Genet. Sel. Evol. 30 (1998) 241–256

Original article

Bayesian inference in the semiparametric log normal frailty model using Gibbs sampling

Inge Riis Korsgaard*, Per Madsen, Just Jensen

Department of Animal Breeding and Genetics, Research Centre Foulum, Danish Institute of Agricultural Sciences, P.O. Box 50, DK-8830 Tjele, Denmark

(Received 16 October 1997; accepted 23 April 1998)

Abstract – In this paper, a full Bayesian analysis is carried out in a semiparametric log normal frailty model for survival data using Gibbs sampling. The full conditional posterior distributions describing the Gibbs sampler are either known distributions or shown to be log concave, so that adaptive rejection sampling can be used. Using data augmentation, marginal posterior distributions of breeding values of animals with and without records are obtained. As an example, disease data on future AI-bulls from the Danish performance testing programme were analysed. The trait considered was 'time from entering test until first time a respiratory disease occurred'. Bulls without a respiratory disease during the test and those tested without disease at date of analysing data had right censored records. The results showed that the hazard decreased with increasing age at entering test and with increasing degree of heterozygosity due to crossbreeding. Additive effects of gene importation had no influence. There was genetic variation in log frailty as well as variation due to herd of origin by period and year by season. © Inra/Elsevier, Paris

survival analysis / semiparametric log normal frailty model / Gibbs sampling / animal model / disease data on performance tested bulls

Résumé – Inférence Bayésienne dans un modèle de survie semiparamétrique log-normal à partir de l'échantillonnage de Gibbs. Une analyse complètement Bayésienne utilisant l'échantillonnage de Gibbs a été effectuée dans un modèle de survie semiparamétrique log-normal. Les distributions conditionnelles a posteriori mises à profit par l'échantillonnage de Gibbs ont été, soit des distributions connues, soit des distributions log-concaves de telle sorte que l'échantillonnage avec rejet adaptatif a pu être utilisé. En utilisant la simulation des données manquantes, on a obtenu les distributions marginales a posteriori des valeurs génétiques des animaux

* Correspondence and reprints
E-mail: snfirk@genetics.sh.dk or IngeR.Korsgaard@agrsci.dk

avec ou sans données. Un exemple analysé a concerné les données de santé des futurs taureaux d'insémination dans les stations danoises de contrôle de performance. Les taureaux sans maladie respiratoire ou n'en ayant pas encore eu à la date de l'analyse ont été considérés comme porteurs d'une information censurée à droite. Les résultats ont montré que le risque instantané décroissait quant l'âge à l'entrée en station ou le degré d'hétérozygotie lié au croisement croissaient. Les effets additifs des différentes sources de gènes importés n'ont pas eu d'influence. Le risque instantané de maladie a été trouvé soumis à des influences génétiques et non génétiques (troupeau d'origine et année-saison). © Inra/Elsevier, Paris

analyse de survie / modèle semi-paramétrique / échantillonnage de Gibbs / modèle animal / résistance aux maladies

1. INTRODUCTION

When survival data, the time until a certain event happens, is analysed, very often the hazard function is modelled. The hazard function, $\lambda_i(t)$, of an animal i, denotes the instantaneous probability of failing at time t, if risk exists.

In Cox's proportional hazards model [5] it is assumed that $\lambda_i(t) = \lambda_0(t) \exp\{\mathbf{x}_i'\beta\}$, where, in semiparametric models, $\lambda_0(t)$ is any arbitrary baseline hazard function common to all animals. Covariates of animal i, $\mathbf{x}_i$, are supposed to act multiplicatively on the hazard function by $\exp\{\mathbf{x}_i'\beta\}$, where β is a vector of regression parameters. In fully parametric models the baseline hazard function is also parameterized. The proportional hazard model assumes that conditional on covariates, the event times are independent and attention is focused on the effects of the explanatory variables. The baseline hazard function is then regarded as a nuisance factor.

Frailty models are mixed models for survival data. In frailty models it is assumed that there is an unobserved random variable, a frailty variable, which is assumed to act multiplicatively on the hazard function. Sometimes a frailty variable is introduced to make correct inference on regression parameters. In other situations the parameters of the frailty distribution are of major interest.

In shared frailty models, introduced by Vaupel et al. [32], groups of individuals (or several survival times on the same individual) share the same frailty variable. Frailties of two individuals have a correlation equal to 1 if they come from the same group and equal to 0 if they come from different groups. Mainly for reasons of mathematical convenience, the frailty variable is often assumed to follow a gamma distribution. In the animal breeding literature, this method has been used to fit sire models for survival data using fully parametric models (e.g. [8, 10]).

Several papers deal with correlated gamma frailty models (e.g. [22, 26, 30, 31]). In these models individual frailties are linear combinations of independent gamma distributed random variables constructed to give the desired variance covariance matrix among frailties. From a mathematical point of view these models are convenient because the EM algorithm [7] can be used to estimate the parameters. Because of the infinitesimal model often assumed in quantitative genetics, frailties may be log normally distributed; thereby conditional random effects act multiplicatively on the baseline hazard as do covariates. It is not

immediate to use the EM algorithm in log normally distributed frailty models as stated by several authors and shown in Korsgaard [21].

In this paper we show how a full Bayesian analysis can be carried out in a semiparametric log normal frailty model using Gibbs sampling and adaptive rejection sampling. It is shown that by using data augmentation, marginal posterior distributions of breeding values of animals without records can be obtained. The work is very much inspired by the works of Kalbfleisch [19], Clayton [4], Gauderman and Thomas [11] and Dellaportas and Smith [6]. Kalbfleisch [19] presented a Bayesian analysis of the semiparametric regression model. Gibbs sampling was used by Clayton [4] for Bayesian inference in the semiparametric gamma frailty model and by Gauderman and Thomas [11] for inference in a related semiparametric log normal frailty model with emphasis on applications in genetic epidemiology. Finally Dellaportas and Smith [6] demonstrated that Gibbs sampling in conjunction with adaptive rejection sampling gives a straightforward computational procedure for Bayesian inferences in the Weibull proportional hazards model.

The semiparametric log normal frailty model is defined in section 2 of this paper. In this part we show how a full Bayesian analysis is carried out in the special case of the log normal frailty model, where the model of log frailty is a variance component model. The full conditional posterior distributions required for using Gibbs sampling are derived for a given set of prior distributions. In section 3, we analyse disease data on performance tested bulls as an example and section 4 contains a discussion.

