Journal of the Korean Statistical Society (2005), **34**: *3, pp 235–244*

A SIMULATION STUDY OF BAYESIAN PROPORTIONAL HAZARDS MODELS WITH THE BETA PROCESS PRIOR †

JAEYONG LEE[1]

ABSTRACT

In recent years, theoretical properties of Bayesian nonparametric survival models have been studied and the conclusion is that although there are pathological cases the popular prior processes have the desired asymptotic properties, namely, the posterior consistency and the Bernstein-von Mises theorem. In this study, through a simulation experiment, we study the finite sample properties of the Bayes estimator and compare it with the frequentist estimators. To our surprise, we conclude that in most situations except that the prior is highly concentrated at the true parameter value, the Bayes estimator performs worse than the frequentist estimators.

AMS 2000 subject classifications. Primary 62G20, 62N99; Secondary 62F15.
Keywords. Neutral to the right process; Beta process; Markov chain monte carlo.

1. INTRODUCTION

In recent years, asymptotic properties of Bayesian nonparametric survival models have been studied. In particular, with the neutral to the right (NTR) process prior on the unknown distribution, the sufficient conditions of the posterior consistency (Kim and Lee 2001)and Bernstein-von Mises theorem were obtained (Kim and Lee 2004).The essence of these studies is that not all the NTR process priors have the desired asymptotic properties, but all of the popular nonparametric priors, Dirichlet process, beta process and gamma process have the desired asymptotic properties, i.e. the posterior with these priors have compatible asymptotic properties with the frequentiest counter parts.

While the theoretical properties of Bayesian nonparametric survival model are studied, small sample performance of the Bayesian nonparametric survival model

Received April 2005; accepted July 2005.

†This research is supported by Seoul National University Grant 326-20030010.

[1] Department of Statistics, Seoul National University, Sillim-Dong Kwanak-Gu San 56-1, Seoul Korea (e-mail: leej@stats.snu.ac.kr)

has not been studied. In this study, we consider the proportional hazard model as the testbed. With the beta process (Hjort 1990),which is the most popular prior process in the survival model, on the baseline distribution, we study the small sample behaviour of the Bayes estimator and compare with that of the frequentist estimators in a simulation study. Especially, we focus on the behavior of the posterior of the regression coefficients. Similar study has been done in Kim and Ibrahim (2000),but their model is different from ours.

To our surprise, our conclusion of this simulation experiment is that although asymptotically the performance of the Bayes estimator is equivalent to that of the frequentist estimators, the story is quite different in small sample situations. In almost all situations except that the prior is highly concentrated at the true distribution, the Bayes estimator performs worse than the MLE.

In small sample problems, the common Bayesian folklore is that the Bayes estimator performs much better than the frequentist estimators. The rationale behind this folklore lies at the asymptotic approximation of the sampling distribution. Usually the frequentist inference relies on the asymptotic approximation of the sampling distribution, which is often very poor when the sample size is small. On the other hand, the Bayesian inference is exact whether the sample size is small or large. Because of this, it is commonly believed that, although the posterior and sampling distribution are different entities, the Bayesian procedure performs better than the frequentist procedures even with the frequentist criteria. However, to our surprise, in the proportional hazard model, our simulation study shows that the Bayes estimator performs worse than the frequentist estimators in almost all situations.

This paper is organized as follows. In section 2, we describe in detail the components of Bayesian nonparametric survival model: the model, prior, posterior and computation method. We report the simulation result in section 3 and discuss it.

