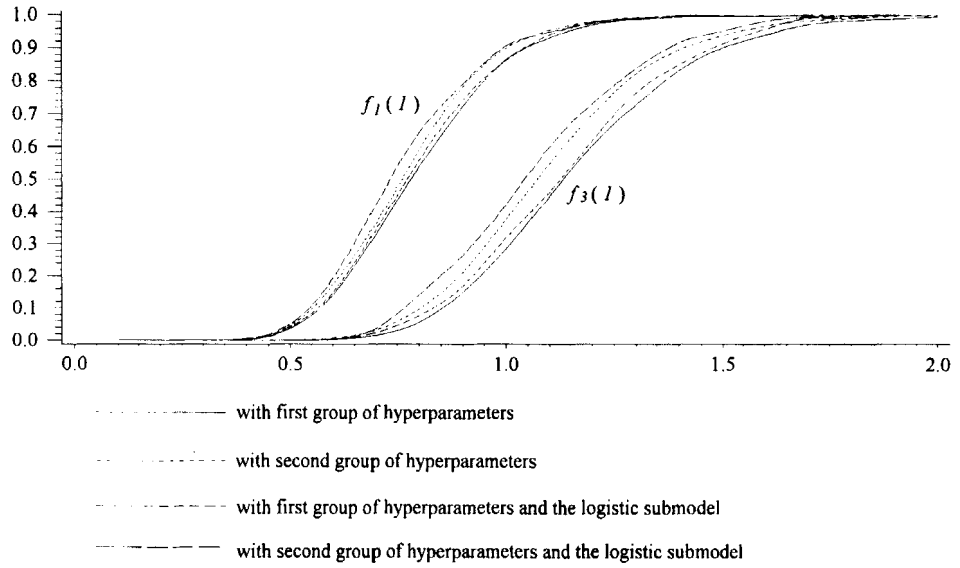

Figure 2. Distribution functions of $f_1(1)$ and $f_3(1)$

A more sophisticated, and perhaps also more efficient, way of assessing the sensitivity of inferences to model specification, and particularly to the specification of the prior, could be based on the ideas presented in Smith and Gelfand.[10]

4.1. Effects of categorical covariates

In the Cox regression model, the relative risk associated with Z_1 (MHC phenotype), say, is e^{β_1}, where β_1 is the regression coefficient. In our model the corresponding ratio of hazard rates, $f_1(1)$, is regarded as a random parameter and therefore the logical way to summarize the findings concerning relative risk is to use the (marginal) posterior distribution of $f_1(1)$. It is displayed in Figure 2, together with the posterior distribution of $f_3(1)$, the relative risk associated with coat colour. In the four simulation experiments, we obtained the values 0·86, 0·90, 0·86 and 0·91 for the approximate posterior probability that $f_1(1)$ is less than one. (Relative risks less than one correspond to a negative value of β_1 in the Cox model, as both indicate that there is lower risk if Z_1 takes value 1 than when it takes value 0). The approximate 95 per cent Bayesian credible intervals (between the 2·5 and 97·5 percentiles of the posterior) for $f_1(1)$ were (0·49, 1·25), (0·47, 1·17), (0·47, 1·20) and (0·45, 1·20) in the four experiments.

For comparison, we estimated the Cox model by using the same covariates as in model (2′) (omitting antibody level) and by deleting the individuals with missing covariate values as in Table 8.6 in Kalbfleisch and Prentice (reference 8, p. 216). The estimate for β_1 was $-0{\cdot}559$ (standard error 0·486), leading to the approximate 95 per cent confidence interval (0·22, 1·48) for the relative risk e^{β_1}. It seems therefore that the statistical conclusions which can be drawn from the marginal posterior of $f_1(1)$ are somewhat stronger than those based on the partial MLE $\hat{\beta}_1$ and a normal approximation. We did not try to assess how much of this difference was due to our using a different hazard model and a different inferential principle, and how much could be attributed to the inclusion of the observations with missing covariate values.

