www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

The BayHaz package for Bayesian estimation of smooth hazard rates in R

Luca La Rocca
www-dimat.unipv.it/luca

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Smooth hazard rate estimation

CPP and BPS priors

Prior elicitation

Posterior computation

Directions for future work

Smooth hazard rate estimation

CPP and BPS priors

Prior elicitation

Posterior computation

Directions for future work

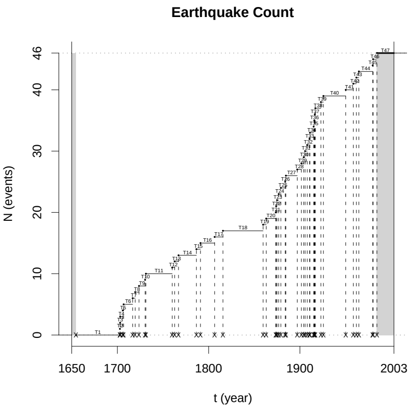
Suppose we observe either $\{T_i = t_i\}$ or $\{T_i > t_i\}$ for $i = 1, \ldots, n$, where

$$T_1, \ldots, T_n | \rho \overset{i.i.d.}{\sim} \rho(t) \exp\left\{-\int_0^t \rho(s) ds\right\} dt$$

are survival times with unknown (non-defective) hazard rate ρ, that is,

$$\rho \geq 0, \qquad \exists t > 0 : \int_0^t \rho(s) ds < \infty, \qquad \int_0^\infty \rho(s) ds = \infty.$$

We want to learn the shape of ρ from data (non-parametric approach) but we know that ρ is smooth.



Events with moment magnitude greater than 5.1 in a very active Italian seismogenic zone...

...the inter-event times can be considered exchangeable; they are available as a data set of `BayHaz` [La Rocca, 2007]:

```
library(BayHaz)
data(earthquakes)
```

Smooth hazard rate estimation

CPP and BPS priors

Prior elicitation

Posterior computation

Directions for future work

A compound Poisson process (CPP) prior hazard rate [La Rocca, 2008] is defined by

$$\rho(t) \quad = \quad \xi_0 k_0(t) + \sum_{j=1}^{\infty} \xi_j k(t - \sigma_j), \qquad t \geq 0,$$

where $\sigma_j, j \geq 1$, are the jump-times of a CPP process with gamma distributed jump-sizes $\xi_j, j \geq 1$, while k is a zero-mean Gaussian density (kernel), ξ_0 is an independent random variable with the same distribution as any jump-time ξ_j, and k_0 is a suitable function such that the mean of $\rho(t)$ does not depend on t.

A first-order autoregressive Bayesian penalized spline (BPS) prior hazard rate, based on [Hennerfeind *et al.*, 2006], is defined by

$$\rho(t) = \exp\left\{\sum_{j=1}^{G+k-2} \eta_j B_j(t)\right\}, \quad 0 \le t \le T_\infty,$$

where η is a normal first order autoregressive stationary process, while $B_j(t)$ is the j-th B-spline basis function of order k, evaluated at t, defined on a grid of $G + 2k - 2$ equispaced knots with first internal knot at 0 and last internal knot at T_∞ (time-horizon of interest); there are G internal nodes, and B-spline basis functions sum to one within them.

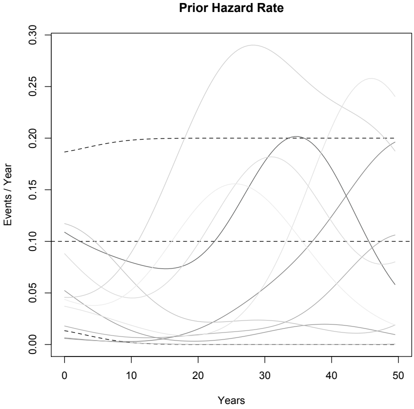
For CPP priors, a time-scale equivariant elicitation procedure is available to assign a constant prior expected hazard rate while controlling prior variability, based on the following quantities:

- `r0` prior mean hazard rate (r_0);
- `H` corresponding (asymptotic) coefficient of variation;
- `T00` time-horizon of interest (T_∞);
- `M00` number of extremes within the time-horizon in a "typical" hazard rate trajectory (M_∞).

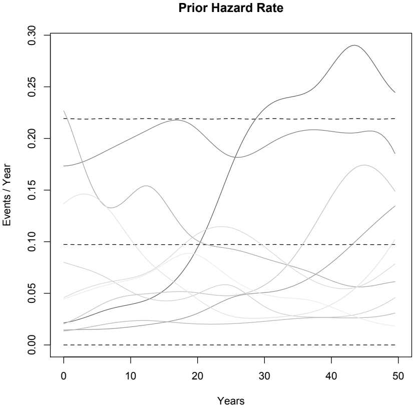
There is a technical issue (disregarded in these slides) concerning the number of CPP jumps needed to cover the time-horizon of interest.