2. Bayesian Proportional Hazard Model

2.1. Proportional Hazard Model

Let $X_1, \ldots, X_n$ be survival times with covariates $Z_1, \ldots, Z_n \in \mathcal{R}^p$. In the proportional hazard model we assume the distribution F_i of X_i with covariate Z_i is given by

$$1 - F_i(t) = (1 - F(t))^{\exp(\beta^T Z_i)},$$

where $\beta \in R^p$ is the unknown regression coefficient and F is the baseline distribution. The cumulative hazard function (CHF) A of the distribution F is defined as $dA(t) = dF(t)/(1 - F(t-))$. The CHF of X_i, A_i, is defined similarly,

$$dA_i(t) = 1 - (1 - dA(t))^{\exp(\beta^T Z_i)}. \tag{2.1}$$

If A is absolutely continuous with respect to the Lebesgue measure, there exists a hazard function a such that $A(t) = \int_0^t a(s)ds$ and the hazard function of F_i is given by $a_i(t) = a(t)e^{\beta^T Z_i}$.

The main feature of the survival data is that the survival times are subject to right censoring, and only (T_1, δ_1, Z_1), ..., (T_n, δ_n, Z_n) are observed, where $T_i = \min(C_i, X_i), \delta_i = I(X_i \leq C_i)$ and $C_1, \ldots, C_n$ are independent censoring variables. Based on the data, (T_1, δ_1, Z_1), ..., (T_n, δ_n, Z_n), we wish to make inference on β and F.

For the following sections, we introduce some notation. For $i = 1, 2, \cdots, n$, define counting processes $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(T_i \geq t)$. Let $N(t) = \sum_{i=1}^n N_i(t)$, $\Delta N(t) = N(t) - N(t-)$, and $Y(t) = \sum_{i=1}^n Y_i(t)$. Let q_n be the number of distinct uncensored observations and $t_1 < t_2 < \cdots < t_{q_n}$ the distinct ordered uncensored observations. Define

$$D_n(t) = \{i : T_i = t, \delta_i = 1, \quad i = 1, \ldots, n\},$$

$$R_n(t) = \{i : t \leq T_i, \quad i = 1, \ldots, n\},$$

and $R_n^+(t) = R_n(t) - D_n(t)$.

2.2. Prior and Posterior

In the proportional hazard model, we have two parameters, the CHF A and regression coefficient β. For the simulation study, we use the beta process with parameter $A_0(t)$ and $c(t)$, $BP(A_0, c)$, as the prior for A and $N(0, \sigma^2 I)$ for β.

The beta process $BP(A_0, c)$ is a nondecreasing independent increment (NII) process with Lévy measure

$$\nu(dx, dt) = \frac{c(t)}{x}(1 - x)^{c(t)-1} dx dA_0(t).$$

See Hjort (1990)for the original definition of the beta process and Kim and Lee (2001, 2003)for the detailed discussion of the NII process and its Lévy measure.

