Approximate Bayesian Inference for Survival Models

Author(s): SARA MARTINO, RUPALI AKERKAR and HÅVARD RUE

Source: *Scandinavian Journal of Statistics*, Vol. 38, No. 3 (September 2011), pp. 514-528
Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics
Stable URL: https://www.jstor.org/stable/23015578
Accessed: 19-05-2019 13:25 UTC

Scandinavian Journal of Statistics, Vol. 38: 514–528, 2011
doi: 10.1111/j.1467-9469.2010.00715.x

Approximate Bayesian Inference for Survival Models

SARA MARTINO, RUPALI AKERKAR and HÅVARD RUE
Department of Mathematical Sciences, NTNU

ABSTRACT. Bayesian analysis of time-to-event data, usually called survival analysis, has received increasing attention in the last years. In Cox-type models it allows to use information from the full likelihood instead of from a partial likelihood, so that the baseline hazard function and the model parameters can be jointly estimated. In general, Bayesian methods permit a full and exact posterior inference for any parameter or predictive quantity of interest. On the other side, Bayesian inference often relies on Markov chain Monte Carlo (MCMC) techniques which, from the user point of view, may appear slow at delivering answers. In this article, we show how a new inferential tool named integrated nested Laplace approximations can be adapted and applied to many survival models making Bayesian analysis both fast and accurate without having to rely on MCMC-based inference.

Key words: approximate inference, Bayesian hazard rate model, geoadditive hazard regression, Laplace approximation, latent Gaussian fields

1. Introduction

Since its introduction in the seminal work of Cox (1972), the proportional hazard or Cox model is the default choice when dealing with continuous time-to-event data. In its basic form, it leaves the baseline hazard function unspecified (thus allowing for some flexibility) but requires all covariates to have linear effects. While classical analysis have to rely on parameter inference based on partial likelihood, and on a postestimate of the baseline hazard, the Bayesian approach allows to use information from the full likelihood and to jointly estimate all unknown elements in the model. In general, Bayesian methods permit a full and exact posterior inference for any parameter or predictive quantity of interest.

The last years have seen an increasing interest in Bayesian analysis of time-to-event data, mainly because of improvements in both modelling techniques and computational power. Several extensions to the basic Cox model have been proposed in the Bayesian literature to account for different characteristic of the data, such as within-group correlation, spatial patterns or non-linear covariate effects. The book by Ibrahim *et al.* (2001) provides a good overview of Bayesian survival models. Banerjee *et al.* (2003) discuss parametric Weibull baseline hazard and adds a spatial component using a geostatistical model, whereas Carlin & Banerjee (2003) and Banerjee & Carlin (2003) do similarly with a semi-parametric estimation of baseline hazard. Hannerfeind *et al.* (2006) extend the work of Fahrmeir & Lang (2001) and Lang & Brezger (2004) and propose a geoadditive Cox model where the linear predictor is extended to include spatial components, unknown form of the (log)baseline hazard and semi-parametric effect of covariates. The spatial effect is modelled via geostatistical and conditional autoregressive priors while B-splines are used to model the unknown smooth functions. Inference is made using Markov chain Monte Carlo (MCMC) algorithms. Kneib (2006) extends the geoadditive Cox model to deal with interval-censored survival times. Kneib & Fahrmier (2007) propose a mixed model-based methodology for geoadditive Cox models, which can be interpreted as an empirical Bayes version of the full Bayesian approach in Hannerfeind *et al.* (2006). A joint frequentistic analysis of survival and longitudinal data

was proposed by Handerson *et al.* (2000), central feature is to postulate a latent bivariate Gaussian process and assume that, given such process, the longitudinal measurements and the survival data are conditionally independent. Guo & Carlin (2004) discuss a Bayesian version of the same model using MCMC for inference.

Although there exist software offering general solutions for wide classes of models, like WinBUGS (Lunn *et al.*, 2000) and BayesX (Brezger *et al.*, 2003), the use of MCMC-based inference still carries a large computational cost and requires interaction from the user to diagnose convergence and accuracy of the estimates. All these additional costs become more prominent when applied to more advanced models including spatial and/or semi-parametric (smooth) effects. We conclude that Bayesian inference for survival models is indeed possible, but the current computational solutions is not yet at the level which gives the end-user a smooth experience, both in terms of speed and simplicity.

The aim of this article is twofold. First, we want to show that many of the Cox-type models proposed in the literature can be seen as *latent Gaussian models*. These are a wide class of statistical models whose latent variables are jointly Gaussian and are partially observed through data. Some hyperparameters might also be present; see Rue *et al.* (2009) for more examples. There are two main advantages in viewing survival models as latent Gaussian models; 'complicated' components in the models, like smooth effects of continuous covariates, spatial/temporal effects, various frailty effects, are easy to add and appear only as trivial changes in the Gaussian part of the model. Moreover, Bayesian inference for latent Gaussian models can be made using *integrated nested Laplace approximations* (INLA), see Rue *et al.* (2009). INLA provides fast and accurate approximations to the posterior marginals through a clever use of Laplace approximations and advanced numerical methods taking computational advantage of sparse matrices. The result is that posterior marginals can be estimated in a small fraction of the time required by MCMC, with a relative error not additive error as for MCMC. Moreover, INLA can be used almost as a black box and does not require interaction from the user to check convergence. To show how INLA can be adapted to fit survival models is the second aim of this article. Although the implementation of INLA is quite involved, an open-source version written in C based on the GMRFLib-library (Rue & Held, 2005) is available. An interface from R named INLA, is also available; see http://www.r-inla.org for documentation and worked through examples. The R codes used to implement all the examples in this article are also available from the same web page.

The rest of the article is organized as follows: in section 2 we introduce latent Gaussian models and give a short description of the INLA approach to approximate posterior marginals. In section 3, we discuss Weibull hazard regression. Increasing the model complexity, we discuss a joint model for longitudinal and survival data in section 4. Semi-parametric models for the baseline hazard are discussed in section 5. We end with a discussion of various forms for censoring in section 6 and a general discussion in section 7.

2. Latent Gaussian models and INLA

In general, latent Gaussian models are hierarchical models where we assume a n-dimensional latent field $\mathbf{x}$ to be point-wise observed through $n_d \leq n$ data $\mathbf{y}$. The latent field $\mathbf{x}$ is assumed to have Gaussian density conditionally on some hyperparameters $\boldsymbol{\theta}_1$,

$$\mathbf{x} \mid \boldsymbol{\theta}_1 \sim \mathcal{N}(0, \mathbf{Q}^{-1}(\boldsymbol{\theta}_1)).$$

The data $\mathbf{y}$ are assumed to be conditionally independent given the latent field $\mathbf{x}$ and, possibly, some additional hyperparameters $\boldsymbol{\theta}_2$ in the likelihood. The model definition is completed

with a prior density for the hyperparameters $\boldsymbol{\theta}=\{\theta_1,\theta_2\}$. In addition, some linear constraints of the form $\boldsymbol{A}\mathbf{x}=\mathbf{e}$, where the $h\times n$ matrix $\boldsymbol{A}$ has rank $h\ll n$, may be imposted.

