survBayes: A introduction into the package

Volkmar Henschel, Christiane Heiß, Ulrich Mansmann
University of Munich
Department of Medical Informatics, Biometry and Epidemiology
Marchioninistr. 15, 81377 Munich, Germany

June 9, 2005

Abstract

This software fits a multivariate proportional hazards model to interval censored event data by a Bayesian approach. Right and interval censored data and a lognormal or gamma frailty term can be fitted. An example is studied and the output analysed.

1 The basic model

The data, based on a sample of size n, consists of the triple $(t_i, \delta_i, \mathbf{x}_i), i = 1, \ldots, n$ where t_i is the time on study for subject i, δ_i is the event indicator for subject i ($\delta_i = 1$ if event has occurred, $\delta_i = 0$ if the observation is right censored), $\mathbf{x}_i$ is the r-dimensional vector of covariate values for subject i.
The likelihood contribution of the i-th single observation is given by

$$\lambda_0(t_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i) = \exp\left\{\delta_i[h(t_i) + \beta'\mathbf{x}] - \mathrm{e}^{\beta'\mathbf{x}} \int_0^{t_i} \exp[h(s)]\mathrm{d}s\right\}$$

where $h(s) = \ln[\lambda_0(s)]$. The infinite dimensional problem gets to a finite dimensional one by partitioning the time axis $[0, \infty[$ into disjoint intervals $I_k = [\theta_{k-1}, \theta_k[$ for $k = 1, ..., K+1$ where θ_k is the time of the k-th event and $\theta_0 = 0$. The largest event time observed is θ_K and I_{K+1} is taken as the interval $[\theta_K, \infty[$. The function h is constant on the intervals I_k and is set to $-\infty$ on $[\theta_K, \infty[$. The integral in the likelihood contribution of the i-th observation can be written as a sum.
The priors for the components of the vector β will be independently normal distributed with mean 0 and a small precision $\tau = 0.001$. The prior for step function h will be a autoregressive process of order one with prior information on smoothness. Writing $h_k = h(\theta_k)$, $k = 1, \ldots, K$ the first order process is defined as $h_k = h_{k-1} + \epsilon_k$ with $\epsilon_k \sim N(0, \sigma_k^2)$ and $h_0 \sim N(0, \sigma_0^2)$, where h_0 and ϵ_k, $k = 1, \ldots, K$ are pairwise independent. The variances are chosen as $\sigma_k^2 = \Delta_k \sigma_1^2$ and Δ_k may be defined by $\theta_k - \theta_{k-1}$ for $k > 1$ with $\theta_0 = 0$. The inverse of the covariance matrix, Σ^{-1}, is a bandmatrix of bandwidth one. The parameters $\frac{1}{\sigma_0^2} = \tau_0$ and $\frac{1}{\sigma_1^2} = \tau_1$ are treated as hyperparameters with flat gamma priors setting both parameters equal to 0.001.

2 Sampling procedure

Sampling for the parameter vector

Aitkin and Clayton [1] pointed out that the proportional hazards model can interpreted as a generalized linear model.
Gamerman [2] describes how one can effectively sample the vector of covariates in generalizes linear mixed models in a block updating step. This is a combination of the iterated least squares method (IWLS) as it is known in fitting such models with a Metropolis-Hasting sampling.

Sampling for the baseline hazard

With the given structure of the log baseline hazard function one has to sample from a Gaussian Markov Random Field (GMRF), see Rue [5].

Sampling for the dispersion parameters

For the dispersion parameters σ_0^2 and σ_1^2 a flat Gamma prior with rate κ and shape ν is chosen. This leads to Gamma posteriors.

3 Extensions of the basic model

Data augmentation and a multiplicative frailty model is used to analyze clustered interval censored event data. Data augmentation is used to interfere unobserved event times. The potential clustering of event times within a statistical unit is modeled by introducing an unit specific random effect or frailty term into the proportional hazards model.