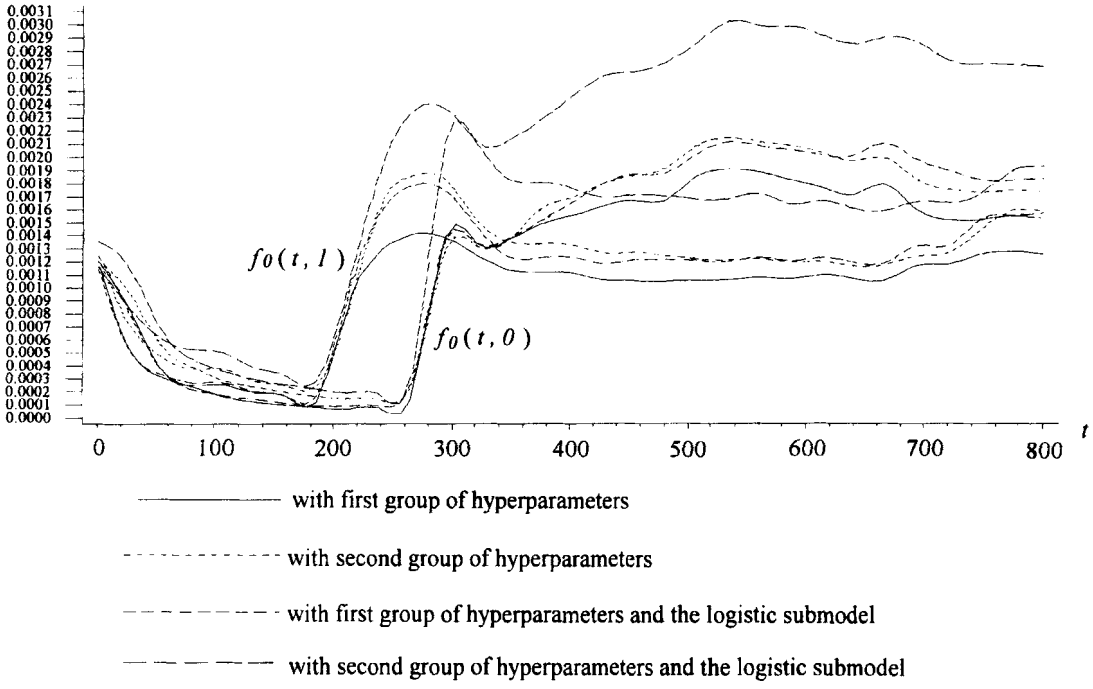

Figure 3. Averaged realizations of $f_0(t, 0)$ and $f_0(t, 1)$

4.2. Checking the packaging assumption

We expect that our model, which uses packaged non-parametric relative risk functions, would fit the data at least as well as the Cox model, which is based on completely multiplicative parametric (exponential) relative risks. At present, there do not seem to be direct diagnostic procedures available for checking the performance of such non-parametric Bayesian models. Work in this direction is currently in progress, however.

In Figures 3 and 5 we just plotted the averaged realizations (from the 1500 (respectively 1000, 1000, 500) realizations generated by the simulation) of our four random functions of a continuous variable. If the (unknown) functions $f_2(0, .)$ and $f_2(1, .)$ were proportional, proportionality would hold also approximately for the averages from the simulation. However, Figure 5 indicates that a multiplicativity assumption on the effects of Gpd-1 phenotype and virus level would be inappropriate. In this sense the packaging method appears to have been worthwhile. An additional conclusion is that, as the virus level increases, both functions f_2 increase, and $f_2(0, Z_6)$ is apparently always above $f_2(1, Z_6)$. (In fact, it might have been reasonable to build such monotonicity properties into the specification of the prior, see Arjas and Gasbarra[11]). It is less apparent from Figure 3 how close to proportionality $f_0(t, 0)$ and $f_0(t, 1)$ would be. There seems to be some evidence that the hazard first decreases with age, then suddenly assumes a much higher value, and that the change occurs somewhat earlier in female than in male mice. However, we should be cautious when trying to draw conclusions from Figure 3 when t is beyond 300 days, or from Figure 5 when Z_6 is beyond 10,000, for in these areas the data are quite sparse.

Such uncertainty is actually easily quantified. For each fixed time t, $f_0(t, 0)$ is a real valued random parameter, and its Bayesian credible interval can be calculated from our simulation