2. BAYESIAN INFERENCE IN THE SEMIPARAMETRIC LOG NORMAL FRAILTY MODEL – USING GIBBS SAMPLING

Let T_i and C_i be the random variables representing the survival time and the censoring time of animal i, respectively. Then data on animal i are (y_i, δ_i), where y_i is the observed value of $Y_i = \min\{T_i, C_i\}$ and δ_i is an indicator random variable, equal to 1 if $T_i \leq C_i$, and 0 if $C_i < T_i$. In the semiparametric frailty model, it is assumed that, conditional on frailty $Z_i = z_i$, the hazard function, $\lambda_i(t)$, of T_i; $i = 1, \ldots, n$, is given by

$$\lambda_i(t) = \lambda_h(t) z_i \exp\{\mathbf{x}_i'(t)\beta\} \tag{1}$$

where $\lambda_h(t)$ is the common baseline hazard function of animals that belong to the hth stratum, $h = 1, \ldots, H$, where H is the number of strata. $\mathbf{x}_i(t)$ is a vector of possible time-dependent covariates of animal i and β is the corresponding vector of regression parameters. Z_i is the frailty variable of animal i. This is an unobserved random variable assumed to act multiplicatively on the hazard function. A large value, z_i, of Z_i increases the hazard of animal i throughout the whole time period.

Definition: let $\mathbf{w} = (w_1, \ldots, w_n)'$; if $\mathbf{w}|\boldsymbol{\Sigma} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$ and the frailty variable Z_i in equation (1) be given by $Z_i = \exp\{w_i\}$, i.e. Z_i is log normally distributed; $i = 1, \ldots, n$. Then the model given by equation (1) is called a semiparametric log normal frailty model.

This is the definition of a semiparametric log normal frailty model in broad generality. However, special attention is given to a subclass of models where the distribution of log frailty is given by a variance component model:

$$\mathbf{w} = \mathbf{Q}_u\mathbf{u} + \mathbf{Q}_a\mathbf{a} + \mathbf{e} \tag{2}$$

or in scalar form, $w_i = u_j + a_i + e_i$ where j is the class of the random effect, $\mathbf{u}$, that animal i belongs to; $j \in \{1, \ldots, q\}$. a_i is the random additive genetic value and e_i the random value of environmental effect not already taken into account. It is assumed that $\mathbf{u}|\sigma_u^2 \sim N_q(\mathbf{0}, \mathbf{I}_q\sigma_u^2)$, $\mathbf{a}|\sigma_a^2 \sim N_N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ and $\mathbf{e}|\sigma_e^2 \sim N_n(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$. $\mathbf{Q}_u$ and $\mathbf{Q}_a$. $\mathbf{Q}_u$ and $\mathbf{Q}_a$ are known design matrices of dimension $n \times q$ and $n \times N$, respectively, where N is the total number of animals defining the additive genetic relationship matrix, $\mathbf{A}$, and n is the number of animals with records. Here, $(\mathbf{u}, \sigma_u^2)$, $(\mathbf{a}, \sigma_a^2)$ and $(\mathbf{e}, \sigma_e^2)$ are assumed to be mutually independent. Generalizations will be discussed later. From equation (2), the hazard of T_i is:

$$\lambda_i(t) = \lambda_0(t)\exp\{\mathbf{x}_i'\beta + u_j + a_i + e_i\} \tag{3}$$

assuming that the covariates are time independent and that there is no stratification. The vector of parameters and hyperparameters of the model is $\psi = (\Lambda_0(\cdot), \beta, \mathbf{u}, \sigma_u^2, \mathbf{a}, \sigma_a^2, \mathbf{e}, \sigma_e^2)$, where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the integrated hazard function.

Note that log frailty, w_i, of animal i, is an unobserved quantity which is modelled. This is analogous to the threshold model (e.g. [28]), where an unobserved quantity, the liability, is modelled. In the threshold model, a categorical trait is considered, but heritability is defined for the liability of the trait. In the semiparametric log normal frailty model the trait is a survival time, but heritability is defined for log frailty of the trait. The semiparametric log normal frailty model is not a log linear model for the survival times T_i, $i = 1, \ldots, n$. The only log linear models that are also proportional hazards models are the Weibull regression models (including exponential regression models), where the error term is ε/p, with p being a parameter of the Weibull distribution and having the extreme value distribution [20]. Without restriction on the baseline hazard, the proportional hazard model postulates no direct relationship between covariates (and frailty) and time itself. This is unlike the threshold model, where the observed value is determined by a grouping on the underlying scale.

2.1. Prior distributions

In order to carry out a full Bayesian analysis, the prior distributions of all parameters and hyperparameters in the model must be specified. A priori, it is assumed (by definition of the log normal frailty model) that $\mathbf{u}$, given the hyperparameter σ_u^2, follows a multivariate normal distribution: $\mathbf{u}|\sigma_u^2 \sim N_q(\mathbf{0}, \mathbf{I}_q\sigma_u^2)$. Similarly, it is assumed that $\mathbf{a}|\sigma_a^2 \sim N_N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ and $\mathbf{e}|\sigma_e^2 \sim N_n(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$. A

priori elements in β are assumed to be independent and each is assumed to follow an improper uniform distribution over the real numbers; i.e. $p(\beta_b) \propto 1$; $b = 1, \ldots, B$, where B is the dimension of β. The hyperparameters σ_u^2, σ_a^2 and σ_e^2 are assumed to follow independent inverse gamma distributions; i.e. $\sigma_u^2 \sim IG(\mu_u, \nu_u)$, $\sigma_a^2 \sim IG(\mu_a, \nu_a)$ and $\sigma_e^2 \sim IG(\mu_e, \nu_e)$, where μ_u, ν_u, μ_a, ν_a, and μ_e, ν_e, are values assigned according to prior belief. The convention used for inverse gamma distributions is given in the Appendix. The baseline hazard function $\lambda_0(t)$ will be approximated by a step function on a set of intervals defined by the different ordered survival times, $0 < t_{(1)} < \ldots < t_{(M)} < \infty$: $\lambda_0(t) = \lambda_{0m}$ for $t_{(m-1)} < t \leq t_{(m)}$; $m = 1, \ldots, M$, with $t_{(0)} = 0$ and M the number of different uncensored survival times. The integrated hazard function is then continuous and piecewise linear. A priori it is assumed that $\lambda_{01}, \ldots, \lambda_{0M}$ are independent and that the prior distribution of λ_{0m} is given by $p(\lambda_{0m}) \propto \lambda_{0m}^{-1}$; $m = 1, \ldots, M$. The prior distribution of $\Lambda_{0m} = \Lambda_0(t_{(m)}) - \Lambda_0(t_{(m-1)}) = \lambda_{0m}(t_{(m)} - t_{(m-1)})$ is then $p(\Lambda_{0m}) \propto (\Lambda_{0m})^{-1}$ and $p(\Lambda_{01}, \ldots, \Lambda_{0M}) \propto \prod_{m=1}^{M} \Lambda_{0m}^{-1}$, by having assumed independence of $\lambda_{01}, \ldots, \lambda_{0M}$ a priori. Based on these assumptions and, assuming furthermore that a priori $(\Lambda_{01}, \ldots, \Lambda_{0M})$, β, $(\mathbf{u}, \sigma_u^2)$, $(\mathbf{a}, \sigma_a^2)$ and $(\mathbf{e}, \sigma_e^2)$ are mutually independent, the prior distribution of ψ can be written

