

Discrete time Markov gamma processes and time dependent covariates in survival analysis

Luis E. Nieto-Barajas
ITAM, Department of Statistics
Río Hondo 1
Mexico D.F. 01000, Mexico
lnieto@itam.mx

1. Introduction

The proportional hazards model of Cox (1972) is widely used to deal with covariate information in survival analysis. This model assumes that hazard rates between two individuals are proportional, having a common hazard rate multiplied by a factor which depends on the covariates. To state the model, let T_i be a nonnegative random variable that represents the time to the event of interest (failure time) of individual i , and $Z_i(t) = (Z_{i1}(t), \dots, Z_{ip}(t))$ the vector of p time varying covariates. Then, the hazard rate function of individual i is taken to be

$$(1) \quad h_i(t) = h_0(t) \exp\{Z_i(t)'\theta\},$$

where $h_0(t)$ is a baseline hazard rate function and $\theta' = (\theta_1, \dots, \theta_p)$ is a vector of regression coefficients. In this case, the survival function for individual i becomes

$$S_i(t) = \exp\left\{-\int_0^t h_i(s)ds\right\}.$$

Cox (1972) proposed this model with the baseline hazard function $h_0(\cdot)$ being unspecified. Therefore this model has a semi-parametric nature. In a classical perspective, statisticians have concentrated in estimating the regression coefficients θ , maximising a partial likelihood (Cox, 1995), with none or little interest in the baseline hazard rate function. In a Bayesian approach, on the other hand, interest has focused on both, the baseline hazard $h_0(t)$ and the regression coefficients θ . Kalbfleisch (1978), for instance, modelled the baseline cumulative hazard function with a Gamma process prior. Laud et al. (1998) used a beta process prior to model the cumulative hazard function. More recently, Mezzetti and Ibrahim (2000) used the Markov gamma process prior of Nieto-Barajas and Walker (2002) to model the baseline hazard rate and called it correlated gamma process.

However, none of the Bayesian semi-parametric approaches of the proportional hazards model have explicitly considered the case when the covariates vary with time. The objective of this paper is to study the case of time varying covariates using the Markov gamma process prior of Nieto-Barajas and Walker (2002).

In Section 2, the construction of the Markov gamma process prior is reviewed. Section 3 presents the semi-parametric model and posterior distributions are derived. Finally, in Section 4 a sensitivity analysis is carried out using a simulated data set.

2. Markov gamma process prior

Let $0 = \tau_0 < \tau_1 < \tau_2 < \dots$ be a partition of the time axis into intervals, and h_k the constant hazard rate in the interval $(\tau_{k-1}, \tau_k]$, that is,

$$h(t) = \sum_{k=1}^{\infty} h_k I_{(\tau_{k-1}, \tau_k]}(t).$$

Then, the Markov gamma process prior is defined by a first order Markov process in $\{h_k\}$ through the use of a latent process $\{u_k\}$ in the following way (Nieto-Barajas and Walker 2002):

$$h_1 \sim \text{Ga}(\alpha_1, \beta_1), \quad u_k | h_k \sim \text{Po}(c_k h_k) \quad \& \quad h_{k+1} | u_k \sim \text{Ga}(\alpha_{k+1} + u_k, \beta_{k+1} + c_k),$$

for $k = 2, 3, \dots$. If $\alpha_k = \alpha_1$ and $\beta_k = \beta_1$ for all k then the process $\{h_k\}$ is stationary and

$$\text{Corr}(h_k, h_{k+1}) = c_k / (\beta_1 + c_k).$$

Although the process $\{h_k\}$ is defined in discrete time, the Markov gamma process prior $h(t)$ is defined in continuous time and assigns probability one to the set of continuous distribution functions.

3. Semi-parametric model

Differing from most of the previous Bayesian analysis of the proportional hazards model, which model the baseline cumulative hazard function with a stochastic process, in this paper we model the baseline hazard rate with a stochastic process. Considering the hazard rate (1) and using the Markov gamma process prior described in the previous section to model the baseline hazard rate $h_0(t)$, the cumulative hazard function for individual i becomes

$$H_i(t) = \sum_{k=1}^{\infty} h_k W_{i,k}(t, \theta),$$

where,

$$W_{i,k}(t, \theta) = \begin{cases} \int_{\tau_{k-1}}^{\tau_k} e^{Z_i(s)' \theta} ds & \text{if } t > \tau_k \\ \int_{\tau_{k-1}}^t e^{Z_i(s)' \theta} ds & \text{if } t \in (\tau_{k-1}, \tau_k] \\ 0 & \text{otherwise.} \end{cases}$$