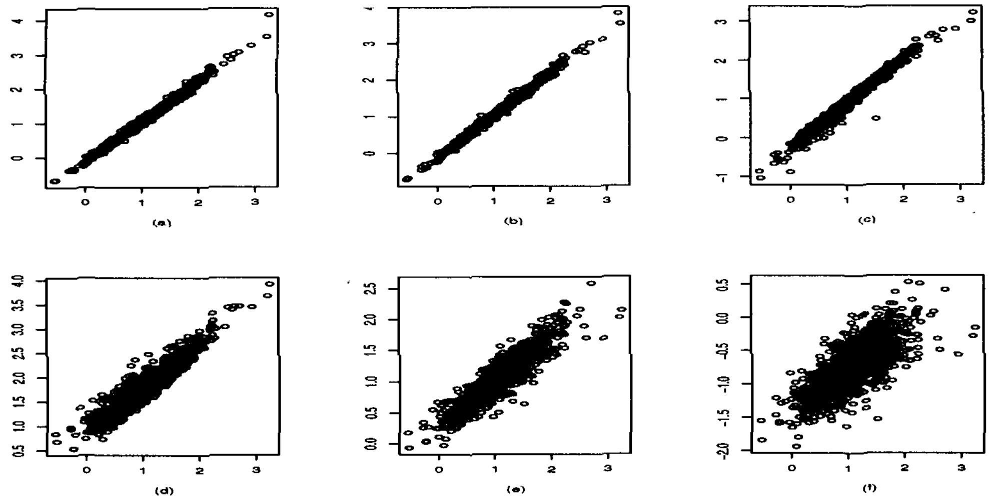

FIGURE 2.1 *The scatter plot of the MLE (x-axis) and the posterior mean (y-axis) with sample size* 30 *and (a)* $c = 0.1, a = 0.1$*, (b)* $c = 0.1, a = 1.0$ $c = 0.1, a = 10.0$ $c = 10.0, a = 0.1$ $c = 10.0, a = 1.0$ $c = 10.0, a = 10.0$.

If $A \sim BP(A_0, c)$, the expectation and variance of A are given as follows:

$$\begin{aligned} \mathrm{E}A(t) &= A_0(t) \\ \mathrm{Var}A(t) &= \int_0^t \frac{dA_0(s)}{c(s)+1}. \end{aligned}$$

Thus, A_0 is the center of the prior and c governs the variance of the the prior, *i.e.*, smaller the value of c larger the variance of A. In the simulation experiment, we will study the behavior of the posterior with different values of A_0 and c.

For the statistical analysis, we need to extract information from the posterior. Since the posterior of A and β is not of a simple form, to extract information from the posterior we need to sample from the posterior using Markov chain Monte Carlo (MCMC). An MCMC scheme for the proportional hazard model with beta process is described in Lee and Kim (2004).For the completeness of the paper, we describe the algorithm here. Note $A = A_d + A_c$ where A_d and A_c are stochastically discrete and continuous part of A, respectively and in the sampling scheme A_d and A_c are sampled separately.

- **Sampling A_c given β and data:**

  a. $M \sim Poisson(\lambda)$ with $\lambda = \frac{1}{\epsilon}\int_0^T c(s)dA_0(s)$.

b. For $i = 1, \cdots, M$, $r_i \propto c(r)dA_0(r)I(0 \leq r \leq \tau)$ and let $s_i = r_{(i)}$.

c. For $i = 1, \cdots, M$, $x_i \sim Beta(\epsilon, R_n(s_i, \beta) + c(s_i))$.

- **Sampling A_d given β and data:** For $i = 1, \cdots, q_n$, let $v_i = -\log(1 - u_i)$. Two auxiliary random variables, y_i and $w_i = (w_{i(1)}, \cdots, w_{i(k_i)})$ are introduced.

  a. $y_i \sim Geometric(1 - e^{-v_i})$.

  b. For $j = 1, \cdots, k_i$, $w_{ij} \propto \exp(-\exp(\beta^T z_{i(j)})v_i w_{ij})I(0 < w_{ij} < 1)$.

  c. $v_i \sim Gamma(k_i + 1, c(t_i) + R_n^+(t_i, \beta) + y_i + \sum_{j=1}^{k_i} w_{ij} e^{\beta^T z_{i(j)}})$.

  d. Set $u_i = 1 - e^{-v_i}$.

- **Sampling β given A and data:** We use the random-walk Metropolis-Hastings algorithm. Let β^* be a candidate value generated from random-walk kernel $q(\beta, \beta^*)$. Then, the acceptance ratio is

$$\frac{\pi(\beta^*)}{\pi(\beta)} \frac{\prod_{j=1}^{q_n} \left(\prod_{j=1}^{k_i} (1 - (1 - u_i)^{e^{\beta^{*T} z_{i(j)}}})(1 - u_i)^{R_n^+(t_i, \beta^*)} \right) \times \prod_{j=1}^{M} (1 - x_j)^{R_n(s_j, \beta^*)}}{\prod_{j=1}^{q_n} \left(\prod_{j=1}^{k_i} (1 - (1 - u_i)^{e^{\beta^T z_{i(j)}}})(1 - u_i)^{R_n^+(t_i, \beta)} \right) \times \prod_{j=1}^{M} (1 - x_j)^{R_n(s_j, \beta)}} \frac{q(\beta^*, \beta)}{q(\beta, \beta^*)}.$$