$$
\begin{aligned}
p(\psi) &= p(\Lambda_{01}, \ldots, \Lambda_{0M}) \left[\prod_{b=1}^{B} p(\beta_b) \right] p(\mathbf{u}|\sigma_u^2) p(\sigma_u^2) p(\mathbf{a}|\sigma_a^2) p(\sigma_a^2) p(\mathbf{e}|\sigma_e^2) p(\sigma_e^2) \\
&\propto \left[\prod_{m=1}^{M} \Lambda_{0m}^{-1} \right] \times (\sigma_u^2)^{-q/2} \exp\left\{ -\frac{1}{2\sigma_u^2} \mathbf{u}'\mathbf{u} \right\} (\sigma_u^2)^{-(\mu_u+1)} \exp\left\{ -(\sigma_u^2 \nu_u)^{-1} \right\} \\
&\times |\mathbf{A}|^{-1/2} (\sigma_a^2)^{-N/2} \exp\left\{ -\frac{1}{2\sigma_a^2} \mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \right\} (\sigma_a^2)^{-(\mu_a+1)} \exp\left\{ -(\sigma_a^2 \nu_a)^{-1} \right\} \\
&\times (\sigma_e^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_e^2} \mathbf{e}'\mathbf{e} \right\} (\sigma_e^2)^{-(\mu_e+1)} \exp\left\{ -(\sigma_e^2 \nu_e)^{-1} \right\} \qquad (4)
\end{aligned}
$$

2.2. Likelihood and joint posterior distribution

The usual convention that survival times tied to censoring times, precede the censoring times is adopted. Furthermore, as in Breslow [3], it is assumed that censoring occurring in the interval $[t_{(m-1)}, t_{(m)})$ occurs at $t_{(m-1)}$; $m = 1, \ldots, M+1$, with $t_{(M+1)} = \infty$.

Under the assumption, where, conditional on $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{e}$, censoring is independent (e.g. [1, 2]), the partial conditional (censoring omitted) likelihood is given by

$$
p((\mathbf{y}, \delta)|\psi) = \prod_{i=1}^{n} \left[(\lambda_0(y_i) z_i \exp\{\mathbf{x}_i'\beta\})^{\delta_i} \exp\left\{ -\int_0^{y_i} z_i \exp\{\mathbf{x}_i'\beta\} \lambda_0(t) \mathrm{d}t \right\} \right] \qquad (5)
$$

(e.g. [15]). Under the assumptions given above, equation (5) becomes

$$\left[\prod_{m=1}^{M} \lambda_{0m}^{d(t_{(m)})}\right] \times \left[\prod_{m=1}^{M} \prod_{i \in D(t_{(m)})} (z_i \exp\{\mathbf{x}_i'\beta\})\right] \times \exp\left\{-\sum_{m=1}^{M} \Lambda_{0m} \sum_{i \in R(t_{(m)})} z_i \exp\{\mathbf{x}_i'\beta\}\right\} \quad (6)$$

where $D(t_{(m)})$ is the set of animals that failed at time $t_{(m)}$, $d(t_{(m)})$ is the number of animals that failed at time $t_{(m)}$, and $R(t_{(m)})$ is the set of animals at risk of failing at time $t_{(m)}$. Furthermore assuming that, conditional on $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{e}$, censoring is non-informative for ψ, then the joint posterior distribution of ψ is, using Bayes' theorem, obtained up to proportionality by multiplying the conditional likelihood and the prior distribution of ψ

$$p(\psi|(\mathbf{y}, \delta)) \propto p((\mathbf{y}, \delta)|\psi)p(\psi) \quad (7)$$

where $p((\mathbf{y}, \delta)|\psi)$ is the conditional likelihood given by equation (6) and $p(\psi)$ is the prior distribution of parameters and hyperparameters given by equation (4).

2.3. Marginal posterior distributions and Gibbs sampling

If φ is a parameter or a subset of parameters of interest from ψ, the marginal posterior distribution of φ is obtained by integrating out the remaining parameters from the joint posterior distribution. If this can not be performed analytically for one or more parameters of interest, Gibbs sampling [12, 14] can be used to obtain samples from the joint posterior distribution, and thereby also from any marginal posterior distribution of interest. Gibbs sampling is an iterative method for generation of samples from a multivariate distribution which has its roots in the Metropolis–Hastings algorithm [17, 24]. The Gibbs sampler produces realizations from a joint posterior distribution by sampling repeatedly from the full conditional posterior distributions of the parameters in the model. Geman and Geman [14] showed that, under mild conditions, and after a large number of iterations, samples obtained are from the joint posterior distribution.

2.4. Full conditional posterior distributions

In order to implement the Gibbs sampler, the full conditional posterior distributions of all the parameters in ψ must be derived. The following notation is used: that $\psi_{\setminus\varphi}$ denotes ψ except φ; e.g. if $\varphi = \beta$, then $\psi_{\setminus\beta}$ is $(\Lambda_{01}, \ldots, \Lambda_{0M}, \mathbf{u}, \sigma_u^2, \mathbf{a}, \sigma_a^2, \mathbf{e}, \sigma_e^2)$. The full conditional posterior distribution of φ given data and all the remaining parameters, $\psi_{\setminus\varphi}$, is proportional to the joint posterior distribution of ψ given by equation (7).