4 Example

Meisel et al. [4] present data on the shrinkage of aneurisms associated with cerebral arteriovenous malformations (cAVM) after embolization treatment. The time to a shrinkage of the aneurism to below 50% of the baseline volume was of interest. Several patients had multiple aneurisms. Each patient was inspected at a random inspection time $obs.t$. The censoring variable z was set to one, if at the inspection time sufficient shrinkage was observed, else the censoring indicator was set to zero.
Two covariates were considered: the degree of cAMV occlusion by embolization (dichotomized at 50%, variable mo) and the location of the aneurism, whether at the midline arteries or at other afferent cerebral arteries, variable lok.
The single aneurisms are not independent because aneurisms within a patient may shrink in the same way (because the share the same "environment"). Multiple aneurisms were observed per patient. This clustering of aneurisms is indicated by the grouping variable gr.
The data is loaded and inspected for the first five patients.

```
> library(survBayes)
```

```
Loading required package: survival
Loading required package: splines
Loading required package: coda
```

```
> data(AA.data)
> AA.data[1:11, ]
```

```
   z mo gr lok    t.left  t.right
1  0  0  1   1 1.7698630       NA
2  0  1  2   1 0.9972603       NA
3  0  1  2   1 0.9972603       NA
4  0  1  2   1 0.9972603       NA
5  0  0  3   0 1.0712329       NA
6  0  0  3   1 1.0712329       NA
7  0  0  4   1 5.6547945       NA
8  0  0  5   1 1.5780822       NA
9  1  0  5   0 0.0000000 1.578082
10 1  0  5   0 0.0000000 1.578082
11 1  0  5   1 0.0000000 1.578082
```

The data is analyzed by applying the `survBayes` algorithm. The fit with `survBayes` gives an object which stores all sampled values in the required number after the burn in. The str function gives a survey over the output. The low number for the sample is only due to fast checking of the package in the CRAN. Please choose at least 5000.

```
> AA.res <- survBayes(Surv(t.left, t.right, z * 3, type = "interval") ~
+     mo + lok + frailty(gr, dist = "gauss"), data = AA.data, burn.in = 0,
+     number.sample = 10)
> str(AA.res)
```

```
List of 7
 $ t.where        : num [1:49] 0.0000 0.0281 0.0611 0.1305 0.1580 ...
 $ lbh            : mcmc [1:10, 1:49] -1.395 -1.328 -0.904 -0.795 -0.593 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:10] "lbh" "lbh" "lbh" "lbh" ...
  .. ..$ : NULL
  ..- attr(*, "mcpar")= num [1:3] 1 10 1
  ..- attr(*, "class")= chr "mcmc"
 $ beta           : mcmc [1:10, 1:2] -0.435 -0.647 -0.647 -1.144 -1.283 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:2] "mo" "lok"
  ..- attr(*, "mcpar")= num [1:3] 1 10 1
  ..- attr(*, "class")= chr "mcmc"
 $ sigma.lbh      : mcmc [1:10, 1:2] 1917.150    1.870  598.689    4.518    0.551 ...
  ..- attr(*, "mcpar")= num [1:3] 1 10 1
  ..- attr(*, "class")= chr "mcmc"
 $ alpha.cluster  : mcmc [1:10, 1:83]  0.000000 -0.001091  0.000736 -0.001478 -0.001493 ..
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:10] "alpha.cluster" "alpha.cluster" "alpha.cluster" "alpha.cluster" ...
```

```
  .. ..$ : NULL
  ..- attr(*, "mcpar")= num [1:3] 1 10 1
  ..- attr(*, "class")= chr "mcmc"
 $ sigma.cluster  :Class 'mcmc'  atomic [1:10] 2.00e-06 4.32e-06 7.68e-06 1.28e-05 1.22e-0
  .. ..- attr(*, "mcpar")= num [1:3] 1 10 1
 $ m.h.performance: num [1:3] 7 9 9
```

The components are

t.where: the time points which were chosen; the range of the Kaplan Meier estimate is divided by the number of grid points and transformed back to the time axis;

lbh: samples of the log baseline hazard at the grid points;

beta: samples of the vector of covariates;

sigma.lbh: samples of sigma.lbh.0 and sigma.lbh.1;

alpha.cluster: samples of the frailty values;

sigma.cluster: samples of frailty variance;

m.h.performance: number of the successful performances of the Metropolis-Hastings step for beta, lbh and, if appropriate, alpha