3. A SIMULATION STUDY

For the simulation experiment, we generated the sample from two baseline distributions: $F_1 = Exponential(1)$ and a discrete distribution F_2 which has support $\{0.3, 0.5, 1\}$ with probability 1/3 each. We consider only the one dimensional regression coefficient and its true value is fixed at 1. The covariate x is generated from $Bernoulli(0.5)$. The right censoring variable is generated from $Exponential(0.5)$. We generated 1000 set of samples with size $n = 10$ and 30. Because the sample sizes are small, there were non-negligible portion of samples that do not have unique MLE. We discarded these samples. This is usually due to constancy of the likelihood in some direction. For a set of sufficient conditions for the existence of the MLE, see Andersen et. al (1993).

We took $N(0, 100^2)$ for the prior for β. We chose sufficiently large variance for the prior of β so that in the simulation the effect of the prior of β is negligible.

For the prior for A, we put two different types of beta processes:

- BP1: $BP(A_0(dt) = adt, c(t) = c)$
- BP2: $BP(A_0(dt) = adt, c(t) = cexp(-at)$.

TABLE 3.1 *Frequentist Performance with sample size* 10 *and* $F_0 = Exponential(1)$.

	$n = 10$		Posterior Mean		Posterior Median		Mean Length	Coverage
Prior	c	a	bias	RMSE	bias	RMSE	of CI	Rate
BP1	.1	0.1	.07020	1.2505	.007372	1.1881	3.642	.856
		1.0	.0008926	1.2284	-.06191	1.1735	3.589	.860
		10.0	-.4500	1.2884	-.4777	1.2692	3.153	.804
	10.0	0.1	1.3807	1.5981	1.3511	1.5473	3.168	.541
		1.0	-.06002	.5940	-.03156	.5722	2.283	.957
		10.0	-2.2104	2.2814	-2.1609	2.2304	1.910	.018
BP2	.1	0.1	.04568	1.2489	-.01219	1.1922	3.630	.860
		1.0	-.09261	1.2787	-.1385	1.2441	3.466	.828
		10.0	-.7164	1.2094	-.7400	1.1883	1.208	.293
	10.0	0.1	1.3800	1.6006	1.3477	1.1922	3.183	.540
		1.0	-.05414	.6039	-.03067	.5798	2.353	.961
		10.0	-1.3942	1.5135	-1.3796	1.4982	1.006	.061
	MLE		-.1349	0.8195			3.104	.935

BP1 is used in Lee and Kim (2004) as an example and BP2 is used in Laud, Damien and Smith (1998).In fact, BP2 is the Dirichlet process at the parametric center *Exponential*(a). We considered prior combinations of $c = 0.1, 10$ and $a = 0.1, 1.0, 10.0$. Here $c = 0.1$ represents small prior variance or weak prior knowledge and $c = 10$ represents strong prior belief. The value a represents that the center of the prior processs is at *Exponential*(a). Thus, when the true baseline is F_1, $a = 1$ represents that the prior is centered at the truth. Even when the true baseline is F_2, the center of the prior is closest to the truth when $a = 1$.

The MCMC for the Bayesian analysis was run for 2200 iterations with thinning number 2 with the first 200 iterations were discarded as burnin. We wish we could have run for longer chain, but the simulation time is prohibitively long as one can imagine. In some cases we ran much longer chains, but the simulation results were similar to what we have here.

Tables 3.1 and 3.2 are the simulation results for the true baseline F_1 with sample sizes 10 and 30, respectively; Tables 3.3 and 3.4 are for the true baseline F_2.

In each simulation iteration, we computed two Bayes estimators: the posterior mean and median, and the MLEs. There are three different ways to deal with

TABLE 3.2 *Frequentist Performance with sample size* 30 *and* $F_0 = Exponential(1)$.

