Bayesian Statistics and Its Applications
Edited by S.K. Upadhyay, U. Singh and D.K. Dey

Semiparametric Accelerated Failure Time Models for Censored Data

Sujit K.Ghosh and Subhashis Ghosal

Department of Statistics, North Carolina State University Raleigh, NC 27695-8203

Abstract

An accelerated failure time (AFT) semiparametric regression model for censored data is analyzed as an alternative to the widely used proportional hazards survival model. The statistical inference is based on a nonparametric Bayesian approach that uses a Dirichlet process prior for the mixing distribution. Consistency of the posterior distribution of the regression parameters in the Euclidean metric is established under certain conditions. Finite sample parameter estimates along with associated measure of un-certainties can be computed by a MCMC method. Simulation studies are presented to provide empirical validation of the new method. Analysis of data from two studies are provided to show the easy applicability of the proposed method.

1. Introduction

The framework of generalized linear models (GLM) popularized by McCullagh and Nelder (1989) is arguably one of the most flexible statistical tools and is used extensively in almost all fields of applications. However, such a flexible framework is not widely used in regression models for survival data with censoring. A key motivation of this article is to develop a flexible class of regression models that are similar in spirit to the GLM but retain the semiparametric structure of the widely used proportional hazard (PH) models. It is known that such an objective may be achieved by a class of mixture models where the mixing distribution is left unspeci fied (see Heckman and Singer, 1984, Ishwaran, 1996).

The Cox proportional hazard regression model (Cox, 1972) and the associated partial likelihood theory of estimation was a breakthrough in developing a flexible method of regression for censored data. The huge success of PH models testify to the many needs for this type of semiparametric regression models. The structure of PH is quite different from the usual GLM for regression, in that the link function is not speci fied via the mean but rather through the hazard function. The proportionality structure is interesting but it may be hard to interpret the regression coefficients. Sir D. Cox himself once remarked (Reid, 1994) that

"Of course, another issue is the physical or substantive basis for the proportional hazards model. I think that's one of its weakness, that accelerated life models are in many ways more appealing because of their quite direct physical interpretation, particularly in an engineering context."

An accelerated failure time (AFT) model is characterized by specifying the conditional survival

function, $S(t|\mathbf{Z}=z)$ of survival time, T given the covariates, $\mathbf{Z}=z$, as $S(t|\mathbf{z}) = S_0(t_g(\mathbf{z}^T\boldsymbol{\beta}))$, where $S_0(\cdot)$ is an unspecified baseline survival function. If $S_0(\cdot)$ is specified parametrically (with possibly some unknown finite-dimensional parameter), then we get a parametric AFT model. Under such parametric models, estimates of the regression coefficient $\boldsymbol{\beta}$ can be obtained by the maximum likelihood or Bayesian method, for a given completely known link function $g(\cdot)$. The main contributions of this article are (a) a semiparametric formulation of the accelerated failure time model and (b) development of asymptotic justifications of the model.

Among the various extensions of the traditional linear model, AFT models and the method of least squares to accommodate censored data seems very appealing, simply because the model is well known, widely used, well understood and well tested (Wei, 1992). Using a U-statistic representation, Koul et al. (1981) showed that their estimates are consistent and asymptotically normal under some regularity conditions. Following on the simple idea of using synthetic data, several extensions of the method have appeared in the literature that use a more efficient ways to obtain estimated responses (Zeng, 1984, Lai et al., 1995, Zhou, 1992 and references therein). These developments have been very exciting but generally lack stability of the estimators and hence are not as widely used as the PH model.

From a Bayesian perespective a wide class of semiparametric AFT regression models have been developed by, among others, Christensen and Johnson (1988), Johnson and Christensen (1989), Kuo and Mallick (1998), Walker and Mallick (1999), Campolieti (2001) and Hanson and Johnson (2004). However, to the best of our knowledge, none of the previous authors have attempted to formally prove the posterior consistency based on censored data from semiparametric AFT models. For the case of no censoring, Ishwaran (1998) presented some interesting results for Weibull semiparametric mixture under Dirichlet process and Ishwaran (1996) points out an important relationship between the constraint on mixing distribution and rates of estimation for the parameters in the model.

The proposed method of parameter estimation is quite similar in spirit to the previous approaches in terms of using the popular Markov Chain Monte Carlo (MCMC) methods. Moreover, we demonstrate that such Monte Carlo iterative methods can be easily implemented (see Appendix for a code) using freely available softwares like WinBUGS (Spiegelhalter et al., 1999). In addition to the computational stability and simplicity of the proposed method, we show that the posterior distribution is consistent under certain compactness assumptions on the parameter space.

