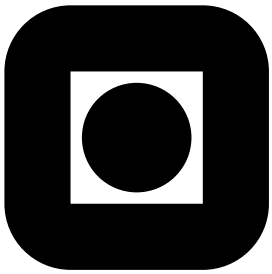
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Approximate Bayesian Inference for Survival Models

by

Sara Martino Rupali Akerkar and Håvard Rue

PREPRINT
STATISTICS NO. 3/2010



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL http://www.math.ntnu.no/preprint/statistics/2010/S3-2010.pdf
Sara Martino has homepage: http://www.math.ntnu.no/∼martino
E-mail: martino@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Approximate Bayesian Inference for Survival Models

Sara Martino, Rupali Akerkar and Håvard Rue
Department of Mathematical Sciences
NTNU, Norway
Email: {martino,akerkar,hrue}@math.ntnu.no

February 16, 2010

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 paper, we show how a new inferential tool named Integrated Nested Laplace approximations (INLA) can be adapted and applied to many survival models making Bayesian analysis both fast and accurate without having to rely on MCMC based inference.

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 post-estimate 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 due to improvements in both modeling techniques and computational power. Several extensions to the basic Cox model have been proposed in the Bayesian literature in order 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 and Banerjee (2003) and Banerjee and Carlin (2003) do similarly with a semi-parametric estimation of baseline hazard. Hannerfeind et al. (2006) extend the work of Fahrmeir and Lang (2001) and Lang and 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 modeled via geostatistical and conditional autoregressive priors while B-splines are used to model the unknown smooth functions. Inference is done using MCMC algorithms. Kneib (2006) extends the geoadditive

Cox model to deal with interval censored survival times. Kneib and Fahrmeir (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 Henderson 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 and 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 diagose 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 paper 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 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 a trivial changes in the Gaussian part of the model. Moreover, Bayesian inference for latent Gaussian models can be done using *Integrated Nested Laplace approximations* (INLA), see Rue et al. (2009). INLA provide fast and accurate approximations to the posterior marginals through a clever use of Laplace approximations and advanced numerical methods taking computational advantage 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. The second aim the this paper is to show how INLA can be adapted to fit survival models. Although the implementation of INLA is quite involved, an open-source version written in `C` based on the `GMRFLib`-library (Rue and Held, 2005) is available. An interface from `R` named `INLA`, is also available; see `www.r-inla.org` for documentation and worked through examples.

The rest of the paper is organized as follows: in Section 2 we introduce latent Gaussian models and give a short description of the INLA approach to approximate the 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 $\boldsymbol{x}$ to be point-wise observed through $n_d \leq n$ data $\boldsymbol{y}$. The latent field $\boldsymbol{x}$ is assumed to have Gaussian density conditionally on some hyperparameters $\boldsymbol{\theta}_1$,

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

The data $\boldsymbol{y}$ are assumed to be conditionally independent given the latent field $\boldsymbol{x}$ and, possibly, some additional hyperparameters $\boldsymbol{\theta}_2$ in the likelihood. The model definition is completed the prior density for the hyperparameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$. In addition, some linear constraints of the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$, where the $k \times n$ matrix $\boldsymbol{A}$ has rank $k \ll n$, may be imposted.

Latent Gaussian models are a subset of all Bayesian structured additive models (see Fahrmeir and 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:

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

Here, the $\{\beta_k\}$'s represent the linear effect of covariates $\boldsymbol{z}$. The $\{f^{(j)}(\cdot)\}$'s are unknown functions of the covariates $\boldsymbol{u}$: 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 $\{f^{(j)}\}$'s functions. The w_{ij} are known weights defined for each observed data point. Finally, ϵ_i's are unstructured random effects. A latent Gaussian model is obtained by assigning $\boldsymbol{x} = \{\{f^{(j)}(\cdot)\}, \{\beta_k\}, \{\eta_i\}\}$ a Gaussian prior with precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_1)$.