Latent Gaussian models are a subset of all Bayesian structured additive models (see Fahrmeir & Tutz, 2001, for a review). Here the likelihood for the ith observation, $\pi(y_i\,|\,\eta_i,\boldsymbol{\theta}_2)$, depends on some structured additive predictor η_i and, possibly, on hyperparameters $\boldsymbol{\theta}_2$. Rue *et al.* (2009) consider $\pi(y_i\,|\,\eta_i,\boldsymbol{\theta}_2)$ to belong to an exponential family with the mean μ_i linked to the structured predictor η_i through a known link function. In survival analysis applications, the likelihood does not belong to an exponential family but depends on the survival application considered.

The structured predictor η_i accounts for effects of various covariates in an additive way:

$$\eta_i=\beta_0+\sum_{j=1}^{n_f} w_{ij}f^{(j)}(u_{ij})+\sum_{k=1}^{n_\beta}\beta_k z_{ki}+\varepsilon_i. \tag{1}$$

Here, the $\{\beta_k\}$s represent the linear effect of covariates $\mathbf{z}$. The $\{f^{(j)}(\cdot)\}$s are unknown functions of the covariates $\mathbf{u}$: non-linear effects of continuous covariates, time trends, seasonal effects, i.i.d. 'random' intercepts and slopes, group-specific random effects (frailties) and spatial random effects can all be represented through the $\{f^{(j)}\}$ functions. The w_{ij} are known weights defined for each observed data point. Finally, ε_is are unstructured random effects. A latent Gaussian model is obtained by assigning $\mathbf{x}=\{\{f^{(j)}(\cdot)\},\{\beta_k\},\{\eta_i\}\}$, a Gaussian prior with precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$.

The posterior distribution then reads:

$$\pi(\mathbf{x},\boldsymbol{\theta}\,|\,\mathbf{y})\propto\pi(\boldsymbol{\theta})\pi(\mathbf{x}\,|\,\boldsymbol{\theta})\prod_{i\in\mathcal{I}}\pi(y_i\,|\,\mathbf{x},\boldsymbol{\theta}), \tag{2}$$

where the likelihood for y_i depends only on η_i and $\boldsymbol{\theta}_2$. As the likelihood often is not Gaussian, this posterior density is not analytically tractable. The aim is to infer the posterior marginal distributions for the latent field $\pi(x_i\,|\,\mathbf{y})$, $i=1,\ldots,n$ and for the hyperparameters $\pi(\boldsymbol{\theta}\,|\,\mathbf{y})$.

INLA provides a recipe for computing such posterior marginals in a fast and accurate way (Rue *et al.*, 2009). The approximations $\tilde{\pi}(\boldsymbol{\theta}\,|\,\mathbf{y})$ and $\tilde{\pi}(x_i\,|\,\boldsymbol{\theta},\mathbf{y})$, $i=1,\ldots,n$ are based on a clever use of Laplace approximations. Rue *et al.* (2009) describe three different approximations for $\tilde{\pi}(x_i\,|\,\boldsymbol{\theta},\mathbf{y})$, Gaussian, simplified Laplace and Laplace, ordered here by accuracy and computational cost. The default option in the INLA library is the simplified Laplace approximation and this is used in all the examples in this article. Posterior marginals for the latent variables $\tilde{\pi}(x_i\,|\,\mathbf{y})$ are then computed via numerical integration:

$$\begin{aligned}\tilde{\pi}(x_i\,|\,\mathbf{y})&=\int\tilde{\pi}(x_i\,|\,\boldsymbol{\theta},\mathbf{y})\tilde{\pi}(\boldsymbol{\theta}\,|\,\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}\\&\approx\sum_{k=1}^{K}\tilde{\pi}(x_i\,|\,\boldsymbol{\theta}_k,\mathbf{y})\tilde{\pi}(\boldsymbol{\theta}_k\,|\,\mathbf{y})\Delta_k,\end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_k$ are points accurately chosen in the $\boldsymbol{\theta}$ space and Δ_k are integration weights. Posterior marginals for the hyperparameters $\tilde{\pi}(\theta_j\,|\,\mathbf{y})$, $j=1,\ldots,M$ can also be derived by numerical integration. The output of INLA consists of posterior marginal distributions, which can be summarized via means, variances and quantiles. As a bi-product of the main computations, INLA can compute the deviance information criteria (DIC; Spiegelhalter *et al.*, 2002), a measure of complexity and fit useful to compare different models.

For the INLA methodology to work in an efficient way, latent Gaussian models have to satisfy some additional properties which will be assumed throughout this article. First, the latent field $\mathbf{x}$, often of large dimension, admits conditional independence properties. In other words it is a latent Gaussian Markov random field (GMRF) with a sparse precision

matrix **Q** (Rue & Held, 2005). The efficiency of INLA relies, in fact, on algorithms for sparse matrices computation. Almost all latent Gaussian models in the literature satisfy these conditions. The second condition to be satisfied is that the dimension of the hyperparameter vector $\boldsymbol{\theta}$ should not be too large. This is necessary for the integral in (3) to be computationally feasible. Rue *et al.* (2009) propose an integration scheme named central composite design (CCD) which allows to explore the $\boldsymbol{\theta}$ space using a limited number of points. Such scheme, which is the default choice in the INLA package, allows fast computations also for moderate numbers of hyperparameters (say ≤ 20). Finally, each data point y_i should depend on the latent field $\mathbf{x}$ only through the predictor η_i, that is, $\pi(y_i \mid \mathbf{x}, \boldsymbol{\theta}_1) = \pi(y_i \mid \eta_i, \boldsymbol{\theta}_1)$. This is a technical requirement because of the software design of the GMRFLib library upon which the INLA library is based and not a condition necessary to the INLA methodology itself.

3. Parametric proportional hazard models: Weibull regression

Here we consider survival data in their most common form: for each individual i in study the lifetime T_i and the censoring time C_i are independent random variables. The observed time is $t_i = \min(T_i, C_i)$ and δ_i denotes the censoring observation. In the Cox proportional hazard model the hazard rate for individual i is:

$$h_i(t) = h_0(t)\exp(\eta_i), \quad t > 0 \tag{4}$$

where $h_0(\cdot)$ is the baseline hazard, and η_i is the predictor. One common approach is to assume a Weibull distribution for the baseline hazard:

$$h_0(t) = \alpha t^{\alpha - 1}, \quad \alpha > 0. \tag{5}$$

The contribution to the log-likelihood of observation (t_i, δ_i) is:

$$\begin{aligned} l_i &= \delta_i \log h(t_i) - \int_0^{t_i} h(u)\,\mathrm{d}u \\ &= \delta_i(\log\alpha + (\alpha - 1)\log t_i + \eta_i) - \exp(\eta_i)t_i^{\alpha}. \end{aligned} \tag{6}$$

In the basic Cox model we have $\eta_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i$ where $\boldsymbol{\beta}$ is a vector of unknown parameters and $\mathbf{z}_i$ is a vector of observed covariates. Following Hannerfeind *et al.* (2006), we let the predictor η_i take the structured additive form in (1). We assign Gaussian priors to all elements on the right end side of (1), so that $\mathbf{x} = \{\{f^{(j)}(\cdot)\}, \{\beta_k\}, \{\eta_i\}\}$ is a Gaussian field with precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$.