From equation (7) it then follows that the full conditional posterior distribution of u_j, $j = 1, \ldots, q$ up to proportionality is given by

$$p(u_j|(\mathbf{y}, \delta), \psi_{\setminus u_j}) \propto \exp\{u_j d(u_j)\} \exp\{-u_j^2/(2\sigma_u^2)\} \exp\left\{-\exp\{u_j\} \sum_{i \in S(u_j)} \Omega_i^u\right\} \quad (8)$$

where $\Omega_i^u = \sum_{m:t_{(m)} \le y_i} \exp\{a_i + e_i + \mathbf{x}_i'\beta\}\Lambda_{0m}$ and $d(u_j)$ is the number of animals that failed from the jth class of $\mathbf{u}$ and $S(u_j)$ is the set of animals belonging to the jth class of $\mathbf{u}$. For i, an animal with records, the full conditional posterior distribution of a_i is given by

$$p(a_i|(\mathbf{y},\delta),\psi_{\setminus a_i}) \propto \exp\left\{a_i\left(\delta_i - \frac{1}{\sigma_a^2}\sum_{\substack{j=1\\ j\neq i}}^{N} A^{ij}a_j\right)\right\}\exp\left\{-a_i^2\frac{A^{ii}}{2\sigma_a^2}\right\}\exp\{-\exp(a_i)\Omega_i^a\} \quad (9)$$

where $\Omega_i^a = \sum_{m:t_{(m)} \le y_i} \exp\{u_j + e_i + \mathbf{x}_i'\beta\}\Lambda_{0m}$ and $\{A^{ij}\}$ are the elements of $\mathbf{A}^{-1}$. For an animal, i, without record, the full conditional posterior distribution of a_i follows a normal distribution according to

$$a_i|(\mathbf{y},\delta),\psi_{\setminus a_i} \sim N\left(-\frac{1}{A^{ii}}\sum_{\substack{j=1\\ j\neq i}}^{N} A^{ij}a_j,\quad \sigma_a^2/A^{ii}\right) \quad (10)$$

The full conditional posterior distribution of e_i, $i = 1,\ldots,n$, is, up to proportionality, given by

$$p(e_i|(\mathbf{y},\delta),\psi_{\setminus e_i}) \propto \exp\{e_i\delta_i\}\exp\{-e_i^2/(2\sigma_e^2)\}\exp\{-\exp(e_i)\Omega_i^e\} \quad (11)$$

where $\Omega_i^e = \sum_{m:t_{(m)} \le y_i} \exp\{u_j + a_i + \mathbf{x}_i'\beta\}\Lambda_{0m}$, and the full conditional posterior distribution of each regression parameter β_b, $b = 1,\ldots,B$ is given by

$$p(\beta_b|(\mathbf{y},\delta),\psi_{\setminus\beta_b}) \propto \exp\left\{\beta_b\sum_{i=1}^{n} x_{ib}\delta_i\right\}\exp\left\{-\sum_{i=1}^{n}\sum_{m:t_{(m)} \le y_i}\exp\{u_j + a_i + e_i + \mathbf{x}_i'\beta\}\Lambda_{0m}\right\} \quad (12)$$

The full conditional posterior distribution of each of the hyperparameters σ_u^2, σ_a^2 and σ_e^2 is inverse gamma, according to:

$$\sigma_u^2|\mathbf{y},\delta,\psi_{\setminus\sigma_u^2} \sim IG\left(q/2 + \mu_u, (1/\nu_u + \mathbf{u}'\mathbf{u}/2)^{-1}\right) \quad (13)$$

and

$$\sigma_a^2|\mathbf{y},\delta,\psi_{\setminus\sigma_a^2} \sim IG\left(N/2 + \mu_a, (1/\nu_a + \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}/2)^{-1}\right) \quad (14)$$

$$\sigma_e^2|\mathbf{y},\delta,\psi_{\setminus\sigma_e^2} \sim IG\left(n/2 + \mu_e, (1/\nu_e + \mathbf{e}'\mathbf{e}/2)^{-1}\right) \quad (15)$$

and the full conditional posterior distribution of Λ_{0m}, $m = 1,\ldots,M$, is gamma:

$$\Lambda_{0m}|\mathbf{y},\delta,\psi_{\setminus\Lambda_{0m}} \sim \Gamma\left(d(t_{(m)}), \left[\sum_{i\in R(t_{(m)})}\exp\{u_j + a_i + e_i + \mathbf{x}_i'\beta\}\right]^{-1}\right) \quad (16)$$

Sampling from gamma, inverse gamma and normalely distributed random variables is straightforward. The full conditional posterior distribution of u_j, of a_i, for i, an animal with records, of e_i and of regression parameters, given by equations (8), (9), (11) and (12), respectively, can all be shown [21] to be log concave, and therefore adaptive rejection sampling [16] can be used to sample from these distributions. Adaptive rejection sampling is useful in order to sample efficiently from densities of complicated algebraic form. It is a method for rejection sampling from any univariate log-concave probability density function, which need only be specified up to proportionality.

3. AN EXAMPLE

3.1. Data

As an example, disease data on future AI-bulls from the Danish performance testing programme for beef traits of dairy and dual purpose breeds were analysed. The trait considered was 'time from entering test until first time a respiratory disease occurred'. The bulls of the Danish Red breed were all performance tested in the 15-year period 1982–1996 and entered the Aalestrup test station between 23 and 74 days of age. Bulls which did not experience a respiratory disease during the test period or which were still undergoing testing, on the date of data analysis have right censored records. For these animals, it is only known that the time at first occurrence of a respiratory disease, T_i, will be greater than the time at censoring, C_i, that is, either the time at the end of the test (336 days of age) or the time at the date of data analysis or the time at being culled before end of test (a very rare event). Data on animal i; $i = 1, \ldots, n$ is (y_i, δ_i), where y_i is the observed value of $Y_i = \min\{T_i, C_i\}$ and δ_i is a random indicator variable, equal to 1 if a respiratory disease occurred during test, and 0 otherwise. Data on all animals is $(\mathbf{y}, \delta)$.

3.2. Model

It is assumed that the hazard function, $\lambda_i(t)$, of T_i, is given by

$$\lambda_i(t) = \lambda_0(t)\exp\{\mathbf{x}_i'\beta + h_j + s_k + a_i + e_i\} \tag{17}$$

where t is time (in days) from entering test. In (17), $\lambda_0(t)$ is the baseline hazard function; $\mathbf{x}_i' = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ is a vector of covariates of animal i; x_{i1} ranges between 23 and 74 days of age in the data and is the animal's age at entering test; x_{i2} ranges between 0.0 and 1.0 and x_{i3} ranges between 0.0 and 0.78125 and are proportions of genes from foreign populations (American Brown Swiss and Red Holstein cattle) and x_{i4} (which ranges between 0.0 and 1.0) is the degree of heterozygosity due to crossbreeding. x_{i1} is included in order to take into account that bulls are entering test at different ages; x_{i2} and x_{i3} in order to take additive effects of gene importation into account and x_{i4} in order to take account of heterosis due to dominance. $\beta' = (\beta_1, \ldots, \beta_4)$ is the corresponding vector of regression parameters. $Z_i = \exp\{h_j + s_k + a_i + e_i\}$ is the log normally distributed frailty variable of animal i. h_j is the effect of the jth herd of origin by period combination (one period is 5 years), $j = 1, \ldots, J$, where J is the

number of herd of origin by period combination, and s_k is the effect of entering test in the kth yearseason (one season is 1 month), $k = 1, \ldots, K$, where K is the number of yearseasons. a_i is an additive genetic effect of animal i and e_i is an effect of environment not already taken into account; $i = 1, \ldots, n$, where n is the number of animals with records. In this example J is 540, K is 170 and n is 1 635. The relationship among the test bulls was traced back as far as possible, leading to a total of $N = 5\,083$ animals defining the additive genetic relationship matrix.