The posterior distribution then reads:

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

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 infere the posterior marginal distributions for the latent field $\pi(x_i|\boldsymbol{y})$, $i = 1, \ldots, n$ and for the hyperparameters $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Integrated Nested Laplace approximation (INLA) provides a recipe for computing in a fast and accurate way, such posterior marginals (Rue et al., 2009). The approximations $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ and $\tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, $i = 1, \ldots, n$ are based on a clever use of Laplace approximations. Posterior marginals for the latent variables $\tilde{\pi}(x_i|\boldsymbol{y})$ are then computed via numerical integration:

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

Posterior marginals for the hyperparameters $\tilde{\pi}(\theta_j|\boldsymbol{y})$, $j = 1, \ldots, M$ can also be derived via 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.

In order 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 paper. First, the latent field $\boldsymbol{x}$, 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 $\boldsymbol{Q}$, (Rue and Held, 2005). The efficiency of INLA relies, in fact, on algorithms for sparse matrices computations. Almost all latent Gaussian models in the literature satisfy this conditions. The second condition to be satisfied is that the dimension of the hyperparameter vector $\boldsymbol{\theta}$ should be not be too large. This is necessary for the integral in Eq. (3) to be computationally feasible. Finally, each data point y_i should depend on the latent field $\boldsymbol{x}$ only through the predictor η_i, i.e. $\pi(y_i|\boldsymbol{x}, \boldsymbol{\theta}_1) = \pi(y_i|\eta_i, \boldsymbol{\theta}_1)$. This is a technical requirement due to 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:

$$h(t_i) = h_0(t_i) \exp(\eta_i) \tag{4}$$

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

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

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

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

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

The extended Weibull regression model described above can easily be seen as a latent Gaussian model with latent field $\boldsymbol{x}$, 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 $\boldsymbol{x}$ only through the predictor η_i, as can be seen from Eq. (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 and Aisbett, 1991; Spiegelhalter et al., 1995). The data set contains, for each patient, the first and second infection time t_{ij} and a set of three covariates: sex, age and the type of disease. Spiegelhalter et al. (1995) propose a Weibull model with to analyze 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:

$$h(t_{ij}) = \alpha t_{ij}^{\alpha-1} \exp(\eta_{ij}); \qquad i = 1, \ldots, 38 \qquad j = 1, 2$$

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 $\boldsymbol{\beta} = \{\beta_{\text{sex}}, \beta_{\text{age}}, \beta_{\text{dis2}}, \beta_{\text{dis3}}, \beta_{\text{dis4}}\} \sim \mathcal{N}(0, \boldsymbol{I})$. 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:

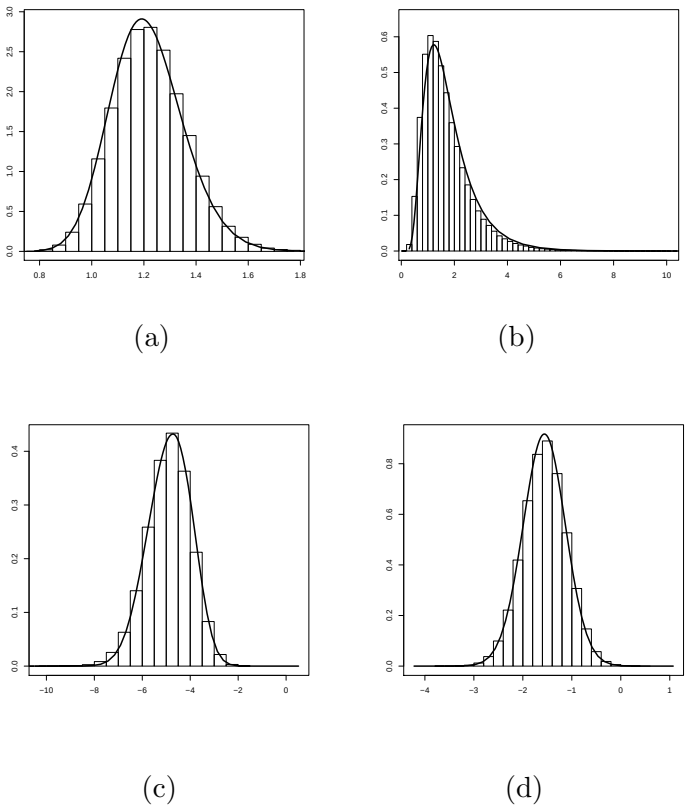
```
formula = inla.surv(time,event) ~ age + sex + dis2 + dis3 + dis4 +
          f(ID, model="iid", param=c(1, 1))
```