The extended Weibull regression model described thus can easily be seen as a latent Gaussian model with latent field $\mathbf{x}$, hyperparameter vector $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, with $\boldsymbol{\theta}_2 = \alpha$, and observed data (t_i, δ_i), $i = 1, \ldots, n_d$. The likelihood for (t_i, δ_i) depends on the latent field $\mathbf{x}$ only through the predictor η_i, as can be seen from (6); therefore, INLA can be directly applied to such model as shown in the following example.

3.1. *Example: the kidney infection data*

Our first example concerns the well-known study of times to kidney infection for a set of 38 patients (McGilchrist & Aisbett, 1991; Spiegelhalter *et al.*, 1995). The data set contains, for each patient, the first and second infection times t_{ij} and a set of three covariates: sex, age and the type of disease. Spiegelhalter *et al.* (1995) propose a Weibull model to analyse the data

set. Log-normal frailties are used to model the association between the two survival times related to the same patient. The hazard model is:

$$h_{ij}(t) = \alpha t^{\alpha-1} \exp(\eta_{ij}), \quad i = 1, \ldots, 38, \; j = 1, 2,$$

where

$$\eta_{ij} = \beta_0 + \beta_{\text{sex}} \text{sex}_i + \beta_{\text{age}} \text{age}_i + \beta_{\text{dis2}} \text{dis2}_i + \beta_{\text{dis3}} \text{dis3}_i + \beta_{\text{dis4}} \text{dis4}_i + \log(u_i).$$

We assign $b_i = \log(u_i) \sim \mathcal{N}(0, \tau^{-1})$, $\beta_0 \sim \mathcal{N}(0, 0.001^{-1})$ and $\boldsymbol{\beta} = \{\beta_{\text{sex}}, \beta_{\text{age}}, \beta_{\text{dis2}}, \beta_{\text{dis3}}, \beta_{\text{dis4}}\} \sim \mathcal{N}(0, \mathbf{I})$. Further we assume Gamma priors $\Gamma(a, b)$ with mean a/b and variance a/b^2, for both τ and α. In particular, $\tau \sim \Gamma(1, 1)$ and $\alpha \sim \Gamma(1, 1)$.

We implement the model using INLA by defining the formula:

```
formula = inla.surv(time, event) ~ age + sex + dis2 + dis3 + dis4 +
          f(ID, model = "iid", param = c(1, 1))
```

and then using the inla() function:

```
mod = inla(formula, family = "weibull", data = Kidney,
           control.data = list(param = c(1,1)), control.fixed = list(prec = 1))
```

The 'inla.surv()' function is used to describe censored data, and always appears on the left side of a model formula when dealing with survival models. The 'param' argument specifies the parameters a and b in the Gamma priors while the 'prec' argument specifies the precision for the prior of the $\boldsymbol{\beta}$ vector.

In Fig. 1, the INLA posterior marginals for α, τ, β_0 and β_{sex} are compared with histograms based on long MCMC runs using WinBUGS. All examples in the article are implemented on on a dual-core 2.5-GHz laptop and the execution times refer to such machine. The estimates are practically indistinguishable despite the fact that the computing time for inla() was only two seconds while WinBUGS needed around five minutes. Results for the other elements of the $\boldsymbol{\beta}$ vector are similar.

The core of the INLA methodology is a Gaussian approximation of the full conditional density for the latent field, given the data and the hyperparameters. Therefore, for INLA to achieve accurate results we either have to have a proper Gaussian prior for $\mathbf{x}$, or a large enough ratio between the number of data points and the total number of model parameters. In this example, there are relatively few data compared with the total number of parameters (76 observations and a total of 46 parameters) so choosing flat priors makes the problem difficult for INLA. To illustrate, let us use the same priors as in the WinBUGS manual: $\boldsymbol{\beta} \sim \mathcal{N}(0, 10^4 \mathbf{I})$, $\tau \sim \Gamma(10^{-3}, 10^{-3})$ and $\alpha \sim \Gamma(1, 10^{-3})$. The very low precision for $\boldsymbol{\beta}$ and the vague prior assigned to τ make the prior density for $\mathbf{x} \mid \boldsymbol{\theta}$ resembling more a uniform than a normal density. To check how INLA behaves in such a challenging case we have compared INLA and MCMC results.

Now the two estimates present some discrepancy. We quantify them using both the difference between the estimated posterior means relative to the estimated standard deviation $(\mu_{\text{inla}} - \mu_{\text{mcmc}})/\sigma_{\text{mcmc}}$ and the ratio of the estimated standard deviations $(\sigma_{\text{inla}}/\sigma_{\text{mcmc}})$. Results are reported in Table 1. Although there are differences in the two estimates, these are rather small and could be ignored for practical use. Despite this being a quite difficult case for the INLA methodology, we get reliable estimates in only two seconds when the MCMC sampler needed around 10 minutes.