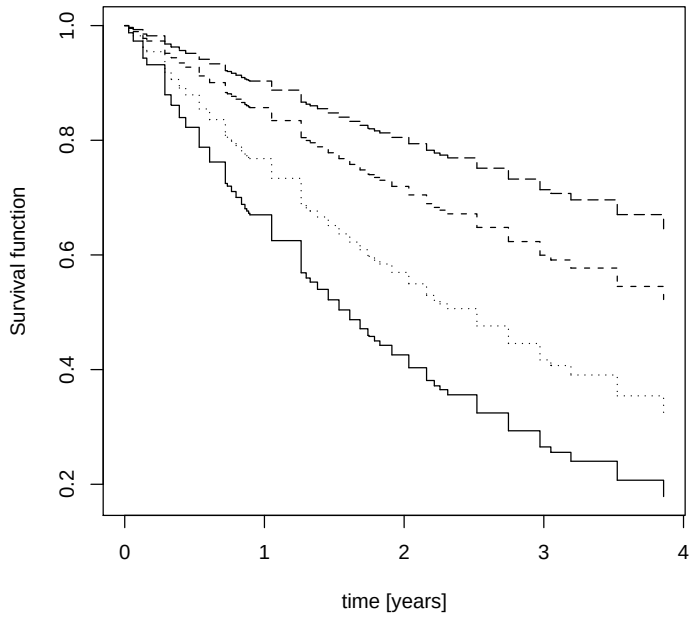
The convergence is diagnosed by mean of CODA. The Raftery-Lewis diagnostic gives a good description of the convergence, see [3].

```
> raftery.diag(AA.res$beta)

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

You need a sample size of at least 3746 with these values of q, r and s

> raftery.diag(AA.res$sigma.lbh)

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

You need a sample size of at least 3746 with these values of q, r and s

> raftery.diag(AA.res$sigma.cluster)

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

You need a sample size of at least 3746 with these values of q, r and s

> raftery.diag(AA.res$alpha.cluster)
```

```
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

You need a sample size of at least 3746 with these values of q, r and s
```

This indicates that the sample size should be increased to at least 30000 samples.
The estimated coefficients and cumulative baseline hazard can be used to estimated and plot group specific survival curves.

```
> beta.est <- apply(AA.res$beta, 2, mean)
> lambda0 <- exp(apply(AA.res$lbh, 2, mean))
> Lambda0 <- c(0, cumsum(diff(AA.res$t.where) * lambda0[-length(lambda0)]))
> surv.base <- exp(-Lambda0)
> plot(AA.res$t.where, surv.base, type = "s", xlab = "time [years]",
+     ylab = "Survival function", lty = 1)
> lines(AA.res$t.where, surv.base^exp(beta.est["mo"]), type = "s",
+     lty = 2)
> lines(AA.res$t.where, surv.base^exp(beta.est["lok"]), type = "s",
+     lty = 3)
> lines(AA.res$t.where, surv.base^exp(sum(beta.est[c("mo", "lok")])),
+     type = "s", lty = 5)
> leg.names <- c("mo=0, lok=0", "mo=1, lok=0", "mo=0, lok=1", "mo=1, lok=1")
> legend(4, 1, leg.names, lty = c(1, 2, 3, 5), bty = "n")
```



This work was supported by DFG grant MA 1723/2-1.

References

[1] M. Aitkin and D. Clayton. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, 29:156–163, 1980.

[2] D. Gamerman. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7:57–68, 1997.

[3] U. Mansmann. Convergence Diagnosis for Gibbs Sampling Output. *Medical Infobahn for Europe, A. Hasman et al. (Eds.)*, IOS Press, 83–87, 2000.

[4] H. J. Meisel, U. Mansmann, H. Alvarez, G. Rodesch, M. Brock, and P. Lasjaunias. Cerebral arteriovenous malformations and associated aneurysms: Analysis of 305 cases from a series of 662 patients. *Neurosurgery*, 46:793–802, 2000.

[5] H. Rue. Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society B*, 63:325–338, 2001.

[6] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: extending the Cox model*. Springer, New York, 2000.

[7] B. Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69:169–173, 1974.