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 Figure 1 the INLA posterior marginals for α, τ, β_0 and β_{sex} are compared to histograms based on long MCMC runs using WinBUGS. All examples in the paper are implemented on on a dual-core 2.5GHz laptop and the execution times refers to such machine. The estimates are practically indistinguishable despite the fact that the computing time for `inla()` was only 2 seconds while WinBUGS needed around 5 minutes. Results for the other elements of the $\boldsymbol{\beta}$ vector are similar.



(a) (b)

(c) (d)

Figure 1: Kidney example: Posterior marginal distributions approximated by INLA (solid line) and MCMC based density estimates (histogram) for α (a), τ (b), β_0 (c) and β_{sex} (d)

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 $\boldsymbol{x}$, or a large enough ratio between the number of data points and the total (or effective) number of model parameters. In this example there are relatively few data compared to the total number of parameters (76 observation and a total of 46 parameters) so choosing flat priors makes the problem difficult for INLA. To illustrate, lets use the priors chosen in the WinBUGS manual: $\boldsymbol{\beta} \sim \mathcal{N}(0, 10^4\boldsymbol{I})$, $\tau \sim \Gamma(10^{-3}, 10^{-3})$ 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 $\boldsymbol{x}|\boldsymbol{\theta}$ 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.

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

Table 1: The Kidney Infection example: Comparison between INLA and MCMC based estimates in the case of very flat prior densities.

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 2 seconds when the the MCMC sampler needed around 10 minutes.

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 set 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 and 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 $\boldsymbol{z}$ 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 $\mu_i(s) = \boldsymbol{z}_{1i}^T(s)\boldsymbol{\beta}_1$ and $W_{1i}(s) = \boldsymbol{d}_{1i}(s)\boldsymbol{U}_i$ incorporates subject specific random effects. The vectors $\boldsymbol{z}_{1i}^T(s)$ and $\boldsymbol{\beta}_1$ represent (possibly time varying) explanatory variables and their correspond-

ing regression coefficients. The $\boldsymbol{U}_i$ are vectors of Gaussian random effects corresponding to the explanatory variables $\boldsymbol{d}_{1i}(s)$.

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

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

where the vectors $\boldsymbol{z}_{2i}(t)$ 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 $\boldsymbol{z}_{1i}$ 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 $(W_{1i}, W_{2i})^T$. 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 Eq. (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 $(U_{1i}, U_{2i})^T$ are assumed to have a bivariate Gaussian distribution $\mathcal{N}(0, \boldsymbol{Q}_U^{-1})$ while the U_{3i} are independent frailty terms with $\mathcal{N}(0, \tau_{U_3}^{-1})$ prior and independent of the $(U_{1i}, U_{2i})^T$'s. Vague Gaussian priors are assigned to $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, γ_1, γ_2 and γ_3 in Eq. (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 $\boldsymbol{Q}_U$.