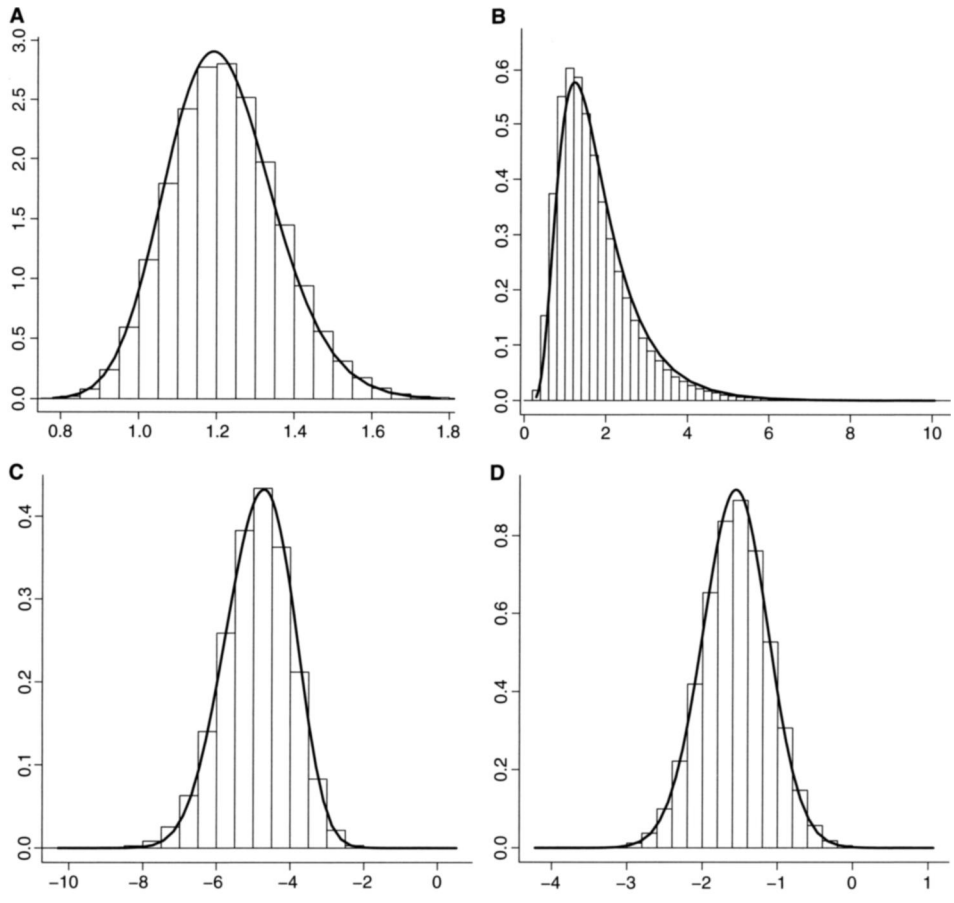

Fig. 1. Kidney example: posterior marginal distributions approximated by integrated nested Laplace approximation (solid line) and Markov chain Monte Carlo-based density estimates (histogram) for α (A), τ (B), β_0 (C) and β_{sex} (D).

Table 1. *The Kidney infection example: comparison between integrated nested Laplace approximations (INLA) and Markov chain Monte Carlo (MCMC)-based estimates in the case of very flat prior densities*

Parameter	$(\mu_{inla} - \mu_{mcmc})/\sigma_{mcmc}$	$\sigma_{inla}/\sigma_{mcmc}$
β_0	0.015	0.897
β_{age}	0.008	0.937
β_{sex}	0.015	0.911
β_{dis2}	−0.002	0.935
β_{dis3}	−0.009	0.935
β_{dis4}	0.007	0.936
α	−0.174	0.800
$\log \tau$	−0.093	0.835

4. Model for joint analysis of survival and longitudinal data

Many scientific investigations generate both longitudinal data (repeated measurement of a response variable at a number of time points) and survival data. Often the longitudinal variable is linked to the mechanism generating the survival data, then joint study of the two data

sets is of interest. A flexible model for such analysis is presented in Henderson *et al.* (2000). The authors argue that a joint model for longitudinal and survival data should incorporate the most commonly used assumption for both subject. Thus, they model longitudinal data by including fixed effects, random effects, serial correlations and pure measurement error, and the survival data by using a parametric proportional hazard with or without frailty terms. The longitudinal and the survival processes are then connected by a latent bivariate Gaussian process and assumed conditional independent given such latent process and any available covariate. While Henderson *et al.* (2000) propose a classical maximum likelihood approach for this model, Guo & Carlin (2004) assume a Bayesian perspective and rely on MCMC algorithms. In this section, we show how this rather complex model reduces again to a latent Gaussian model where the observations have different likelihoods.

4.1. Model specification

Suppose we have a set of m subject followed over a time interval $[0, \tau]$. The ith subject provides a set of (possibly missing in part) longitudinal quantitative measurements $\{y_{ij}, j=1,\ldots,n_i\}$ at times $\{s_{ij}, j=1,\ldots,n_i\}$ and a (possibly) censored survival time t_i. Moreover, a set of covariates $\mathbf{z}$ is recorded. The joint model is composed of two sub-models, one for each type of data.

The longitudinal data y_{ij} are assumed to have a Gaussian likelihood with unknown precision τ_1 and mean:

$$\eta_{ij}^{l} = \mu_i(s_{ij}) + W_{1i}(s_{ij}), \tag{7}$$

where $\mu_i(s) = \mathbf{z}_{1i}^{\mathrm{T}}(s)\boldsymbol{\beta}_1$ and $W_{1i}(s) = \mathbf{d}_{1i}(s)\mathbf{U}_i$ incorporates subject-specific random effects. The vectors $\mathbf{z}_{1i}^{\mathrm{T}}(s)$ and $\boldsymbol{\beta}_1$ represent (possibly time varying) explanatory variables and their corresponding regression coefficients. The $\mathbf{U}_i$ are vectors of Gaussian random effects corresponding to the explanatory variables $\mathbf{d}_{1i}(s)$.

The survival observations t_i, $i=1,\ldots,m$ are assumed to have a Weibull likelihood with unknown shape parameter α and predictor

$$\eta_i^{s} = \mathbf{z}_{2i}^{\mathrm{T}}(t_i)\boldsymbol{\beta}_2 + W_{2i}(t_i), \tag{8}$$

where the vectors $\mathbf{z}_{2i}(t)$ and $\boldsymbol{\beta}_2$ represent (possibly time-dependent) explanatory variables and their corresponding regression coefficients. They may or may not have elements in common with $\mathbf{z}_{1i}$ and $\boldsymbol{\beta}_1$ in the longitudinal model. The form of $W_{2i}(t)$ is similar to $W_{1i}(s)$, including subject-specific covariate effects and intercept (frailty).

Henderson *et al.* (2000) introduce association between models (7) and (8) by using a latent zero-mean bivariate Gaussian process to model $(W_{1i}, W_{2i})^{\mathrm{T}}$. The random variables are assumed to be independent across different subjects. Specifically, they propose:

$$W_{1i}(s) = U_{1i} + U_{2i}s \tag{9}$$

and

$$W_{2i}(t) = \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3(U_{1i} + U_{2i}t) + U_{3i}. \tag{10}$$

The parameters γ_1, γ_2 and γ_3 in (10) measure the association between the two sub-models induced by the random intercepts, slopes and fitted longitudinal value at the event time $W_{1i}(t)$, respectively. The pairs $(U_{1i}, U_{2i})^{\mathrm{T}}$ are assumed to have a bivariate Gaussian distribution $\mathcal{N}(0, \mathbf{Q}_U^{-1})$ while the U_{3i} are independent frailty terms with $\mathcal{N}(0, \tau_{U_3}^{-1})$ prior and independent of the $(U_{1i}, U_{2i})^{\mathrm{T}}$s. Vague Gaussian priors are assigned to $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, γ_1, γ_2 and γ_3 in

(10), while Gamma priors are assigned to τ_1, τ_{U_3} and α. Finally, a Wishart prior is assigned to the 2×2 precision matrix $\mathbf{Q}_U$.

This rather complex model reduces to a latent Gaussian field. Define the vector of hyperparameters to be $\boldsymbol{\theta}=\{\alpha, \tau_1, \tau_{U_3}, \gamma_1, \gamma_2, \gamma_3, \mathbf{Q}_U\}$. The latent field $\mathbf{x}=(\{\eta_{ij}^l\}, \{\eta_i^s\}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \{(U_{1i}, U_{2i})\}, \{U_{3i}\})$, conditioned on $\boldsymbol{\theta}$, has Gaussian distribution with precision matrix $\mathbf{Q}(\boldsymbol{\theta})$. Finally, the observations $(\{y_{ij}\}, \{t_i\})$ have a likelihood which, conditional on the hyperparameters $\boldsymbol{\theta}$, depends on the latent field $\mathbf{x}$ only through the predictor, η_{ij}^l or η_i^s according to the particular data point we are considering. The fact that not all data points have the same likelihood does not pose any challenge to the INLA methodology; each data point could have a different likelihood.