A procedure to find a matching BPS prior is also available.

Prior elicitation




```
hypCPP <- CPPpriorElicit(r0 = 0.1, H = 1,
                         T00 = 50, M00 = 2)
priorCPP <- CPPpriorSample(ss = 10,
                           hyp = hypCPP)
CPPplotHR(priorCPP, tu = "Year")
```




```
hypBPS <- BPSpriorElicit(r0 = 0.1, H = 1,
                         T00 = 50, G = 9)
priorBPS <- BPSpriorSample(ss = 10,
                           hyp = hypBPS)
BPSplotHR(priorBPS, tu = "Year")
```

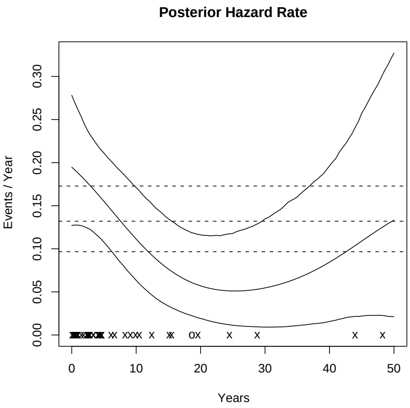
Markov chain Monte Carlo (MCMC) posterior approximation:

- Gibbs-type sampler for CPP posteriors, introducing a latent label per exact observation ⇒ hazard-driven probabilistic clustering;
- tailored proposal density Metropolis-Hastings sampler for BPS posteriors;

```
# CPP posterior sample (about three quarters of an hour on my MacBook2,1)
postCPP <- CPPpostSample(hypCPP, times = earthquakes$ti, obs = earthquakes$ob,
                         mclen = 10000, burnin = 50000, thin = 20, lab = FALSE)
# BPS posterior sample (about one fourth of the time on the same machine)
postBPS <- BPSpostSample(hypBPS, times = earthquakes$ti, obs = earthquakes$ob,
                         mclen = 10000)
```

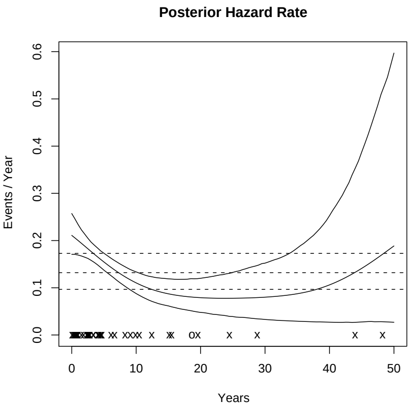
Interface to package `coda` [Plummer et al., 2007] for output diagnostics:

```
MCMCpostCPP <- CPPpost2mcmc(postCPP) # package 'coda' is automatically loaded
MCMCpostBPS <- BPSpost2mcmc(postBPS) # and an MCMC object is created
```



Pointwise posterior mean and equal tail 95% credible band: solid lines refer to the CPP posterior; dashed lines refer to the posterior obtained by means of a constant hazard rate model (using a conjugate gamma prior and letting its shape and rate parameters tend to zero). Exact observations are marked with "x", whereas censored observations are marked with "o".

```
CPPplotHR(postCPP, tu = "Year")
```



Pointwise posterior mean and equal tail 95% credible band: solid lines refer to the BPS posterior; dashed lines refer to the posterior obtained by means of a constant hazard rate model (using a conjugate gamma prior and letting its shape and rate parameters tend to zero). Exact observations are marked with "x", whereas censored observations are marked with "o".

```
BPSplotHR(postBPS, tu = "Year")
```

Smooth hazard rate estimation

CPP and BPS priors

Prior elicitation

Posterior computation

Directions for future work

Interesting directions for future work include:

- dealing with semiparametric models, e.g., using CPP priors at least for the single binary covariate proportional hazards model [LaRocca, 2004];
- implementing other prior hazard rates;
- revising `R` code and documentation, possibling using `C` code for posterior sampling.

Needless to say, suggestions are welcome... thank you!

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006)
Geoadditive survival models.
Journal of the American Statistical Association **101**, 1065–1075.

La Rocca, L. (2008)
Bayesian Non-parametric Estimation of Smooth Hazard Rates for Seismic Hazard Assessment.
Scandinavian Journal of Statistics, Online Early.

La Rocca, L. (2007)
BayHaz: R Functions for Bayesian Hazard Rate Estimation.
R package version 0.1-3.
`http://www-dimat.unipv.it/luca/bayhaz.htm`

La Rocca, L. (2004)
On Bayesian Analysis of the Proportional Hazards Model.
In: *Proceedings of the XLII Meeting of the Italian Statistical Society*, Università degli Studi di Bari, Italy, June 9–11, 2004.
CLEUP, Padua, Italy.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2007)
coda: Output analysis and diagnostics for MCMC.
R package version 0.13-1.

R Development Core Team (2008)
R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria.
`http://www.R-project.org`