	$n = 30$		Posterior Mean		Posterior Median		Mean Length	Coverage
	c	a	bias	RMSE	bias	RMSE	of CI	Rate
BP1	.1	0.1	.1083	.5851	.09506	.5734	1.706	.870
		1.0	.09196	.5778	.07934	.5681	1.687	.864
		10.0	-.04145	.5712	-0.05090	.5620	1.620	.852
	10.0	0.1	.8141	.4701	.8053	.4609	1.647	.515
		1.0	.05104	.3805	.05377	.3773	1.341	.923
		10.0	-1.7453	.3628	-1.7336	.3617	1.039	.004
BP2	.1	0.1	.1137	.6287	.1000	.6137	1.705	.854
		1.0	.06120	.6255	.05336	.6230	1.623	.829
		10.0	-.5840	.8920	-.5884	.8886	.5586	.291
	10.0	0.1	.8023	.9411	.7923	.9271	1.655	.543
		1.0	.05624	.4075	.05680	.4043	1.387	.926
		10.0	-.9700	1.0334	-.9687	1.0330	.4287	.111
	MLE		.05736	0.5053			1.632	.911

ties in the data, Efron, Breslow and exact. For a detailed explanation of these estimator, see Klein and Moeschberger (2003).When there are no ties, these three estimators are exactly the same; thus we report only one as in Tables 3.1 and 3.2. For each estimator, we compute the bias and the root mean square error (RMSE). Also we computed 90% Bayesian credible set and frequentist confidence interval and compare their empirical coverage rate and the mean length of the interval estimates.

Our observations from the simulation study are in order.

- First of all, in most cases except that the prior is centered near the truth, the RMSE of the Bayes estimators are larger than that of the frequentist estimators whether the baseline distributions are continuous or discrete. Similar phenomenon appears for the Bayesian interval estimators. The empirical coverage rate is smaller than the nominal coverage rate and the mean length is usually larger than that of the frequentist counter part. As is expected, when the prior is centered near the truth however, the Bayesian estimators and interval estimator performs superior.

- In all simulation experiments, these two Bayes estimators do not show much difference in the sense of RMSE, but the RMSE of the posterior median is

TABLE 3.3 *Frequentist Performance with sample size* 10 *and* F_0 *discrete.*

	$n = 10$		Posterior Mean		Posterior Median		Mean Length	Coverage
	c	a	bias	RMSE	bias	RMSE	of CI	Rate
BP1	.1	0.1	2.500	5.588	2.352	5.518	7.455	.680
		1.0	2.056	4.694	1.901	4.592	6.933	.675
		10.0	.2821	1.961	.1175	1.740	4.692	.794
	10.0	0.1	2.197	2.697	2.0268	2.429	4.686	.416
		1.0	.04656	.4802	.06856	.4583	2.428	.992
		10.0	-2.287	2.312	-2.238	2.261	1.947	.000
BP2	.1	0.1	2.518	5.766	2.321	5.619	7.587	.674
		1.0	1.834	4.478	1.661	4.336	6.553	.672
		10.0	-.7255	1.568	-.7536	1.498	.6368	.065
	10.0	0.1	2.1688	2.671	1.9994	2.402	4.688	.425
		1.0	.01674	.5033	.03284	.4781	2.515	.987
		10.0	-1.511	1.642	-1.499	1.628	.7625	.006
		efron	-.2283	.7984			2.936	.926
	MLE	breslow	-.4085	.7136			2.919	.954
		exact	4.855	11.935			3.0×10^5	.985

consistently smaller than that of the posterior mean except a few cases.

- When c is small ($c = 0.1$), the MLE and Bayes estimator match closely, but for a larger c value they do not match. See Figure 2.1.
- When the baseline distribution is discrete, the performance of exact is extremely poor. We don't know yet why this occur. But it seems that when there are many ties, the exact method should not be used.