4.2. Example: AIDS clinical trial

We reconsider the data in Guo & Carlin (2004). For each of the $m=467$ patients enrolled in the study, the times to death ($\delta_i=1$) or to leave the study ($\delta_i=0$) are recorded: (t_i, δ_i), $i=1,\ldots,m$. Moreover, the square root of the number of CD4 cells/ml of blood y_{ij} is recorded at time s_{ij}. There is a maximum of five observations per patient. Four explanatory variables are also recorded: the gender, the type of drug, the AIDS diagnosis at study entry (PrevOI) and the stratum.

The longitudinal sub-model assumes a Gaussian likelihood for y_{ij} with unknown precision τ_1 and mean

$$\eta_{ij}=\beta_{11}+\beta_{12}s_{ij}+\beta_{13}s_{ij}\text{drug}_i+\beta_{14}\text{gender}_i+\beta_{15}\text{prevOI}_i+\beta_{16}\text{stratum}_i+W_{1i}(s_{ij}).$$

The survival sub-model assumes an exponential likelihood for (t_i, δ_i) with predictor:

$$\eta_i=\beta_{21}+\beta_{22}\text{drug}_i+\beta_{23}\text{gender}_i+\beta_{24}\text{prevOI}_i+\beta_{25}\text{stratum}_i+W_{2i}(t). \tag{11}$$

Guo & Carlin (2004) propose a variety of joint models with different forms of the latent processes $W_1(s)$ and $W_2(s)$ and compare them using the DIC. This is a measure of complexity and fit, introduced in Spiegelhalter *et al.* (2002) and used to compare complex hierarchical models. It is defined as:

$$\text{DIC}=\bar{D}+p_D,$$

where $\bar{D}$ is the posterior mean of the deviance of the model and p_D is the effective number of parameters. Smaller DIC values indicate a better trade-off between complexity and fit. We have implemented all models in Guo & Carlin (2004), with the exception of models X and XII which, for technical reasons, cannot be implemented in the current version of INLA. The computing time goes from six seconds needed for model I to 206 for model XI.

In Table 2, the computed $\bar{D}$, p_D and DIC are reported for the nine different joint models and model XI emerges with the smallest DIC.

Having selected a final model we compare results obtained under the separate (i.e. ignoring any latent association introduced by W_2) and the joint model. The estimated parameters, together with 95 per cent credible intervals are reported in Table 3. Our results appear to be equal to those obtained by Gibbs sampling in Guo & Carlin (2004) up to two-digit accuracy.

5. Semi-parametric baseline hazard models

In this section, we consider a semi-parametric model for the baseline hazard rate $h_0(t)$, the piecewise log-constant proportional hazard model (Breslow, 1972). To construct this model we start from a finite partition of the time axis, $0=s_0<s_1<s_2<\cdots<s_K$ with $s_K>t_i$ for all

Table 2. *Model selection for the AIDS clinical trial data*

Model	$W_1(s)$	$W_2(s)$	Mean of the deviance	Effective number of parameters	DIC
No random effects					
I	0	0	9812.1	11.6	9823.7
II	0	U_3	9812.1	12.2	9824.4
Random intercepts					
III	U_1	0	7507.8	432.2	7940.0
IV	U_1	U_3	7507.7	432.1	7939.9
V	U_1	$\gamma_1 U_1$	7438.1	433.3	7871.4
VI	U_1	$\gamma_1 U_1 + U_3$	7439.9	430.8	7870.7
Random intercepts and random slopes					
VII	$U_1 + U_2 s$	0	7109.0	734.6	7843.6
VIII	$U_1 + U_2 s$	$\gamma_1 U_1$	7056.9	736.7	7793.6
IX	$U_1 + U_2 s$	$\gamma_2 U_2$	7053.6	757.4	7811.0
XI	$U_1 + U_2 s$	$\gamma_1 U_1 + \gamma_2 U_2$	6979.3	760.1	7739.4

DIC, deviance information criteria.

Table 3. *Separate and joint analysis for the AIDS clinical trial data*

	Separate analysis		Joint analysis	
Parameter	Posterior mean	95% CI	Posterior mean	95% CI
Longitudinal sub-model				
Intercept (β_{11})	8.05	(7.36, 8.74)	8.05	(7.36, 8.74)
Time (β_{12})	−0.20	(−0.29, −0.10)	−0.27	(−0.36, −0.17)
Time × drug (β_{13})	0.05	(−0.08, 0.19)	0.03	(−0.11, 0.17)
Sex (β_{14})	−0.15	(−0.79, 0.49)	−0.11	(−0.75, 0.54)
PrevOI (β_{15})	−2.33	(−2.81, −1.86)	−2.35	(−2.82, −1.88)
Stratum (β_{16})	−0.10	(−0.57, 0.36)	−0.11	(−0.58, 0.36)
τ_ε	0.35	(0.31, 0.38)	0.35	(0.31, 0.38)
Σ_{11}^{-1}	0.06	(0.06, 0.07)	0.06	(0.06, 0.07)
Σ_{22}^{-1}	2.58	(2.21, 2.99)	2.56	(2.24, 2.92)
$\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$	−0.12	(−0.22, −0.01)	−0.06	(−0.15, 0.03)
Survival sub-model				
Intercept (β_{21})	−3.72	(−4.05, −3.41)	−4.07	(−4.49, −3.67)
Drug (β_{22})	0.21	(−0.08, 0.50)	0.26	(−0.09, 0.61)
Sex (β_{23})	−0.17	(−0.40, 0.08)	−0.13	(−0.41, 0.17)
PrevOI (β_{24})	0.62	(0.40, 0.85)	0.76	(0.51, 1.02)
Stratum (β_{25})	0.08	(−0.08, 0.24)	0.07	(−0.12, 0.27)
γ_1	–	–	−0.20	(−0.25, −0.14)
γ_2	–	–	−1.61	(−1.97, −1.23)

CI, Confidence interval.

$i = 1, \ldots, m$ observed lifetimes, and assume the baseline hazard to be constant in each time interval

$$h_0(t) = \lambda_k \quad \text{for } t \in (s_{k-1}, s_k], \quad k = 1, \ldots, K.$$

Let (t_i, δ_i), $i = 1, \ldots, m$ indicate the (possibly censored) survival times and censoring indicator, and let $\mathbf{z}_i$ indicate the set of covariates recorded for individual i. Then, the hazard rate for individual i in the kth time interval is:

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i + b_k) = \exp(\eta_{ik}), \quad t \in (s_{k-1}, s_k], \tag{12}$$

where $b_k = \log(\lambda_k)$. Assigning a Gaussian prior to the vector $(b_1, \ldots, b_K)$ and to the parameter vector $\boldsymbol{\beta}$ brings us back to a Gaussian distributed predictor η_{ik} and therefore to latent

Gaussian models. Note that extending the predictor η_{ik} in (12) to the general form in (1) does not constitute any problem.