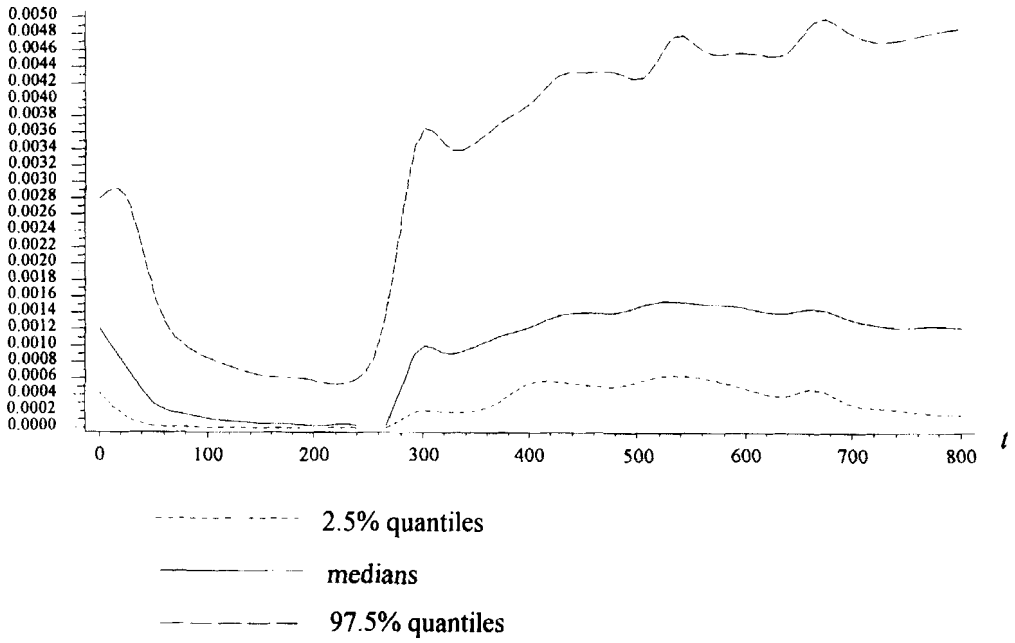

Figure 4. Bayesian credible intervals for $f_0(t, 0)$

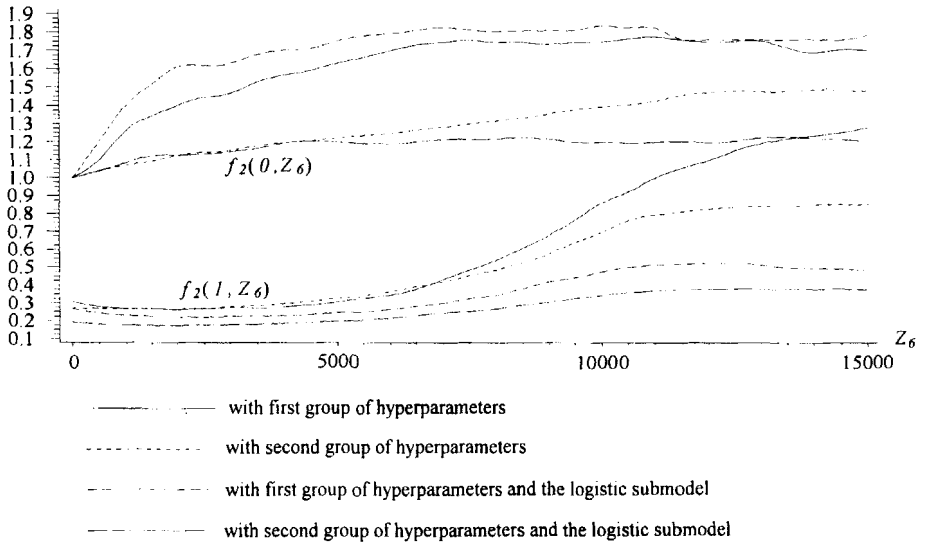

Figure 5. Averaged realizations of $f_2(0, Z_6)$ and $f_2(1, Z_6)$

results. Figure 4 shows the 95 per cent pointwise credible intervals $f_0(t, 0)$ for different values of t, together with the medians (based on our first experiment). This gave us three continuous curves. Notice how the credible interval becomes wider as t increases beyond 300 and the risk set becomes smaller. Similar curves can be drawn for $f_0(t, 1)$, $f_2(t, 0)$ and $f_2(t, 1)$.