3.3 Implementation of the Gibbs sampler and results

The Gibbs sampler was implemented with prior distributions according to the previous section. The prior distributions of the hyperparameters σ_h^2, σ_s^2, σ_a^2 and σ_e^2 were given by inverse gamma distributions with parameters

$$(\mu_h, \nu_h) = (\mu_a, \nu_a) = (2.000001, (0.1 \times 1.000001)^{-1})$$

and

$$(\mu_s, \nu_s) = (\mu_e, \nu_e) = (2.000064, (0.8 \times 1.000064)^{-1})$$

That is, the prior means were of σ_h^2 and σ_a^2 were 0.1 and the prior means of σ_s^2 and σ_e^2 were 0.8. The prior variance of all the hyperparameters is 10 000. The following starting values were assigned to the parameters $\mathbf{h}^{(0)} = (0, \ldots, 0)'$, $\sigma_h^{2(0)} = 0.1$, $\mathbf{s}^{(0)} = (0, \ldots, 0)'$, $\sigma_s^{2(0)} = 0.8$, $\mathbf{a}^{(0)} = (0, \ldots, 0)'$, $\sigma_a^{2(0)} = 0.1$, $\mathbf{e}^{(0)} = (0, \ldots, 0)'$, $\sigma_e^{2(0)} = 0.8$, $\beta^{(0)} = (0, 0, 0, 0)'$. Sampling was carried out from the respective full conditional posterior distributions in the following order, describing one round of the Gibbs sampler:

1) sample Λ_{0m}; $m = 1, \ldots, M$ from the gamma distribution given by equation (16);

2) sample h_j; $j = 1, \ldots, J$ from equation (8) with $u_j = h_j$ and

$$\Omega_i^u = \sum_{m:t_{(m)} \leq y_i} \Lambda_{0m} \exp\{s_k + a_i + e_i + \mathbf{x}_i' \beta\}$$

using adaptive rejection sampling;

3) sample σ_h^2 from the inverse gamma distribution given by equation (13) with, $\sigma_u^2 = \sigma_h^2$, $q = J$, $\mathbf{u} = \mathbf{h}$ and $(\mu_u, \nu_u) = (\mu_h, \nu_h)$;

4) sample a_i from the normal distribution given by equation (10) if i is an animal without records; if i is an animal with records, a_i is sampled from equation (9) with $h_j + s_k$ substituted for u_j in Ω_i^a and using adaptive rejection sampling;

5) sample σ_a^2 from the inverse gamma distribution given by equation (14);

6) sample e_i; $i = 1, \ldots, n$ from equation (11) with $h_j + s_k$ substituted for u_j in Ω_i^e using adaptive rejection sampling;

7) sample σ_e^2 from the inverse gamma distribution given by equation (15);

8) sample β_b; $b = 1, 2, 3, 4$ from equation (12) with $h_j + s_k$ substituted for u_j using adaptive rejection sampling;

9) sample s_k; $k = 1, \ldots, K$ from equation (8) with $u_j = s_k$ and

$$\Omega_i^u = \sum_{m:t_{(m)} \leq y_i} \Lambda_{0m} \exp\{h_j + a_i + e_i + \mathbf{x}_i' \beta\}$$

using adaptive rejection sampling;

10) sample σ_s^2 from the inverse gamma distribution given by (13) with $\sigma_u^2 = \sigma_s^2$, $q = K$, $\mathbf{u} = \mathbf{s}$ and $(\mu_u, \nu_u) = (\mu_s, \nu_s)$.

After 40 000 rounds of the Gibbs sampler, 8 000 samples of model parameters were saved with a sampling interval of 20; i.e. a total chain length of 200 000. After each round of the Gibbs sampler, the following standardized parameters, of log frailty, were computed

$$h^2 = \sigma_a^2/\sigma^2, \; c_h^2 = \sigma_h^2/\sigma^2, \; c_s^2 = \sigma_s^2/\sigma^2 \quad \text{and} \quad e^2 = \sigma_e^2/\sigma^2$$

where $\sigma^2 = \sigma_h^2 + \sigma_s^2 + \sigma_a^2 + \sigma_e^2$ is the variance of log frailty (not of survival time). Summary statistics of selected parameters are shown in *table I.*

Table I. Means and standard deviations of the marginal posterior distributions of σ_a^2, standardized parameters h^2, c_h^2, c_s^2 and e^2 (of log frailty), and regression parameters β_1, β_2, β_3 and β_4, where β_1 is the regression parameter of age at entering test, β_2 and β_3 are the regression parameters of proportions of genes from American Brown Swiss and Red Holstein cattle, respectively, and β_4 is the regression parameter of heterosis due to dominance. N_e is the effective sample size. Lag 1 and Lag 10 are lag 1 and lag 10 correlations of saved sampled values; corresponding to 20 and 200 rounds of the Gibbs sampler, respectively.

Parameter	Mean	Std.	N_e	Lag1	Lag10
σ_a^2	0.0662	0.0396	89	0.978	0.822
h^2	0.1406	0.0735	94	0.941	0.710
c_h^2	0.0913	0.0434	632	0.702	0.096
c_s^2	0.2766	0.0658	603	0.471	0.181
e^2	0.4915	0.1029	286	0.787	0.283
β_1	−0.0061	0.0029	1448	0.484	−0.006
β_2	−0.0026	0.1996	2118	0.308	0.028
β_3	−0.0655	0.3302	8167	0.152	−0.011
β_4	−0.2233	0.1526	5591	0.212	−0.006

The rate of mixing of the Gibbs sampler was investigated by estimating lag-correlations in a standard time series analysis. Lag 1 and lag 10 correlations (lag 1 corresponds to 20 rounds of the Gibbs sampler) are given in *table I.* N_e is the effective sample size, derived from the method of batching (e.g. [13]). The chain of samples from the marginal posterior distribution of σ_a^2 has very slow mixing properties. This is reflected in the standardized parameters as well, whereas all regression parameters have good mixing properties.