In Section 2, the semi-parametric AFT model is presented using a mixture of Dirichlet processes. Section 3 discusses how an MCMC method can be easily implemented to obtain estimates based on the posterior distribution of the parameters. The posterior consistency of the proposed model is discussed in Section 4. Simulation studies are presented in Section 5. Finally, Section 6 presents couple of real data examples to illustrate our method. Proofs and codes are presented in the Appendices.

2. Semiparametric AFT Regression Model

Mixture models have been used for a variety of inference problems including density estimation, clustering analysis and robust estimation; see Lindsay (1995), McLachlan and Basford (1988), Banfield and Raftery (1993), Robert (1996) and Roeder and Wasserman (1997). In the Bayesian context, mixture models for density estimation were introduced by Ferguson (1983) and Lo (1984) who used a Dirichlet process prior on the mixing distribution and obtained expressions for Bayes estimates. Escobar and West (1995) developed this idea further and provided MCMC algorithms for the computation of the posterior distribution of parameters of a normal mixture model.

Consistency and rates of convergence issues for Bayesian density estimation in mixture models

have been studied by Ghosal et al. (1999), Ghosal and van der Vaart (2001), Ghosal (2001) and Petrone and Wasserman (2002). For survival models without covariates, consistency has been studied by Ghosh and Ramamoorthi (1994), Ghosh et al. (1999) and Kim and Lee (2001). In this article we propose using an infinite mixture of standard parametric distributions (such as the Weibull distributions) to model the conditional distribution of a survival time given a set of covariates.

For each subject $i = 1, \ldots, n$, let T_i denote the failure time and C_i denote the censoring time. The observed survival data are $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function; these and all other variables are independent across i. Let $\boldsymbol{Z}_i$ denote a p-dimensional vector of covariates associated with subject i. Assume that for each subject i the conditional survival function of T given $\boldsymbol{Z} = \boldsymbol{z}$ (suppressing the subscript i) is given by,

$$S(t|\boldsymbol{z}, \boldsymbol{\beta}) = \int_0^\infty S_b(t/\mu g(\boldsymbol{z}^T\boldsymbol{\beta}) dH(\mu) \tag{1}$$

where $S_b(\cdot)$ is a survival function with a specified functional form, but may involve additional unknown parameters. The link function $g : \mathbb{R} \to [0, \infty)$ is completely specified. The mixing distribution function $H(\cdot)$ is left unspecified, which will be estimated nonparametrically with the constraint $H(0) = 0$. It follows that the density of T given $\boldsymbol{Z} = \boldsymbol{z}$ is then given by

$$p(t|\boldsymbol{z}, \boldsymbol{\beta}) = \int_0^\infty p_b(t/\mu g(\boldsymbol{z}^T\boldsymbol{\beta})(\mu g(\ddagger^T\beta))^{-1} dH(\mu))$$

where p_b is the density function corresponding to the survival function S_b. Note that (1) leads to an AFT model with $S_0(t) = \int_0^\infty S_b(\frac{t}{\mu}) dH(\mu)$. In other words, the baseline survival function is modeled as a mixture of parametric survival function with an unknown mixing distribution.

Let $m = \int_0^\infty t p_b(t) dt = \int_0^\infty S_b(t) dt$ stand for the mean of the distribution with survival function S_b. Then

$$E(T|\boldsymbol{z}, \alpha, \boldsymbol{\beta}) = m g(\boldsymbol{z}^T\boldsymbol{\beta}) \int_0^\infty \mu dH(\mu) \tag{3}$$

In most applications a logarithmic link is chosen for g^{-1}, that is, $g(w) = e^w$. Putting in the equation (3), we obtain,

$$\log \frac{E(T|\boldsymbol{z}_1, \boldsymbol{\beta})}{E(T|\boldsymbol{z}_2, \boldsymbol{\beta})} = (\boldsymbol{z}_1 - \boldsymbol{z}_2)^T\boldsymbol{\beta}$$

which provides an easy interpretation of the parameter $\boldsymbol{\beta}$ as the unit change in the logarithm of the mean response with respect to a unit change in the covariate. Henceforth, we shall work with the logarithmic link function. Note that the mixing distribution $H(\cdot)$ has been left completely unspecified.