Given a sample of possible right-censored observations T_1, \dots, T_n , where T_1, \dots, T_{n_u} are uncensored and T_{n_u+1}, \dots, T_n are right-censored, the conditional posterior distributions for the parameters of the semi-parametric model are:

- $f(h_k | \text{data}, u, \theta) = \text{Ga}(h_k | \alpha_k + u_{k-1} + u_k + n_k, \beta_k + c_{k-1} + c_k + m_k(\theta))$,
where, $n_k = \sum_{i=1}^{n_u} I(\tau_{k-1} < t_i \leq \tau_k)$ and $m_k(\theta) = \sum_{i=1}^n W_{i,k}(t_i, \theta)$
- $f(u_k | \text{data}, h, \theta) \propto \{c_k(c_k + \beta_{k+1})h_k h_{k+1}\}^{u_k} / \{\Gamma(u_k + 1)\Gamma(\alpha_{k+1} + u_k)\}$,
for $u_k = 0, 1, \dots$
- $f(\theta | \text{data}, h, u) \propto f(\theta) \exp\{\sum_{i=1}^{n_u} Z_i(t_i)' \theta - \sum_{k=1}^{\infty} h_k m_k(\theta)\}$.

Furthermore, if we want to introduce more flexibility within the prior process $\{h_k\}$, we can incorporate a hyper prior process for the $\{c_k\}$ such that $c_k \stackrel{IID}{\sim} \text{Ga}(1, \xi_k)$. The set of full conditional posterior distributions can then be extended to include

- $f(c_k | \text{data}, h, u, \theta) \propto (\beta_{k+1} + c_k)^{\alpha_{k+1} + u_k} c_k^{u_k} \exp\{-(\lambda_{k+1} + \lambda_k + \xi_k)c_k\}$,
for $c_k > 0$.

Posterior inference can be obtained by implementing a Gibbs sampling scheme. Simulating from the conditional posterior distributions of h_k and u_k is straightforward. If the prior distribution for θ is log-concave in each argument, simulation from the conditional posterior distributions of θ and c_k can be achieved using the adaptive rejection sampling (Gilks and Wild, 1992).

4. Sensitivity analysis

In this section we illustrate the results with a full Bayesian analysis and undertake a sensitivity analysis for a simulated data set. This simulated data come from a model with a single time-varying covariate with the following specifications: Let $Z_i(t) = z_i \log(t)$ be the covariate for individual i , and $h_0(t) = \lambda t$ the baseline hazard rate. It is straightforward to show that, in this case, the failure time $T_i \sim \text{We}(\theta z_i + 2, \lambda/(\theta z_i + 2))$.

We simulated a sample of size $n = 100$ with $\lambda = 1$, $\theta = 2$, $z_i = i/50$, for $i = 1, \dots, 100$. To specify the prior for $h(t)$, we took $\alpha_k = 0.001$ and $\beta_k = 0.01$ for all k to have relative noninformative conditions (see Nieto-Barajas and Walker, 2002), and took a range of values $c_k = 0, 1, 5, 10, 20, 50$ to assess the sensitivity for the posterior estimates of the parameter θ . In addition, we chose a prior distribution for $\theta \sim \text{No}(\mu_0, \sigma_0^2)$, with $\mu_0 = 0$ and a large value for the variance $\sigma_0^2 = 9$. The Gibbs sampling was run for 50,000 iterations with a burn-in of 5,000.