The log-likelihood contribution for data point (t, δ) with $t \in (s_{k-1}, s_k]$ is:

$$l = \delta \log h(t) - \int_0^t h(u)\,\mathrm{d}u = \delta\eta_k - (t - s_k)\,\mathrm{e}^{\eta_k} - \sum_{j=1}^{k-1}(s_{j+1} - s_j)\,\mathrm{e}^{\eta_j} \tag{13}$$

and depends on the Gaussian latent field through the vector of predictors $\eta_1, \ldots, \eta_k$. Hence, INLA is not directly applicable to such model.

To be able to apply INLA we have to rewrite the model so that it fits the INLA framework. Notice that (13) is equivalent to the log-likelihood of k Poisson-distributed data points, of which $k-1$ with mean $(s_{j+1} - s_j)\,\mathrm{e}^{\eta_j}$ observed to be 0, and one with mean $(t - s_k)\,\mathrm{e}^{\eta_k}$ observed to be 0 or 1 according to whether the survival time t is observed or censored. The fact that a Cox model with piecewise log-constant baseline hazard is equivalent to certain Poisson regression model was noted independently by Holford (1980), and Laird & Oliver (1981). The key to apply INLA to a Cox model with piecewise log-constant baseline hazard lies therefore in data augmentation. In practice, each original data point (t, δ) with $t \in (s_{k-1}, s_k]$ is represented by k Poisson-distributed data points in the augmented data set. Such data augmentation brings us back to the latent Gaussian models described in section 2.

For $(b_1, \ldots, b_K)$ we choose a prior that gives smooth realizations. As the baseline hazard is by 'default' constant, the choice falls on an intrinsic first-order random walk (RW1) model (Rue & Held, 2005, ch. 3) with precision τ_b. RW1 models are built by first assuming that the location k of the nodes are all positive integers, that is, $k = 1, 2, \ldots, K$ so that the distance between nodes is constant and equal to 1. Then, increments $b_{k+1} - b_k$ are assumed i.i.d.

$$b_{k+1} - b_k \sim \mathcal{N}(0, \tau_b^{-1}), \quad k = 1, \ldots, K-1. \tag{14}$$

Such models are invariant to the addition of any constant to the overall mean. We assume Gamma prior for τ_b.

A useful extension of the semi-parametric model allows for multiple strata. The strata divide the subjects in $j = 1, \ldots, J$ disjoint groups, each of which has a distinct baseline hazard function but common effects of the covariates. Such models are used when the stratum covariate does not respect the proportional hazard assumption.

The hazard of an individual i belonging to stratum j is:

$$h_{ij}(t) = h_0^j(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i). \tag{15}$$

As before, the baseline hazards $h_0^j(\cdot)$ are assumed to be piecewise constant in each time interval $(s_{k-1}, s_k]$, $k = 1, \ldots, K$:

$$h_0^j(t) = \lambda_k^j = \exp(b_k^j) \quad \text{for } t \in (s_{k-1}, s_k], \ k = 1, \ldots, K, \ j = 1, \ldots, J.$$

We assume RW1 priors with common precision for $(b_1^j, \ldots, b_K^j)$, $j = 1, \ldots, J$:

$$b_{k+1} - b_k \sim \mathcal{N}(0, \tau_b^{-1}), \quad k = 1, \ldots, K-1, \ j = 1, \ldots, J.$$

Such stratified model can be easily implemented in INLA library as shown in the next example.

5.1. Example: leukaemia survival data in North-West England

To illustrate the use of INLA for piecewise log-constant Cox models we consider the data set presented in Henderson *et al.* (2002) and re-proposed in Kneib & Fahrmeir (2007). Both analyses concentrate on the detection of spatial variations but while the first paper retains

the assumption of linear predictor for covariate effects, the second one assumes more flexible smooth effects of covariates.

The data set contains information on $m = 1043$ cases of leukaemia in adults diagnosed between 1982 and 1998 in Northwest England. Almost 16 per cent of cases are right censored. For each patient i the following covariates are recorded: the age of the patient (age_i), the white blood cell counts (wbc_i) at diagnosis, the Townsend deprivation index (tpi_i), which measures the deprivation for the enumeration district of residence, the sex of the patient (sex_i) codified as a dummy variable (1 = female, 0 = male) and district of residence (s_i). We partition the time axis into $K = 20$ equally spaced intervals and assume a Cox model with piecewise log-constant baseline hazard. From the results in Kneib & Fahrmeir (2007), we let age and white blood cell counts have a linear effect while for the tpi index we assume a smooth effect. Moreover, a spatial effect is included. The predictor for patient i at time $t \in (s_{k-1}, s_k]$ is then given by:

$$\eta_{ik} = \beta_0 + \beta_{\text{sex}}\text{sex}_i + \beta_{\text{age}}\text{age}_i + \beta_{\text{wbc}}\text{wbc}_i + f^{(\text{tpi})}(\text{tpi}_i) + f^{(s)}(s_i) + b_k.$$

The tpi values are rounded to $n_{\text{tpi}} = 50$ different values and their effect is modelled as a smooth function $f^{(\text{tpi})}(\cdot)$, parameterized as unknown values $\boldsymbol{f}^{(\text{tpi})} = \{f_1^{(\text{tpi})}, \ldots, f_{n_{\text{tpi}}}^{(\text{tpi})}\}$. The vector $\boldsymbol{f}^{(\text{tpi})}$ is assumed to follow a second-order random walk (Rue & Held, 2005, ch. 3) defined as:

$$\pi(\boldsymbol{f}^{(\text{tpi})} \mid \tau_{\text{tpi}}) \propto \tau_f^{(n_{\text{tpi}}-2)/2} \exp\left\{-\frac{1}{2}\tau_{\text{tpi}} \sum_{i=3}^{n_{\text{tpi}}} (f_i^{(\text{tpi})} - 2f_{i-1}^{(\text{tpi})} + f_{i-2}^{(\text{tpi})})^2\right\}.$$

The model for the spatial term $\boldsymbol{f}^{(s)} = \{f_1^{(s)}, \ldots, f_{n_s}^{(s)}\}$, with $n_s = 24$ being the number of districts, is defined conditionally as:

$$f_1^{(s)} \mid \boldsymbol{f}_{-i}^{(s)}, \tau_s \sim \mathcal{N}\left(\frac{1}{n_i^s}\sum_{j \in \partial_i} f_j^s, \frac{1}{n_i^s \tau_s}\right),$$

where ∂_i is the set of neighbour district to district i, namely those n_i^s district which share a common border with i; see Rue & Held (2005, sec. 3.3.2), for further details on this model. For identifiability reasons, we assume sum-to-zero constraints on both the smooth effect of tpi and the district effect. The model is completed by assigning an $\mathcal{N}(0, 10^4\mathbf{I})$ prior to $\boldsymbol{\beta} = \{\beta_0, \beta_{\text{sex}}, \beta_{\text{age}}, \beta_{\text{wbc}}\}$ and independent $\Gamma(1, 0.001)$ priors to the three hyperparameters $\boldsymbol{\theta} = (\tau_b, \tau_{\text{tpi}}, \tau_s)$.