We shall use a Weibull survival function with an unknown shape parameter as a choice for S_b, i.e., we assume that $S_b(t, \alpha) = e^{t\alpha}$. The distribution with survival function e^{-t^a/λ^α} will be referred to as Weibull (α, λ) distribution, the Weibull distribution with scale parameter λ and shape parameter α. One may use other parametric families such as log normal or gamma. More generally, any parametric family supported on the positive half line may be considered. However, unlike the gamma or the log normal family, the Weibull distribution enjoys the advantage of the survival function being free of transcendental functions, and hence it will be easier to explicitly write down the likelihood for censored data. It may be noted that a mixture of Weibull distribution (as in (2)) can be seen as a location mixture model on a log-scale. More specifically, for Weibull mixtures, when we choose a

log-link for g^{-1}, the conditional density of $Y = \log T$ given $\boldsymbol{Z} = \boldsymbol{z}$ is given by

$$p(y|\boldsymbol{z}, \boldsymbol{\beta}) = \int_0^\infty e^y p_b(e^y/\mu g(\boldsymbol{z}^T\boldsymbol{\beta}))(\mu g(\boldsymbol{z}^T\boldsymbol{\beta}))^{-1} dH(\mu)$$

$$= \int_0^\infty \alpha f_b(\alpha(y - \log\mu - \boldsymbol{z}^T\boldsymbol{\beta}))(\mu g(\boldsymbol{z}^T\boldsymbol{\beta}))^{-1} dH(\mu)$$

where $f_b(t) = \exp(t - e^t)$ is Gumbel's extreme value density. Thus $\log\mu$ takes the role of a location parameter and α (the reciprocal of) a scale parameter in the above location mixture. Therefore, letting $\alpha \to \infty$, mixtures over μ can approximate any arbitrary density in the total variation distance. As the total variation distance is invariant under one-to-one measurable transformations, the same conclusion holds in the original scale. Thus, Weibull mixture can adequately represent an arbitrary density in (2).

A likelihood based method to estimate the parameter $\boldsymbol{\beta}$ would require maximization of over the space of all distributions $H(\cdot)$ with the property $H(0) = 0$ and the shape parameter α. Such an approach may yield efficient estimates for $\boldsymbol{\beta}$, and under regularity conditions, an asymptotic estimate of the standard error of the estimate can be obtained. Optimizations over such large space of parameters may be computationally challenging. Alternatively, one may use a set of estimating equations (see Jin et al. 2003) to obtain parameter estimates. However it turns out that for semiparametric AFT models, such estimating equations can be non-monotone and can lead to multiple solutions. In our nonparametric Bayesian approach, we use MCMC method to obtain the marginal posterior distribution of $\boldsymbol{\beta}$ by integrating out the nuisance parameters $H(\cdot)$ and α.

Prior distributions play a crucial role in Bayesian inference. Informative priors can be used if deemed plausible. However, in the present context, it is desirable to choose the prior distribution by some default mechanism. A priori, it is assumed that $(\alpha, \boldsymbol{\beta})$ independent of $H(\cdot)$, and that $H(\cdot)$ has a Dirichlet Process (DP) Prior, denoted by $DP(M, H_b(\cdot))$ with base measure $H_b(\cdot)$ and precision parameter M. The proposed model is equivalently written as a Bayesian hierarchical model below:

$$\begin{aligned} T_i|\boldsymbol{Z}_i, \alpha, \mu_i\ \boldsymbol{\beta} &\sim \text{Weibull}\ (\alpha, \mu_i \exp\left\{\boldsymbol{Z}_i^T\boldsymbol{\beta}\right\}) \\ \mu|H(\cdot) &\overset{\text{iid}}{\sim} H(\cdot) \\ H(\cdot) &\sim DP(M, H_b(\cdot)) \\ (\alpha, \boldsymbol{\beta}) &\sim \pi(\alpha, \boldsymbol{\beta}) \end{aligned} \tag{4}$$

where the joint density $\pi(\alpha, \boldsymbol{\beta})$ is usually chosen to be of the product type and often taken is to be the diffuse uniform prior on $[a_0, A_0] \times [-B_0, B_0]^p$, where the fixed positive constants a_0, A_0 and B_0 are chosen to approximate a non-informative prior. A small value of $a_0 > 0$ and large values for A_0 and $B_0 > 0$ are usually chosen. A moderate value of $M > 0$ is usually used. In practice, a sensitivity study is generally needed to elicit these constants (see Section 4). Let Π denote the joint prior distribution of $(\alpha, \boldsymbol{\beta}, H)$. Note that from the properties of a DP it follows that

$$E(H(\cdot)|\mu_1, \ldots, \mu_n) = \frac{M}{M+n} H_b(\cdot) + \frac{n}{M+n} H_n(\cdot)$$

where $H_n(\cdot)$ is the empirical distribution function of $\mu_1, \ldots, \mu_n$. However, these latent variables μ_1, $\ldots$, μ_n are not directly observable, and hence the posterior distribution may only be computed by MCMC methods. In the next section we discuss how to obtain the posterior distribution of the parameters given the observed data $\{(X_i, \Delta_i, \boldsymbol{Z}_i), i = 1, \ldots, n.\}$

3. Model Fitting

For the proposed PMR model, the posterior distribution of the parameters can not be obtained in a closed form. Gibbs sampling with a data augmentation step will be used to obtain MCMC samples.