The marginal posterior density of σ_a^2 and the standardized parameters h^2, c_h^2, c_s^2 and e^2 (of log frailty) are shown in *figure 1*. The densities were estimated by the methodology of Scott [27]. The mean of the marginal posterior density of h^2 is 0.14 where h^2 is heritability of log frailty. If herd of origin by period and year by season had been considered as fixed effects rather than random effects, then heritability of log frailty would have been higher.

The marginal posterior densities of β_1, β_2, β_3 and are shown in *figure 2*. Because the marginal posterior mean of β_1, the effect of age at entering test, is negative (-0.0061), the hazard is decreased by increasing age at entering test. That is, for a bull entering test at 23 days of age, the conditional hazard is always $\exp\{-0.0061 \times 23\}/\exp\{-0.0061 \times 74\} = 0.87/0.64 = 1.36$ higher, than that of a bull entering test at 74 days of age; conditional on frailty and other covariates being the same for the two bulls. Similarly, because the marginal posterior mean of β_4, the effect of heterozygosity, is negative (-0.22), the hazard is decreased by increasing the degree of heterozygosity. The marginal posterior means of additive effects of gene immigration from American Brown Swiss and Red Holstein cattle are close to 0.0. The marginal posterior mean of h^2, the heritability of log frailty, is 0.1406; of c_h^2, the proportion of variation in log frailty due to herd of origin by period combination is 0.0913 and of c_s^2, the proportion of variation in log frailty caused by a year by season effect is 0.2766.

4. DISCUSSION

This paper illustrates how a Bayesian analysis can be carried out in the semiparametric log normal frailty model using Gibbs sampling. In the version of the Gibbs sampler implemented here, samples were repeatedly taken from univariate full conditional posterior distributions. This is only one possible implementation. With highly correlated univariate components it could be preferable to sample from the joint conditional posterior distribution of these components. The advantage of this method is its greater speed of convergence to the joint posterior distribution [23]. The methodology is quite general and could obviously be used in full parametric models and in models with stratification and/or time-dependent covariates. It is time consuming to sample thousands of observations from thousands of distributions; but, analysing the relatively small dataset in section 3 left us optimistic about possibilities for analysing larger datasets.

Another possibility is to perform a Bayesian analysis using Laplace integration. This was carried out by Ducrocq and Casella [9] with special emphasis on obtaining the marginal posterior distribution of parameters, τ, of the distribution of frailty terms. Their prior distributions of frailty terms were either gamma or log normal. They point out that the computations may quickly become too heavy and propose to summarize the approximate marginal posterior of τ through its first 3 moments using the Gram–Charlier series expansion of a function. The moments are found using numerical integration based on Gauss–Hermite quadrature followed by an iterative strategy to obtain precise estimates. When frailty terms are gamma distributed, these can be integrated out algebraically to obtain the marginal posterior distribution of the remaining parameters. From a likelihood point of view, gamma distributed frailties can be integrated out algebraically to obtain a likelihood based on observed data.

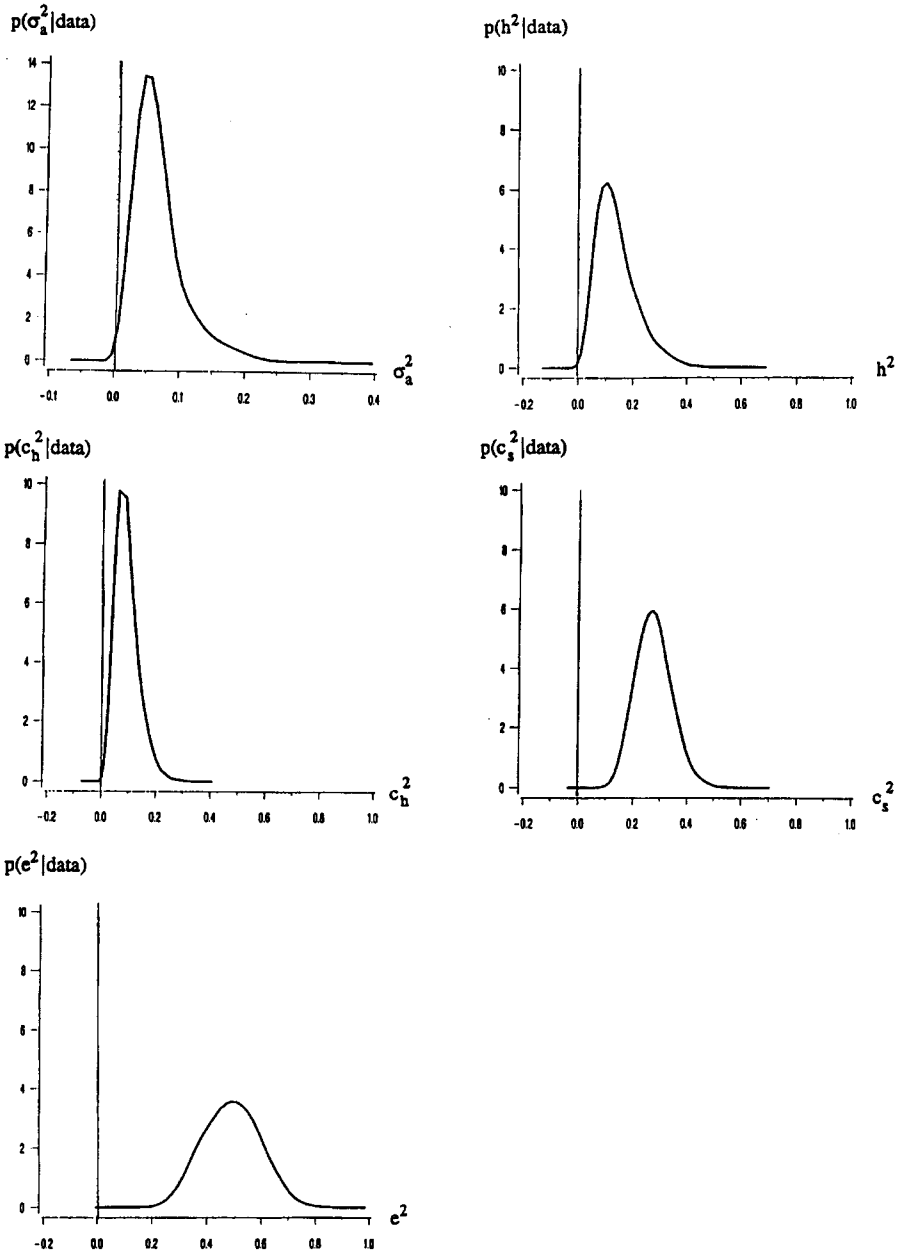

Figure 1. Marginal posterior densities of σ_a^2 and standardized parameters h^2, c_h^2, c_s^2 and e^2 (of log frailty).