In most small sample cases, since the normal approximation to the likelihood is usually poor, the Bayes estimator performs better than the MLE even in the frequentist sense. Accordingly, when we began the simulation experiments, we expected in these small sample experiments the Bayes estimator would perform superior to the MLE. Thus, the less than expected performance of the Bayes estimator came to us as a surprise. We don't understand yet why this behavior occurs. A partial explanation of this is that the partial likelihood is very well approximated by the normal distribution even with a small sample size. But, this does not explain why the performance is the Bayes estimator is so poor. Another

Table 3.4 *Frequentist Performance with sample size* 30 *and* F_0 *discrete.*

	$n = 30$		Posterior Mean		Posterior Median		Mean Length	Coverage
	c	a	bias	RMSE	bias	RMSE	of CI	Rate
BP1	.1	0.1	.3505	1.765	.3221	1.804	2.214	.849
		1.0	.2951	1.426	.2692	1.431	2.128	.839
		10.0	.08613	.8256	.05795	.7644	1.912	.840
	10.0	0.1	.9695	1.207	.9397	1.141	1.904	.470
		1.0	.1643	.3995	.1635	.3941	1.442	.942
		10.0	-1.749	1.764	-1.738	1.753	1.077	.000
BP2	.1	0.1	.3814	1.912	.3573	1.946	2.215	.844
		1.0	.1922	1.230	.1742	1.215	1.734	.728
		10.0	-.4888	.7760	-.4894	.7747	.04321	.014
	10.0	0.1	.9524	1.207	.9232	1.139	1.911	.491
		1.0	.1112	.4099	.1077	.4022	1.491	.941
		10.0	-1.032	1.109	-1.031	1.109	.04460	.003
		efron	-.1262	.4834			1.531	.888
	MLE	breslow	-.3835	.5096			1.520	.845
		exact	.9349	3.179			1.027×10^3	.892

possible explanation for this is that the frequentist model for the derivation of the partial likelihood is basically continuous hazard model while the Bayesian model is the discrete model to accommodate the discreteness of the beta process prior. But, still this does not fully explain the poor performance of the Bayes estimator even in the discrete model. We believe much of theoretical study should be done in small sample case. But the theoretical study of small sample behavior is inherently difficult, because the first order asymptotics do not make difference between the posterior and sampling distribution of the MLE. Thus, to study this behavior we need to study the second order asymptotics of the posterior distribution, which is more technical than the first order asymptotics.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer-Verlag, New York.

Hjort, N. L. (1990). "Nonparametric Bayes estimators based on beta processes in models for life history data", *The Annals of Statistics*, **18**, 1259–1294.

Kim, S. W. and Ibrahim, J. G. (2000). "On Bayesian inference for proportional hazards models using noninformative priors", *Lifetime Data Analysis*, **6**, 331–341.

KIM, Y. AND LEE, J. (2001). "On posterior consistency of survival models", *The Annals of Statistics*, **29**, 666–686.

KIM, Y. AND LEE, J. (2003). "Bayesian analysis of proportional hazard models", *The Annals of Statistics*, **31**, 493–511.

KIM, Y. AND LEE, J. (2004). "A Bernstein-von Mises theorem in the nonparametric right-censoring model", *The Annals of Statistics*, **32**, 1492–1512.

KLEIN, J. P. AND MOESCHBERGER, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, Springer, New York.

LAUD, P. W., DAMIEN, P. AND SMITH, A. F. M. (2004) . "Bayesian nonparametric and covariate analysis of failure time data", In *Practical nonparametric and semiparametric Bayesian statistics* (D. Dey, P. M'uller and D. Sinha, eds), 213–225.

LEE, J. AND KIM, Y. (2004). "A new algorithm to generate beta processes", *Computational Statistics & Data Analysis*, **47**, 441–453.