First assume that T_i's are indeed observed. In this case, a simple strategy to obtain samples from the posterior distribution of $\alpha, \boldsymbol{\beta}, H(\cdot)$) can be described as follows (see MacEachern, 1998):

Gibbs Sampling: Initialize the parameters $\boldsymbol{\beta}^{(0)}$ and $\alpha^{(0)}$. At the k-th iteration of the Gibbs sampler, the full conditional density of μ_i given ($\boldsymbol{\beta}^{(k-1)}$, $\alpha^{(k-1)}$, $\mu_l^{(k-1)}$, $l \neq i, T_i, \boldsymbol{Z}_i$) is given by (upto a proportionality constant),

$$p_b(T_i|Z_i, \mu, \alpha^{(k-1)}, \boldsymbol{\beta}^{(k-1)} h_b(\mu), \qquad \text{with probability } q_{0i}^{(k)}$$

$$\mu_l^{(k-1)} I(l > i) + \mu_l^{(k)} I(l < i), \qquad \text{with probability } q_{li}^{(k)}, \quad l \neq i$$

where

$$q_{li}^{(k)} \propto \begin{cases} \mu_l^{(k)} p_b(T_i|\boldsymbol{Z}_i, \mu_l^{(k)}, \alpha^{(k-1)}, \boldsymbol{\beta}^{(k-1)}), & \text{if } l < i \\ \mu_l^{(k-1)} p_b(T_i|\boldsymbol{Z}_i, \mu_l^{(k-1)}, \alpha^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, & \text{if } l > i \end{cases}$$

and

$$q_{0i}^{(k)} \propto M \int_0^\infty p_b(T_i|\boldsymbol{Z}_i, \mu, \alpha^{(k-1)}, \boldsymbol{\beta}^{(k-1)}) h_b(\mu) d\mu$$

such that $q_{0i}^{(k)} + \sum_{l \neq i} q_{li}^{(k)} = 1$ and $h_b(\mu)$ denotes the density corresponding to $H_b(\mu)$. Note that if we choose a conjugate base measure $H_b(\cdot)$, the above integral can be evaluated analytically and hence samples can be obtained by an inversion method. Otherwise a numerical one-dimensional integral can be used to compute q_{0i}^k's. For our model we may achieve conjugacy by choosing an appropriate gamma distribution for μ_i^α. The full conditional density of (α, β) given $\{(\mu_i^{(k)}, T_i, Z_i),\ i = 1, \ldots, n\}$ is given by (upto a proportionality constant)

$$\prod_{i=1}^n p_b(T_i|\boldsymbol{Z}_i, \mu_i^{(k)}, \alpha, \boldsymbol{\beta}) H_b'(\mu_i^{(k)})$$

which is a log-concave density and hence an adaptive rejection sampling (see Gilks, 1992) can be used to sample from the above density. Thus a Gibbs sampler as described above can be used to obtain approximate samples from the joint posterior distribution of the parameters given the observed data. If desired, the posterior mean of $H(\cdot)$ can be obtained from the MCMC samples using (5).

When censoring is present, an imputation method can be used to "replace" the censored data by the imputed values. This is done as follows. If $\Delta_i = 1$, set $T_i = X_i$ otherwise set $T_i = (U_i + X_i^{\frac{1}{\alpha}})^\alpha$, where U_i's are independently exponentially distributed with means $\mu_i \exp\{\boldsymbol{Z}_i^T \boldsymbol{\beta}\}, i = 1, \ldots, n$.