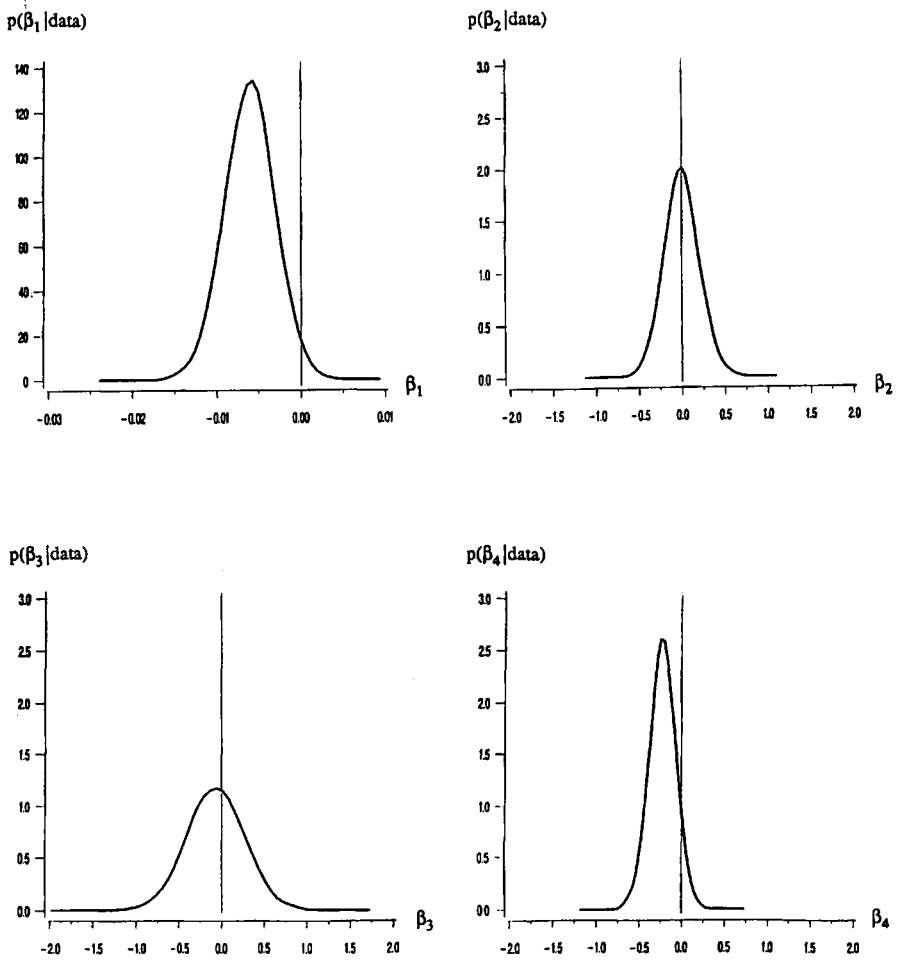

Figure 2. Marginal posterior densities of regression parameters β_1, β_2, β_3 and β_4.

This is so in correlated gamma frailty models as well (e.g. [22]) but not with log normally distributed frailty terms. If τ is a vector of variance components of log frailty terms from a log normal frailty model such as equation (2), Laplace integration may be considered too costly [9]. These authors suggested letting some frailty terms be gamma distributed and others log normally distributed in order to integrate out algebraically some frailty terms. Implementation of the Gibbs sampler in a log normal frailty model avoids making the mathematically tractable but somewhat artificial choice of a gamma distribution of frailty terms.

In this paper some of the parameters are modelled with improper prior distributions. The joint posterior distribution is unavailable in closed form and,

as pointed out by Hobert and Casella [18], the mathematical intractability that necessitates use of the Gibbs sampler also makes demonstrating the propriety of the posterior distribution a difficult task. The fact that the full conditional posterior distributions defining the Gibbs sampler are all proper distributions does not guarantee that the joint posterior distribution will be proper. The question of which improper prior distributions will yield proper joint posterior distributions in hierarchical linear mixed models was addressed by Hobert and Casella [18]. The important question of propriety of the posterior distribution using improper prior distributions must be dealt with in a Bayesian analysis using a Laplace integration as well. One way to avoid improper posteriors is to use proper prior distributions.

The analysis could have been carried out without augmenting [29] by additive genetic values of animals without records. The number of animals defining the additive genetic relationship matrix $\mathbf{A}$ was 5 083, but only 1 635 had records. Let $\widetilde{\mathbf{a}}$ be the vector of additive genetic values of animals with records, then taking the part $\widetilde{\mathbf{A}}$ of $\mathbf{A}$ relating to animals with records, the analysis could have been carried out using the prior $\widetilde{\mathbf{a}}|\sigma_a^2 \sim N_n(\mathbf{0}, \widetilde{\mathbf{A}}\sigma_a^2)$. The number of distributions needing sampling from would have been 5 083–1 635 lower at the expense of obtaining marginal posterior distributions of breeding values on animals without records. However $\widetilde{\mathbf{A}}^{-1}$, is much denser than $\mathbf{A}^{-1}$, and is more difficult to compute.

The model of log frailty, specified by equation (2), can easily be generalized to include more independent variance components and/or the assumption of independence can be relaxed; for example in equation (2), $\mathbf{u}$ could represent a maternal effect. Assuming that $(\mathbf{u}', \mathbf{a}')'|\mathbf{G} \sim N_{2N}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A})$, where $\mathbf{G}$ is the 2×2 matrix of additive genetic covariances, and that the prior distribution of $\mathbf{G}$ is inverse Wishart distributed, then the full conditional posterior distribution of $\mathbf{G}$, required for Gibbs sampling, will also be inverse Wishart distributed. Furthermore, it should be possible to carry out a multivariate analysis of a survival trait, a quantitative trait and a categorical trait, by assuming that the log frailty of the survival trait, the quantitative trait and the liability of the categorical trait follow a multivariate normal distribution. It should also be possible to generalize to an arbitrary number of survival traits, quantitative traits and categorical traits.