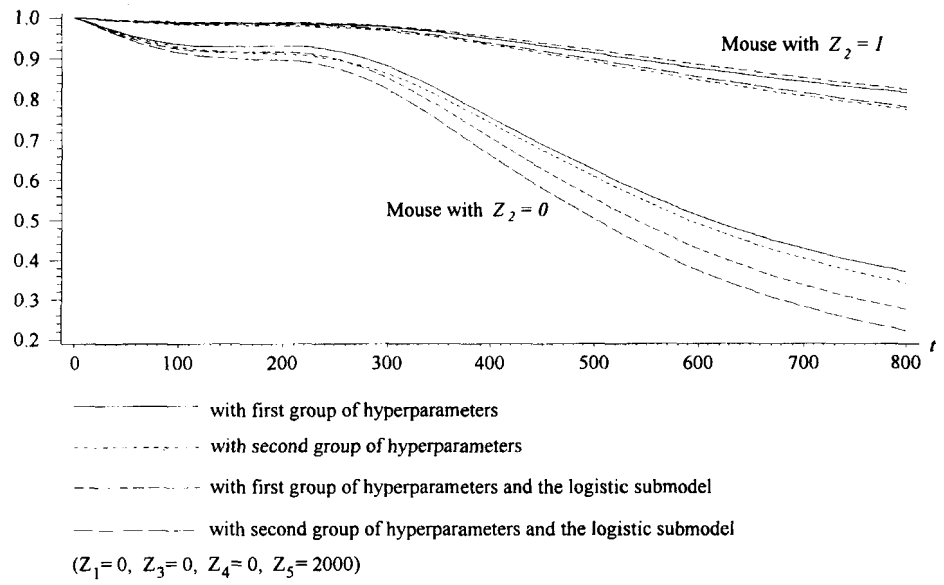

Figure 6. Predictive survival functions for two 'new' mice

4.3. Distribution of missing covariates; comparison of different submodels

Approximate posterior distributions of missing covariates are obtained here as a by-product. Note that in submodel (4), all the 104 missing Z_2's had the same prior, but their posterior distributions need not be the same. For example, considering two particular mice, indexed here respectively by 2 and 18, we arrive at (approximate) posterior probabilities $P(Z_{2,2} = 0) = 0{\cdot}20$ (or 0·24 in the second experiment) and $P(Z_{2,18} = 0) = 0{\cdot}73$ (0·67). This difference in the posterior is explained by these two mice having different observed covariate values and survival times. The corresponding values in experiments 3 and 4, which used model (4′) and thereby explicit information about the covariate values of these same individuals, are $P(Z_{2,2} = 0) = 0{\cdot}17$ (or 0·14) and $P(Z_{2,18} = 0) = 0{\cdot}63$ (0·57). A similar observation can be made regarding the missing Z_6's.

4.4. Survival distributions

One attractive feature of the Bayesian approach to survival analysis is the direct interpretation of the predictive survival distribution. Supposing that we had to predict the survival of yet another mouse, which was not included in the data set but 'similar', we can easily calculate its predictive survival probabilities, conditionally on its covariates and all the observed data. To illustrate this, we considered here two 'new' mice with covariate values $(Z_1, Z_2, Z_3, Z_4, Z_6) = (0, 0, 0, 0, 2000)$ and $(0, 1, 0, 0, 2000)$, respectively, and drew the corresponding predictive survival curves (Figure 6). The survival probabilities differ dramatically, although the only difference in the covariate values is in Z_2 (Gpd-1 phenotype). This result is in good agreement with Figure 4.

ACKNOWLEDGEMENTS

We are grateful to Outi Savolainen for useful discussion of the genetic background to this problem. The work was supported by The Academy of Finland.

REFERENCES

1. Hastie, T. and Tibshirani, R. 'Exploring the nature of covariate effects in the proportional hazards model', *Biometrics*, **46**, 1005–1016 (1990).
2. Besag, J., Green, P., Higdon, D. and Mengersen, K. 'Bayesian computation and stochastic systems (with discussion)', *Statistical Science*, **10**, 3–66 (1995).
3. Arjas, E. and Gasbarra, D. 'Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler', *Statistica Sinica*, **4**, 505–524 (1994).
4. Arjas, E. and Liu, L. 'Assessing the losses caused by an intervention: a hierarchical Bayesian approach', *Applied Statistics*, **44**, 357–368 (1995).
5. Little, R. 'Incomplete data in event history analysis', in *Demographic Applications of Event History Analysis*, Clarendon Press, Oxford, 1992, pp. 209–230.
6. Tanner, M. A. and Wong, W. H. 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association*, **82**, 528–550 (1987).
7. Kong, A., Liu, J. S. and Wong, W. H. 'Sequential imputations and Bayesian missing data problems', *Journal of the American Statistical Association*, **89**, 278–288 (1994).
8. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
9. Roberts, G. O. and Smith, A. F. M. 'Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms', *Stochastic Processes and their Applications*, **49**, 207–216 (1994).
10. Smith, A. F. M. and Gelfand, A. E. 'Bayesian statistics without tears: a sampling-resampling perspective', *American Statistician*, **46**, 84–88 (1992).
11. Arjas, E. and Gasbarra, D. 'Bayesian inference of survival probabilities, under stochastic ordering constraints', *Journal of the American Statistical Association*, (1996, to appear).