Alternatively a finite approximation for DP (Ishawaran and Zarepour, 2002) can be used within the software **BUGS** to implement the Gibbs sampling. In fact, for our simulation study we use a finite approximation technique to implement the required Gibbs sampling using **BUGS**. This is achieved by introducing latent variables $\boldsymbol{L} = (L_1, \ldots, L_n)$ which indicate the group membership for the hidden variables μ_i's along with a probability vector $\boldsymbol{w} = (w_1, \ldots, w_N)^T$. More precisely, (4) can be written

as,

$$\begin{aligned} T_i|L_{i,}\boldsymbol{\alpha},\boldsymbol{\beta} &\sim \text{Weibull } (\alpha, \mu_{L_i} \exp\{\boldsymbol{Z}_i^T\boldsymbol{\beta}\}) \\ L_i|\boldsymbol{w} &\overset{\text{i.i.d}}{\sim} \text{ Multinomial } (\{1,\ldots,N\}, \boldsymbol{w}) \\ \mu_l &\overset{\text{i.i.d}}{\sim} H_b(\cdot), l = 1,\ldots,N \\ \boldsymbol{w} &\sim \text{ Dirichlet } \left(\frac{M}{N},\ldots,\frac{M}{N}\right) \\ (\alpha,\beta) &\sim \pi(\alpha,\beta) \end{aligned}$$

where N is large integer. Notice that, $\Pr[L_i = l] = w_l$ for $l = 1,\ldots,N$ where $\sum_{l=1}^{N} w_l = 1$. The above hierarchical framework which uses a finite dimensional Dirichlet distribution (with N possibly depending on n) is more convenient for programming in BUGS. For instance, Ishwaran and Zarepour (2002) suggested to use $N = \sqrt{n}$ for large n and $N = n$ for small n (see Section 5 of their article). However, more analytical work is necessary to choose the right order of N.

4. Consistency of Posterior Distribution

Consistency is an important desirable large sample property of the posterior distribution which provides a useful validation of a particular Bayesian method in use. For discussion and examples of a of inconsistency, the readers are referred to Diaconis and Freedman (1986) and Ghosh and Ramamoorthi (2003). A celebrated theorem of Schwartz (1965) gives sufficient conditions for posterior consistency in terms of conditions involving the existence of appropriate tests and the prior positivity of a neighborhood defined by the Kullback-Leibler divergence. Useful extensions of Schwartz's theorem are given by Barron et al. (1999) and Ghosal et al. (1999). The goal of the present section is to justify our Bayesian analysis of the AFT model for censored data by posterior consistency. Ishwaran (1998) has studied the consistency of similar posterior distributions for uncensored data.

We shall assume that the domains of $\boldsymbol{Z}, \boldsymbol{\beta}$ and α, and the support of H are compact. The compactness assumption is agreeably somewhat restrictive. Also, we assume, without loss of generality that $\mathbf{0}$ is a possible value of the covariate $\boldsymbol{Z}$. If not, we may shift the covariates to satisfy this condition. We further assume that the true density p_0 of T given $\boldsymbol{Z} = \boldsymbol{z}$ is actually a scale mixture of Weibull:

$$p_0(t|\boldsymbol{z}) = \int \alpha_0(\mu \exp\{\boldsymbol{\beta}_0^T\})^{-\alpha_0} t^{\alpha_0 - 1} \exp(-t/\mu \exp\{\boldsymbol{\beta}_0^T z\})^{\alpha\theta}) dH_0(\mu)$$

so that α_0, $\boldsymbol{\beta}_0$ and H_0 are respectively the true values of the parameters α, $\boldsymbol{\beta}$ and H. The covariates are assumed to be i.i.d. with an absolutely continuous distribution supporting the vector $\mathbf{0}$ on $\mathbb{R}^p$. The density of $\boldsymbol{Z}$ at $\boldsymbol{z}$ will be denoted by $q(\boldsymbol{z})$. Let $f_{\alpha,\boldsymbol{\beta},H}(z,\delta,\boldsymbol{z})$ stand for the joint density of $(X, \Delta, \boldsymbol{Z})$, that is

$$f_{\alpha,\boldsymbol{\beta},H}(x,\delta,\boldsymbol{z}) = \begin{cases} p(x|\alpha,\boldsymbol{\beta},\boldsymbol{z})q(\boldsymbol{z}), & \text{if } \Delta = 1 \\ S(x|\alpha,\boldsymbol{\beta},\boldsymbol{z})q(\boldsymbol{z}), & \text{if } \Delta = 0 \end{cases}$$

where $p(x|\alpha,\boldsymbol{\beta},\boldsymbol{z})$ and $S(x|\alpha,\boldsymbol{\beta},\boldsymbol{z})$ are as defined in equations (2) and (1), respectively. Note that the class of distributions that are supported in a given compact domain is also compact with respect

to the weak topology on the space of probability measures. Hence the parameter space of $(\alpha, \boldsymbol{\beta}, \boldsymbol{H})$ with respect to the product of Euclidean, Euclidean and weak topology, is also compact.