The computing time needed by INLA is around 10 seconds. Estimates for the log-baseline and for the smooth effect of tpi are shown in Fig. 2. The log-baseline decreases over nearly the whole observed period. The increase at the end of the observation time should not be over-interpreted as there are only 26 individual surviving 10 years. The effect of the deprivation index is first increasing and then staying almost constant after reaching a value of about 0.

The estimated spatial effect is shown in Fig. 3(A). Areas with low risk are concentrated in the west part of the country while areas with high risk are more spread. Such path can be seen clearly from the significance map in Fig. 3(B) where white denotes districts with strictly negative 80 per cent credible intervals and black denotes districts with strictly positive 80 per cent credible intervals. These findings correspond to those reported in Kneib & Fahrmier (2007).

To check how the estimate of the log-baseline hazard $h_0(t)$ varies with the number of intervals K, we have repeated the analysis using $K = 50$ and 100. The results are shown in Fig. 4. Increasing the number of intervals K we get a more detailed estimate of the baseline hazard function.

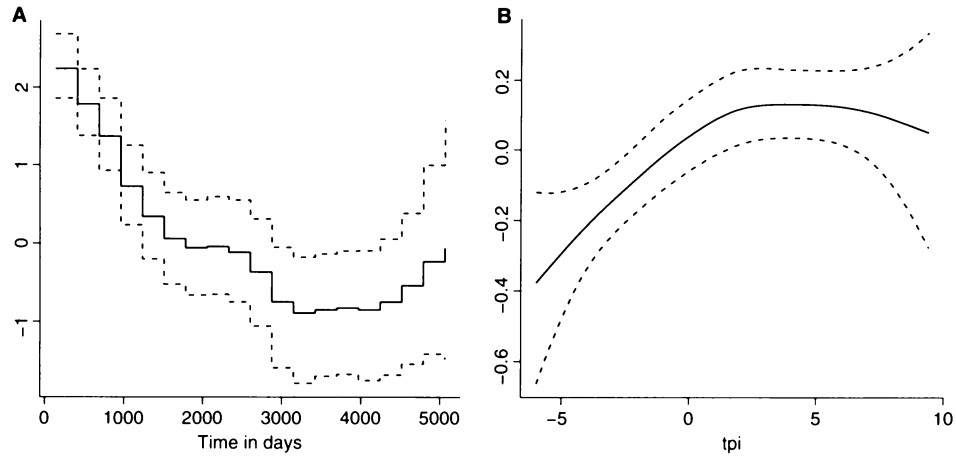

Fig. 2. Leukaemia survival data: posterior means (solid line) and 95 per cent credible intervals (dashed lines) for the log-baseline hazard (A) and the effect of Townsend deprivation index (B).

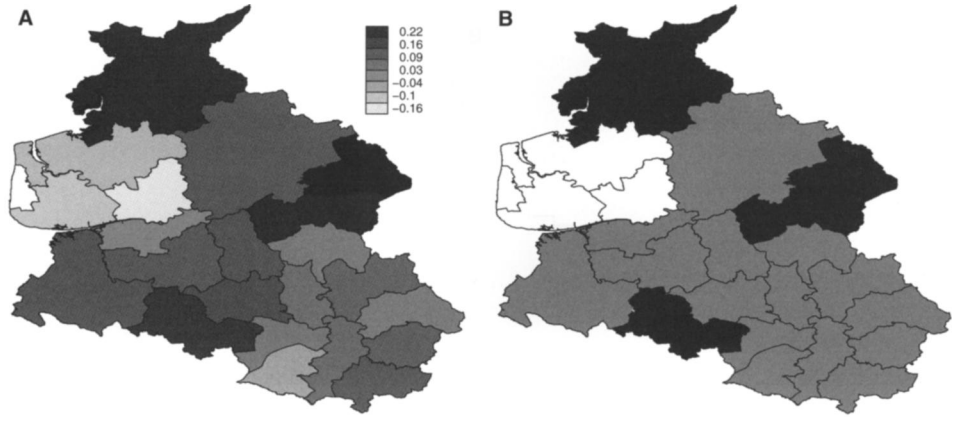

Fig. 3. Leukaemia survival data. (A) Spatial effect on a district level (posterior mean). (B) Point-wise 80 per cent significance map. White denotes districts with strictly negative credible intervals whereas black denotes districts with strictly positive credible intervals.

To check the proportional hazard assumption for men and women, we fit a stratified model where sex is the stratum variable. The estimated log-baseline hazards for women and men are displayed in Fig. 5.

Although the two curves differ, they do so at the end of the observation time where there are only few surviving individuals. Moreover, the confidence bands are wide enough to include both curves. We compare the simple and stratified models using the DIC values reported in Table 4 concluding that the stratified model is not necessary and that the proportional hazard assumption is valid for the sub-populations of men and women. The same conclusion is reached by Kneib & Fahrmier (2007).

6. Interval-censored data

In many applications, the analyst is confronted with more complex censoring schemes than right censoring. Interval-censored survival times T are not observed exactly, but are only

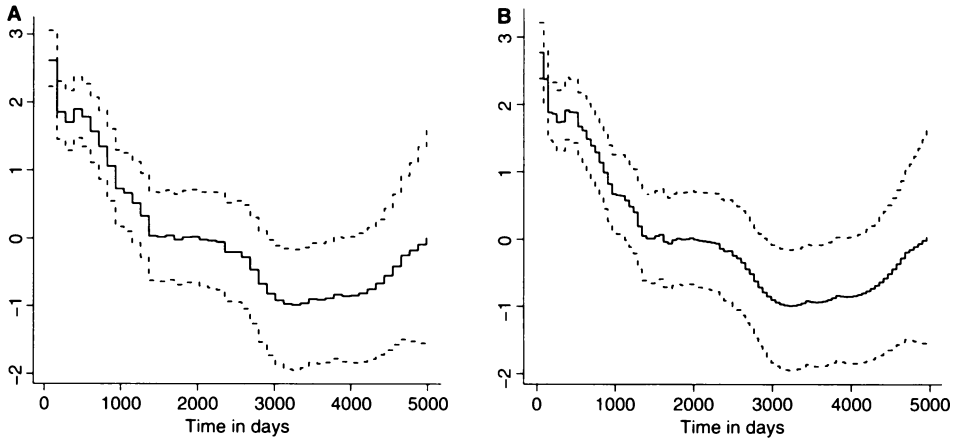

Fig. 4. Leukaemia survival data: posterior means (solid line) and 95 per cent credible intervals (dashed lines) for the log-baseline hazard with $K = 50$ (A) and $K = 100$ (B).