This rather complex model reduces to a latent Gaussian field. Define the vector of hyperparameters to be $\boldsymbol{\theta} = \{\alpha, \tau_1, \tau_{U_3}, \gamma_1, \gamma_2, \gamma_3, \boldsymbol{Q}_U\}$. The latent field $\boldsymbol{x} = (\{\eta_{ij}^l\}, \{\eta_i^s\}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \{(U_{1i}, U_{2i})\}, \{U_{3i}\})$, conditioned on $\boldsymbol{\theta}$, has Gaussian distribution with precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$. Finally the observations $(\{y_{ij}\}, \{t_i\})$ have a likelihood which, conditional on the hyperparameters $\boldsymbol{\theta}$, depends on the latent field $\boldsymbol{x}$ 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 Henderson et al. (2000) referring to a clinical study to compare the efficacy of two antiretroviral drugs, didanosine (ddI) and zelcitabine (ddC) in treating patients who have failed or were intolerant of zidovudine (AZT) therapy. For each of the $m = 467$ patient 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 per ml of blood y_{ij} is recorded at time s_{ij}. There is a maximum of 5 observations per patient. Four explanatory variable 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 and 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 Deviance Information Criterion (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:

$$\text{DIC} = \overline{\text{D}} + p_D$$

where $\overline{\text{D}}$ 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 and Carlin (2004), with the exception of model X and XII which, for technical reasons, cannot be implemented in the current version of `INLA`. The computing time goes from 6 seconds needed for model I to 206 for model XI. The `R` code used to fit the models is available at `www.r-inla.org`. In Table 2 the computed $\overline{\text{D}}$, p_D and DIC are reported for the 9 different joint

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

Table 2: Model selection for the ddI/ddC data

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% credible intervals are reported in Table 3. Our results appear to be equal to those obtained via Gibbs sampling in Guo and Carlin (2004) up to two digit accuracy.

5 Semi-parametric baseline hazard models

In this section we consider a semiparametric 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 $i = 1, \ldots, m$ observed lifetimes, and assume the baseline hazard to be constant in each time interval

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

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

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

Parameter	Separate Analysis Posterior Mean	95%CI	Joint Analysis Posterior Mean	95%CI
	Longitudinal Sub-model		*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)
$\tau_{epsilon}$	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		*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)

Table 3: Separate and Joint analysis for the ddI/ddC data

where $b_k = \log(\lambda_k)$. Assigning a Gaussian prior to the vector $(b_1, \ldots, b_K)$ and to the parameters 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 Eq. (12) to the general form in Eq. (1) does not constitute any problem.

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

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

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.

In order to be able to apply INLA we have to rewrite the model so that it fits the INLA framework. Notice that Eq. (13) is equivalent to the log-likelihood of k Poisson distributed data points, of which $k-1$ with mean $(s_{j+1} - s_j)e^{\eta_j}$ observed to be 0, and one with mean $(t - s_k)e^{\eta_k}$ 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 and 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. Since the baseline hazard is by "default" constant, the choice falls on an intrinsic first-order random walk (RW1) model (Rue and

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, i.e. $k = 1, 2, \ldots, K$ so that the distance between nodes is constant and equal to 1. Then, increments $b_{k+1} - b_k$ are assumed independent and identically distributed

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

Such models are invariant to the addition of any constant to the overall mean. We assume $\tau_b \sim \Gamma(a, b)$.

5.1 Example: Leukemia 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 and Fahrmeir (2007). Both analysis concentrate on the detection of spatial variations but while the first paper retain 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 leukemia in adults diagnosed between 1982 and 1998 in Northwest England. Almost 16% 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), 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 and Fahrmeir (2007), we let age and white blood cell counts have a linear effect while for the Townsend deprivation 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 modeled as a smooth function $f^{(\text{tpi})}(\cdot)$, parametrized 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 and Held (2005), Ch. 3) defined as:

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

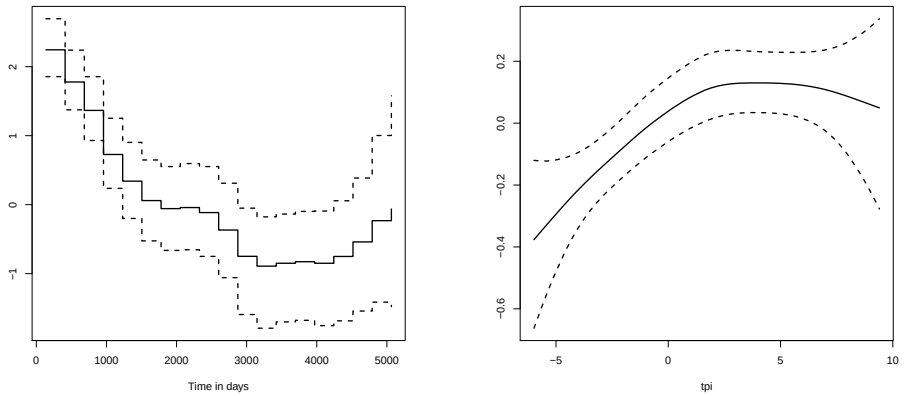
The model for the spatial term $\boldsymbol{f}^{(s)} = \{f_1^{(s)}, \ldots, f_{n_s}^{(s)}\}$, with $n_s = 24$ being the number of districts, is defined conditionally as:

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

where ∂_i is the set of neighbor district to district i, namely those n_i^s district which share a common border with i; see Rue and Held (2005), section 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 a $\mathcal{N}(0, 10^4\boldsymbol{I})$ prior to $\boldsymbol{\beta} = \{\beta_0, \beta_{\text{sex}}, \beta_{\text{age}}, \beta_{wbc}\}$ and independent $\Gamma(1, 0.001)$ priors to the three hyperparameters $\boldsymbol{\theta} = (\tau_b, \tau_{tpi}, \tau_s)$.

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 Figure 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 since 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 Figure 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

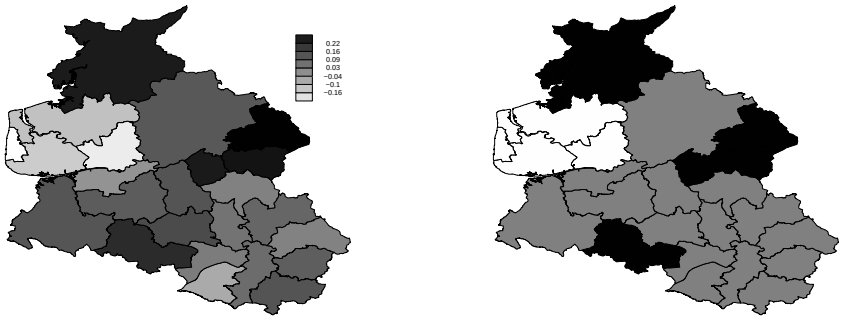


(a) log(baseline) (b) effect of tpi

Figure 2: Leukemia survival data: posterior means (solid line) and 95% credible intervals (dashed lines) for the log-baseline hazard and the effect of Townsend deprivation index.

.

the significance map in Figure 3(b) where white denotes districts with strictly negative 80% credible intervals and black denotes districts with strictly positive 80% credible intervals. These findings



(a) (b)

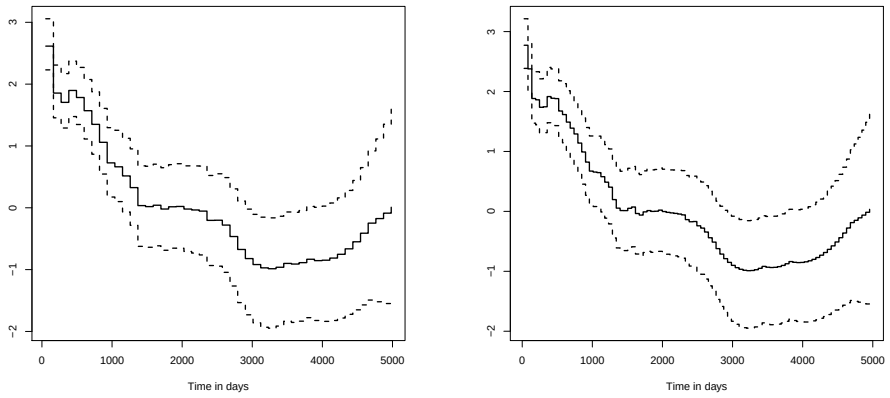
Figure 3: Leukemia survival data. Left panel: spatial effect on a district level (posterior mean). Right panel: point-wise 80% significance map. White denotes districts with strictly negative credible intervals, black denotes districts with strictly positive credible intervals.

correspond to those reported in Kneib and Fahrmeir (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 $K = 100$. The results are shown in Figure 4. Increasing the number of intervals K we get a more detailed estimate of the baseline hazard function.