The following is the main theorem.

Theorem 1. Suppose that the prior density $\pi(\alpha, \boldsymbol{\beta})$ for $(\alpha, \boldsymbol{\beta})$ has compact support containing $(\alpha_0, \boldsymbol{\beta}_0)$ and the base measure H_b of the Dirichlet process has compact support that contains the support of H_0. Then the posterior distribution $\Pi((\alpha, \boldsymbol{\beta}, H) \in .|(X_1, \Delta_1, \ldots (X_n, \Delta_n))$ of $(\alpha, \boldsymbol{\beta}, H)$ given $(X_1, \Delta_1), \ldots (X_n, \Delta_n)$ is consistent with respect to the Euclidean distances on $\boldsymbol{\beta}$ and α and the weak topology on H, that is, given any $\varepsilon > 0$ and a weak neighborhood $\mathcal{N}$ of H_0,

$$\Pi\{(\alpha, \beta, H) : |\alpha - \alpha_0| < \varepsilon, |\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \varepsilon, H \in \mathcal{N}|(X_1, \Delta_1), \ldots, (X_n, \Delta_n)\} \to 1$$

almost surely in $P^{\infty}_{(\alpha_0, \beta_0, H_0)}$-probability.

This gives a large sample justification of our procedure. The proof of Theorem 1 is given in Appendix A.

If the covariates are not random, but arise deterministically from a design, then by largely a similar analysis and using posterior consistency results for independent, non-identically distributed observations as in Amewou-Atisso et al. (2003), consistency will follow if the values of the covariate gradually fill up its range space as the sample size increases. The details are omitted here.

5. A Simulation Study

We performed extensive simulations to explore the sampling properties of the Bayes estimates obtained by the proposed method. We present only some of the significant findings from our simulation experiments. In order to perform the simulation experiments, we generate the data from the model (2). In particular we use the following data generation process (DGP):

DGP for simulation: Fix true values $\boldsymbol{\beta}_0 = (-0.5, 0.5)$, $\alpha_0 = 1.0$ and the sample size $n = 50$, 100 and 200.

1. Generate $U_i \overset{\text{i.i.d.}}{\sim}$ Beta (3, 3) and set $\boldsymbol{Z}_i = 6U_i - 3$. This ensures that support of $\boldsymbol{Z}_i$'s is compact.
2. Generate $\mu_i \overset{\text{i.i.d.}}{\sim}$ Weibull ($\alpha^* = 1$, $\mu^* = 2$). This means that the mixing distribution is also a Weibull distribution.
3. Generate $T_i \overset{\text{ind.}}{\sim}$ Weibull $(\alpha_0 = 1, \mu_i \exp\{\boldsymbol{Z}_i^T \boldsymbol{\beta}_0\})$.
4. Generate $C_i \overset{\text{i.i.d.}}{\sim}$ Weibull $(\alpha^*, 2\mu^*)$. We also perform simulation where C_i's are allowed to depend on $\boldsymbol{Z}_i$'s. In particular we generate $C_i \overset{\text{ind.}}{\sim}$ Weibull $(\alpha^*, 4.5 \exp\{\boldsymbol{Z}_i^T \boldsymbol{\beta}\})$. These choices of censoring variable ensure that the censoring rate is about 30% on average.
5. Set $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, and output the observed data as: $\{(X_i, \Delta_i, \boldsymbol{Z}_i); i = 1, \ldots, n\}$.

We fit the semiparametric Bayesian model to the data generated (at step 5) from the model and repeat (steps 1-5 above) 500 times with a fixed burn-in time of 500 iterations followed by a sample of size 1000 from the posterior distribution of $\boldsymbol{\beta}$ and α. We have tried several combinations of true values, mixing distributions and censoring variables for our simulations. In order to have a compactly supported prior distribution we have used fixed values of the constants $a_0 = 0.01$, $A_0 = 100$, $B_0 = 10$ and $M = 1$ (see description right after (5)). Also, we have used $N = \sqrt{n}$ for our all applications to approximate a DPP. A sensitivity study with $N = n$ revealed no significant difference in the posterior estimates. In this article we present results based on only one set of true values for three different samples sizes and two different censoring mechanisms.