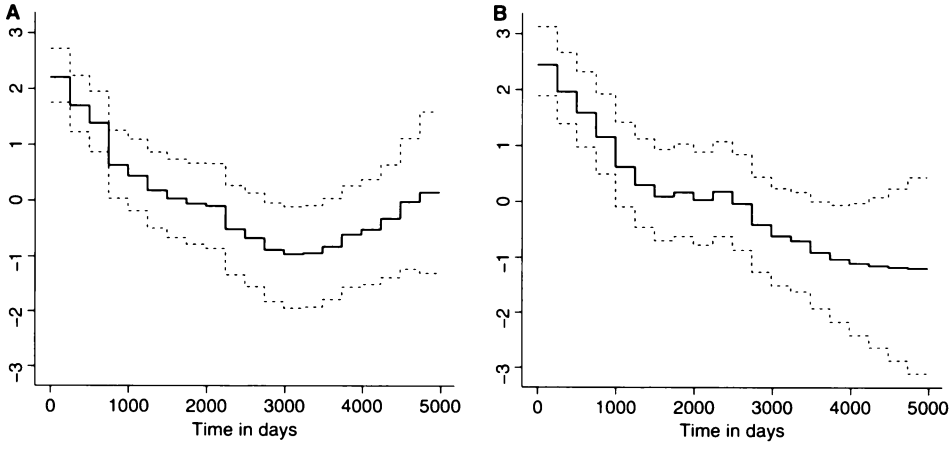

Fig. 5. Leukaemia survival data, stratified model: posterior means (solid line) and 95 per cent credible intervals (dashed lines) for the log-baseline hazard for women (A) and for men (B).

Table 4. *Model selection for the leukaemia example*

Model	Mean of the deviance	Effective number of parameters	DIC
Simple model	5326.60	28.22	5354.82
Stratified model	5325.83	33.52	5359.35

DIC, deviance information criteria.

known to fall into an interval $[T_{lo}, T_{up}]$. If $T_{lo} = 0$, such survival times are referred to as left censored.

Another feature of lifetime data often encountered is that of truncation. While censoring is about leaving the study, truncation is about entering it. We say that an observation is left truncated if the survival time is observed only if it exceeds the truncation time T_{tr}.

In a general framework an observation can be described by a quadruple $(T_{lo}, T_{up}, T_{tr}, \delta)$ with:

$T_{lo} = T_{up}, \delta = 1$ if the observation is uncensored

$T_{lo} = T_{up}, \delta = 0$ if the observation is right censored

$T_{lo} < T_{up}, \delta = 0$ if the observation is interval censored.

Moreover, for left truncated observations we have $T_{tr} > 0$ while $T_{tr} = 0$ indicates that the observation is not truncated. The general log-likelihood contribution for an observation represented by $(T_{lo}, T_{up}, T_{tr}, \delta)$ is given by:

$$l = \delta \log(h(T_{up})) - \int_{T_{tr}}^{T_{up}} h(u)\,\mathrm{d}u + \log\left\{1 - \exp\left(-\int_{T_{lo}}^{T_{up}} h(u)\,\mathrm{d}u\right)\right\}. \tag{16}$$

For the Weibull model discussed in section 3 the general log-likelihood term in (16), for a data point with predictor η as in (1) reduces to:

$$l = \delta \log\{\alpha T_{up}^{\alpha-1} \exp(\eta)\} - \exp(\eta)(T_{up}^{\alpha} - T_{tr}^{\alpha}) + \log\{1 - \exp(-T_{up}^{\alpha} + T_{lo}^{\alpha})\}. \tag{17}$$

The likelihood in (17) depends on the latent field $\mathbf{x}$ only through the predictor η, just like the log-likelihood for right-censored data in (6). Applying INLA to parametric Weibull models is therefore straightforward also for interval-censored data.

As for the piecewise-constant model, the rightmost element in (16), characteristics for interval-censored data, does not allow to use the data augmentation trick discussed in section 5; therefore, INLA cannot be applied in such case.

7. Discussion

In this article we have shown that many models for survival analysis can be considered as latent Gaussian models, which allow us to do the inference using INLA. We have demonstrated that INLA provides a computational solution which gives the end-user a smooth experience, both in terms of speed and simplicity.

The website http://www.r-inla.org contains all the data and R scripts to perform the analysis reported in the article including the INLA software itself.

References

Banerjee, S. & Carlin, B. P. (2003). Semiparametric spatio-temporal frailty modelling. *Environmetrics* **14**, 523–535.

Banerjee, S., Wall, M. M. & Carlin, B. P. (2003). Frailty modeling for spatially correlated survival data with application to infant mortality in Minnesota. *Biostatistics* **4**, 123–142.

Breslow, N. (1972). Discussion on Regression models and life-tables (by D. R. Cox). *J. Roy. Statist. Soc. Ser. B* **34**, 216–217.

Brezger, A., Kneib, T. & Lang, S. (2003). *BayesX: software for Bayesian inference*, Version 1.1 edn. Department of statistics, University of Munich. Available on http://www.stat.uni-muenchen.de/~bayesx/bayesx.html (accessed October 2009).

Carlin, B. P. & Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian statistics, 7* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, M. West), 45–63. Oxford University Press, New York.

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

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *J. Roy. Statist. Soc. Ser. C* **50**, 201–220.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*, 2nd edn. Springer-Verlag, Berlin.

Guo, X. & Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using stadard computer packages. *Amer. Statist.* **58**, 1–9.
Hannerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *JASA* **101**, 1065–1075.
Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.
Henderson, R., Shimakura, S. & Gorst, D. (2002). Modeling spatial variation in leukemia survival data. *J. Amer. Statist. Assoc.* **97**, 965–972.
Holford, T. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* **36**, 299–305.
Ibrahim, J., Chen, M. & Sinha, D. (2001). *Bayesian survival analysis*. Springer, New York.
Kneib, T. (2006). Geoadditive hazard regression for interval censored survival times. *Comput. Statist. Data Anal.* **51**, 777–792.
Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scand. J. Statist.* **34**, 207–228.
Laird, N. & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76**, 231–240.
Lang, S. & Brezger, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13**, 183–212.
Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts structure and extensibility. *Statist. Comput.* **10**, 325–337.
McGilchrist, C. A. & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* **47**, 461–466.
Rue, H. & Held, L. (2005). *Gaussian Markov random fields: theory and applications, Vol. 104 of Monographs on statistics and applied probability*. Chapman & Hall, London.
Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. Roy. Statist. Soc. Ser. B* **71**, 319–392.
Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. R. (1995). *BUGS: Bayesian inference using Gibbs sampling*, version 0.50. MRC Biostatistics Unit, Cambridge.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. Ser. B* **64**, 583–639.

Received February 2010, in final form August 2010

Sara Martino, Department of Mathematical Sciences, NTNU, 7491 Trondheim, Norway.
E-mail: martino@math.ntnu.no