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 known to fall



(a) (b)

Figure 4: Leukemia survival data: posterior means (solid line) and 95% credible intervals (dashed lines) for the log-baseline hazard with $K = 50$ (Panel (a)) and $K = 100$ (Panel(b)).

.

into an interval $[T_{\text{lo}}, T_{\text{up}}]$. If $T_{\text{lo}} = 0$ 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 quatruple $(T_{\text{lo}}, T_{\text{up}}, T_{\text{tr}}, \delta)$ with:

$T_{\text{lo}} = T_{\text{up}} \quad \delta = 1 \quad$ if the observation is uncensored
$T_{\text{lo}} = T_{\text{up}} \quad \delta = 0 \quad$ if the observation is right censored
$T_{\text{lo}} < T_{\text{up}} \quad \delta = 0 \quad$ if the observation is interval censored

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

$$l = \delta \ \log(h(T_{\text{up}})) - \int_{T_{\text{tr}}}^{T_{\text{up}}} h(u)du + \log\left\{1 - \exp\left(-\int_{T_{\text{lo}}}^{T_{\text{up}}} h(u)du\right)\right\} \tag{15}$$

For the Weibull model discussed in Section 3 the general log-likelihood term in Eq. (15), for a data point with predictor η as in Eq. (1) reduces to:

$$l = \delta \ \log\left\{\alpha T_{\text{up}}^{\alpha-1} \exp(\eta)\right\} - \exp(\eta)\left(T_{\text{up}}^{\alpha} - T_{\text{tr}}^{\alpha}\right) + \log\left\{1 - \exp(-T_{\text{up}}^{\alpha} + T_{\text{lo}}^{\alpha})\right\} \tag{16}$$

The likelihood in Eq. (16) depends on the latent field $\boldsymbol{x}$ only through the predictor η, just like the log-likelihood for right censored data in Eq. (6). Applying INLA to parametric Weibull models is therefore straight forward also for interval censored data.

As for the piecewise-constant model , the rightmost element in Eq. (15), 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 paper 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 provide a computational solution which gives the end-user a smooth experience, both in terms of speed and simplicity.

The website `www.r-inla.org` contains all the data and `R`-scripts to perform the analyses reported in the paper including the `INLA`-software itself.

References

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

Banerjee, S., Wall, M. M., and 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). *Journal of the Royal Statistical Society, Series B*, 34:216–217.

Brezger, A., Kneib, T., and Lang, S. (2003). *BayesX: Software for Bayesian inference*. Department of statistics, University of Munich, version 1.1 edition. http://www.stat.uni-muenchen.de/∼lang/bayesx.

Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics, 7*, pages 45–63. Oxford Univ. Press, New York.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C*, 50(2):201–220.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, Berlin, 2nd edition.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using stadard computer packages. *The American Statisticial*, 58(1).

Hannerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *JASA*, 101(475):1065–1075.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Henderson, R., Shimakura, S., and Gorst, D. (2002). Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, 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., and D.Sinha (2001). *Bayesian Survival Analysis*. Springer, New York.

Kneib, T. (2006). Geoadditive hazard regression for interval censored survival times. *Computational Statistics and Data Analysis*, 51:777–792.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34:207–228.

Laird, N. and Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76:231–240.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1).

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing,*, 10:325–337.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, 47:461–466.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(2):583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). BUGS: Bayesian inference using Gibbs sampling. Version 0.50, MRC Biostatistics Unit, Cambridge.