Table 1 Sampling properties of Bayes estimates when the censoring mechanism does not depend on covariates. True values: $\beta_1 = -0.5$, $\beta_2 = 0.5$, $\alpha = 1.0$. The standard deviation of the corresponding posterior summaries is denoted by s.e. (e.g., mean, sd etc.) over 500 replications

	$n = 50$					
	mean	sd	2.5%L	97.5%U	median	cp
β_1	−0.505	0.244	−1.000	−0.036	−0.501	0.927
s.e.	0.273	0.053	0.313	0.288	0.271	
β_2	0.516	0.248	0.055	1.032	0.534	0.928
s.e.	0.253	0.055	0.269	0.289	0.253	
α	1.116	0.241	0.721	1.641	1.134	0.948
s.e.	0.224	0.051	0.148	0.289	0.233	
	$n = 100$					
	mean	sd	2.5%L	97.5%U	median	cp
β_1	−0.491	0.166	−0.821	−0.165	−0.483	0.935
s.e.	0.172	0.027	0.192	0.170	0.173	
β_2	0.494	0.167	0.172	0.828	0.491	0.938
s.e.	0.177	0.027	0.178	0.197	0.176	
α	1.032	0.188	0.716	1.430	1.021	0.954
s.e.	0.176	0.042	0.115	0.235	0.183	
	$n = 200$					
	mean	sd	2.5%L	97.5%U	median	cp
β_1	−0.502	0.121	−0.753	−0.279	−0.504	0.949
s.e.	0.123	0.009	0.113	0.098	0.102	
β_2	0.497	0.118	0.261	0.728	0.496	0.948
s.e.	0.120	0.013	0.131	0.139	0.129	
α	0.996	0.152	0.727	1.305	0.967	0.951
s.e.	0.145	0.034	0.072	0.168	0.128	

Average observed censoring 29.25% (range 10–40%)

The results are presented in Tables 1 and 2 by increasing sample size. In each table we present the average five-number posterior summary values, viz. posterior mean, posterior sd, posterior 2.5% and 97.5% percentiles and the posterior median with the associated standard error (s.e.) from 500 Monte Carlo repetitions. In addition, we also present the coverage probability (cp) based on a 95% equal-tail posterior interval obtained by posterior 2.5% and 97.5% percentiles. For example, in Table 1, for $n = 100$, the entry -0.491 for β_1 is the average of 500 posterior means with s.e. 0.172 (which is the standard deviation of 500 posterior means). Similarly the entry 0.166 is the average of 500 posterior sd's with s.e. 0.027. Finally, a cp of 0.935 means that 93.5% of 500 equal-tail posterior 95% intervals contained the true value of $\beta_1 = -0.5$.

We observe that at an average censoring rate of about 30%, the Bayes estimates (posterior mean and median) are nearly unbiased as sample size increases and also the coverage probability of a 95% posterior interval approaches the nominal value of 95%. In addition, we see that the average length of a 95% posterior interval decreases with increasing sample size, thus making the intervals tighter. For instance, for β_2 the average length of 95% interval drops from 0.977 (for $n = 50$) to an average length of 0.467 (for $n = 200$). Similar observations can be made based on Table 2. These results numerically assert that consistency of posterior distribution of $\boldsymbol{\beta}$ and α. We repeated similar studies for other mixing distributions and for lower and higher censoring rates, the results were very similar to what we see in Tables 1 and 2.

Table 2 Sampling properties of Bayes estimates when the censoring mechanism depends on covariates. True values: $\beta_1 = -0.5$, $\beta_2 = 0.5$, $\alpha = 1.0$. The standard deviation of the corresponding posterior summaries is denoted by s.e. (e.g., mean, sd etc.) over 500 replications

	$n = 50$					
	mean	sd	2.5%L	97.5%U	median	cp
β_1	−0.487	0.234	−0.953	−0.028	−0.487	0.925
s.e.	0.258	0.053	0.283	0.282	0.258	
β_2	0.502	0.233	0.037	0.956	0.503	0.928
s.e.	0.248	0.052	0.266	0.278	0.248	
α	1.121	0.244	0.723	1.654	1.137	0.945
s.e.	0.239	0.053	0.153	0.309	0.247	
	$n = 100$					
	mean	sd	2.5%L	97.5%U	median	cp
β_1	−0.489	0.159	−0.819	−0.175	−0.489	0.939
s.e.	0.165	0.024	0.188	0.168	0.169	
β_2	0.495	0.163	0.179	0.817	0.493	0.940
s.e.	0.167	0.021	0.172	0.190	0.171	
α	1.029	0.179	0.710	1.422	1.018	0.953
s.e.	0.168	0.038	0.108	0.230	0.177	
	$n = 200$					
	mean	sd	2.5%L	97.5%U	median	cp
β_1	−0.502	0.115	−0.746	−0.280	−0.503	0.950
s.e.	0.123	0.009	0.113	0.098	0.102	
β_1	0.501	0.105	0.287	0.709	0.500	0.949
s.e.	0.110	0.010	0.126	0.133	0.124	
α	1.004	0.144	0.719	1.294	0.990	0.949
s.e.	0.140	0.028	0.067	0.160	0.121	