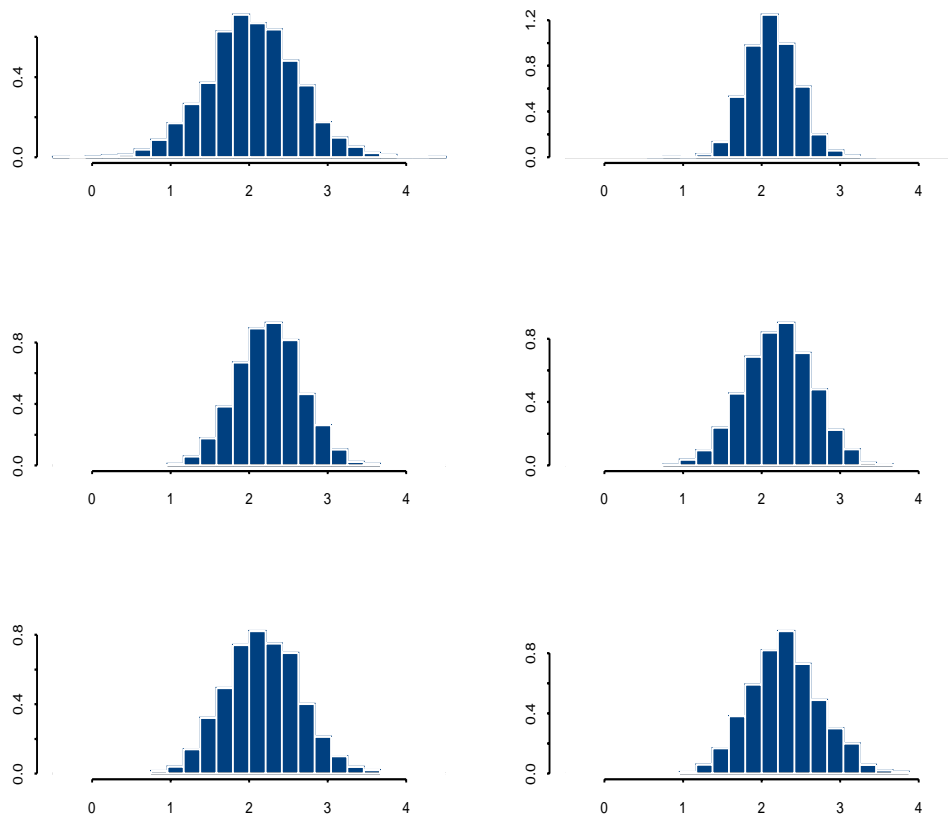


Figure 1. Posterior distributions of θ . From top to bottom and from left to right: $c_k = 0$, $c_k = 1$, $c_k = 5$, $c_k = 10$, $c_k = 20$, $c_k = 50$.

Figure 1 contains a histogram of the posterior distribution of θ for the different c_k 's. From there we can see that all distributions have roughly the same location, but with different variances. This behaviour is also summarised in the following table:

Posterior summaries of θ for different values of c_k

c_k	$E(\theta t)$	$\sqrt{\text{Var}(\theta t)}$	95% HDI
0	2.05	0.57	(1.00, 3.16)
1	2.13	0.32	(1.54, 2.76)
5	2.26	0.41	(1.45, 3.06)
10	2.21	0.43	(1.36, 3.06)
20	2.17	0.47	(1.30, 3.10)
50	2.30	0.45	(1.44, 3.21)

From the previous table we can see that the posterior mean of θ does not change dramatically for the different values of c_k and not far away from $\theta = 2$. The posterior standard deviation is also steady and the 95% high density intervals all clearly contain the true value of θ . Therefore, it can be said that the point estimates for θ are not so sensitive to the choice of the parameter c_k . Regarding the baseline hazard rate, a final comment is that, for larger values of c_k , the posterior estimates are closer to the true one.

REFERENCES

- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-202.
- Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.
- Gilks, W.R. and Wild, P. (1992). Adaptive Rejection sampling for Gibbs sampling. *Applied Statistics - Journal of the Royal Statistical Society, Series C* **41**, 337-348.
- Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214-221.
- Laud, P.W., Damien, P. and Smith, A.F.M. (1998). Bayesian nonparametric and covariate analysis of failure time data. In *Practical nonparametric and semiparametric Bayesian statistics*. D. Dey, P. Müller and D. Sinha (Eds). Springer. New York.
- Mezzetti, M. and Ibrahim, J.G. (2000). Bayesian inference for the Cox model using correlated gamma process priors. *Technical report, Department of Biostatistics, Harvard School of Public Health*.
- Nieto-Barajas, L.E. and Walker, S.G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413-424.

SUMMARY

In this paper we extend the use of the discrete time Markov gamma process of Nieto-Barajas and Walker (2002), to cope with time varying covariates in the proportional hazards model. We carry out a sensitivity analysis for the semi-parametric model using a simulated data set.

RÉSUMÉ

Cet article discute l'extension du modèle Markov-Gamma temps discret de Nieto-Barajas et Walker (2002) aux covariates variables de temps dans le modèle proportionnel de risque. Nous effectuons une analyse de sensibilité pour le modèle semi-paramétrique en utilisant des données simulées.