By definition, the trait considered in the example: 'time until first occurrence of a respiratory disease during test' can occur at most once for each animal. These data are, however, only a subset of the data being collected on bulls during the test period. Each repeated occurrence of a respiratory disease is recorded, as well as other categories of diseases and several quantitative traits and other traits of interest. Oakes [25] gives a survey of frailty models for multiple event time data, and it would be interesting to extend the log normal frailty models to allow for multiple survival times for each animal as well as a multivariate analysis of censored survival time data and other traits of interest.

ACKNOWLEDGEMENTS

We wish to thank Anders Holst Andersen, Department of Theoretical Statistics, University of Aarhus and Daniel Sorensen, Department of Animal Breeding and Genetics, Research Centre Foulum, for carefully reading several drafts of this manuscript.

REFERENCES

[1] Andersen P.K., Borgan Ø., Gill R.D., Keiding N., Statistical Models Based on Counting Processes, Springer, New York, 1992.

[2] Arjas E., Haara P., A marked point process approach to censored failure data with complicated covariates, Scand. J. Stat. 11 (1984) 193–209.

[3] Breslow N.E., Covariance analysis of censored survival data, Biometrics 30 (1974) 89–99.

[4] Clayton D.G., A Monte Carlo method for Bayesian inference in frailty models, Biometrics 47 (1991) 467–485.

[5] Cox D.R., Regression models and life tables (with discussion), J. R. Stat. Soc. B 34 (1972) 187–220.

[6] Dellaportas P., Smith A.F.M., Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, Appl. Statist. 42 (1993) 443–459.

[7] Dempster A.P., Laird N.M., Rubin D.B., Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. R. Stat. Soc. Ser. B 39 (1977) 1–38.

[8] Ducrocq V., Statistical analysis of length of productive life for dairy cows of the Normande breed, J. Dairy Sci. 77 (1994) 855–866.

[9] Ducrocq V., Casella G., A Bayesian analysis of mixed survival models, Genet. Sel. Evol. 28 (1996) 505–529.

[10] Ducrocq V., Quaas R.L., Pollak E.J., Casella G., Length of productive life of dairy cows. 2. Variance component estimation and sire evaluation, J. Dairy Sci. 71 (1988) 3071–3079.

[11] Gauderman W.J., Thomas D.C., Censored survival models for genetic epidemiology: A Gibbs sampling approach, Genet. Epidemiol. 11 (1994) 171–188.

[12] Gelfand A.E., Smith A.F.M., Sampling-based approaches to calculating marginal densities, J. Am. Stat. Assoc. 85 (1990) 398–409.

[13] Gelman A., Carlin J.B., Stern H.S., Rubin D.B., Bayesian Data Analysis, Chapman and Hall, London, 1995.

[14] Geman S., Geman D., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Analysis and Machine Intelligence 6 (1984) 721–741.

[15] Gill R.D., Johansen S., A survey of product-integration with a view towards application in survival analysis, Ann. Stat. 18 (1990) 1501–1555.

[16] Gilks W.R., Wild P., Adaptive rejection sampling for Gibbs sampling, Appl. Stat. 41 (1992) 337–348.

[17] Hastings W.K., Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.

[18] Hobert J.P., Casella G., The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, J. Am. Stat. Assoc. 91 (1996) 1461–1473.

[19] Kalbfleisch J.D., Nonparametric Bayesian analysis of survival time data, J. R. Stat. Soc. B 40 (1978) 214–221.

[20] Kalbfleisch J.D., Prentice R.L., The Statistical Analysis of Failure Time Data, John Wiley, New York, 1980.

[21] Korsgaard I.R., Genetic analysis of survival data, Ph.D. thesis, University of Aarhus, Denmark, 1997.

[22] Korsgaard I.R., Andersen A.H., The additive genetic gamma frailty model, Scand. J. Stat. 25 (1998) 255–269

[23] Liu J.S., Wong W.H., Kong A., Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes, Biometrika 81 (1994) 27–40.

[24] Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E., Equations of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1091.

[25] Oakes D., Frailty models for multiple event-time data, in: Klein J.P., Goel P.K. (eds.), Survival Analysis: State of the Art, Kluwer Academic Publishers, The Netherlands, 1992, pp. 371–380.

[26] Petersen J.H., Andersen P.K., Gill R., Variance component models for survival data, Statistica Neerlandica 50 (1996) 193–211.

[27] Scott D.W., Multivariate Density Estimation: Theory, Applications, and Visualisation, John Wiley, New York, 1992.

[28] Sorensen D.A., Andersen S., Gianola D., Korsgaard I.R., Bayesian inference in threshold models using Gibbs sampling, Genet. Sel. Evol. 27 (1995) 229–249.

[29] Tanner M.A., Wong W.H., The calculation of posterior distributions by data augmentation, J. Am. Stat. Assoc. 82 (1987) 528–540.

[30] Yashin A., Iachine I., Survival of related individuals: an extension of some fundamental results of heterogeneity analysis, Math. Popul. Studies 5 (1995) 321–340.

[31] Yashin A.I., Vaupel J.W., Iachine I.A., Correlated individual frailty: an advantageous approach to survival analysis of bivariate data, Math. Popul. Studies 5 (1995) 145–160.

[32] Vaupel J.W., Manton K.G., Stallard E., The impact of heterogeneity in individual frailty on the dynamics of mortality, Demography 16 (1979) 439–454.

APPENDIX: Convention used for gamma, inverse gamma and log normal distributions

The following convention is used for gamma and inverse gamma distributions: let $X \sim \Gamma(\alpha, \beta)$, with density $p(x) = x^{\alpha-1}\mathrm{e}^{-x/\beta}\left[\Gamma(\alpha)\beta^{\alpha}\right]^{-1}$. Then $E(X) = \alpha\beta$ and $V(X) = \alpha\beta^2 = \beta E(X)$. $Y = X^{-1}$ has an inverse or inverted gamma distribution $Y \sim IG(\alpha, \beta)$ with density $p(y) = y^{-(\alpha+1)}\mathrm{e}^{-(y\beta)^{-1}}$ $\left[\Gamma(\alpha)\beta^{\alpha}\right]^{-1}$ and $E(Y) = [(\alpha-1)\beta]^{-1}$ for $\alpha > 1$ and $V(Y) = [(\alpha-1)^2(\alpha-2)\beta^2]^{-1} = (E(Y))^2(\alpha-2)^{-1}$ for $\alpha > 2$.

If $X \sim N(\mu, \sigma^2)$, then $Y = \exp\{X\}$ is said to have a log normal distribution; the density of Y is given by

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{y}\exp\left\{-\frac{1}{2\sigma^2}(\log(y)-\mu)^2\right\}$$