6. Applications to Data from Two Studies

We now apply our proposed models to some popular real data sets that have been analyzed by other authors. In particular we fit our semi-parametric models to ovarian cancer data set (with $n = 26$, small sample size) and multiple myeloma data set (with $n = 65$, moderate sample size). For both data sets we have used a compactly supported prior with $a_0 = 0.1$, $A_0 = B_0 = 10$ and $M = 1$. Also for Dirichlet prior approximation we have used $N = \sqrt{n}$.

6.1. Analysis of Ovarian Cancer Data

Consider a study on ovarian cancer as reported in Edmunson et al. (1979). In this study $n = 26$ patients were monitored and age for each patient were also recorded. Let X_i denotes the number of days patient i was on the study. For each patient, let A_i denote the age (recorded as number of days/365.25). It was of interest to find a relation between the X_i's and the A_i's using statistical regression methods. However in this study 53.84% of the observations were censored. We obtain the posterior distribution of β, under the model $\log[E(T|A)] = A\beta$.

Using the proposed semiparametric method based on Weibull mixture models, we obtain the summary of the posterior distribution of (β, α) using the Gibbs sampling algorithm described in Section 3. In order to maintain numerical stability, we transformed the X_i to $X_i/500$ and also standardized the covariate age to $Z_i = (A_i - \overline{A})/sd(A)$, where $\overline{A}$ and $sd(A) = 10.1$ denote the sample mean and standard deviation of A_i's. The results are summarized in Table 3. In addition, we can

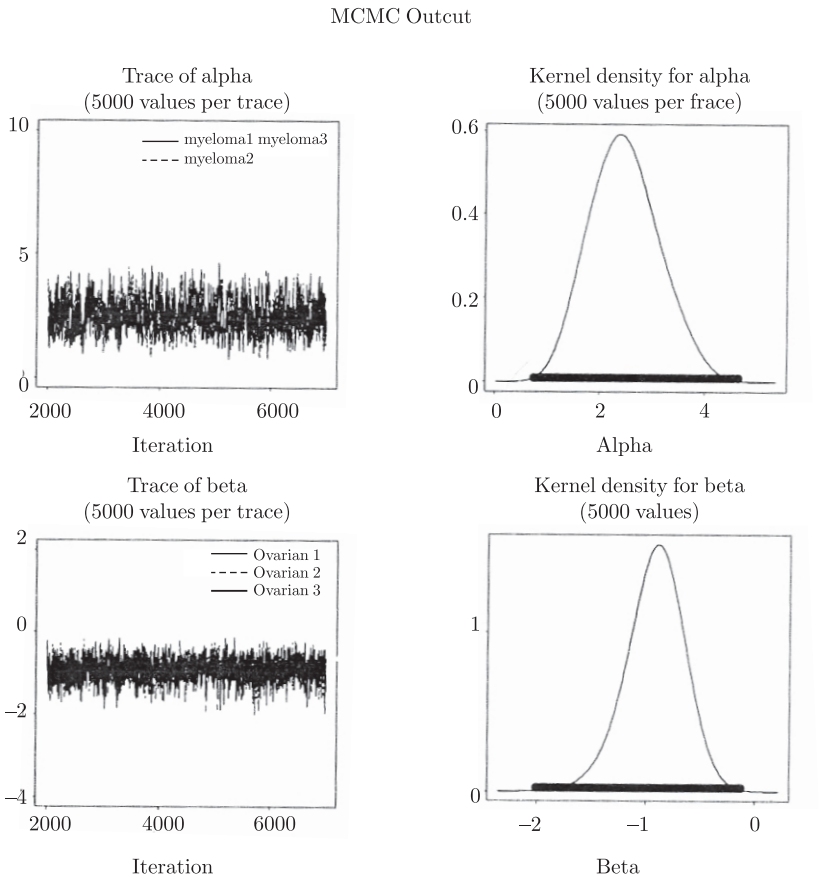

Fig. 1.

obtain the entire posterior density of α and β based on MCMC samples. In Figure 1, we plot the trace and kernel density estimates of (α, β) based 2000 burn-ins and 5000 samples from three independent chains. From the plots it appears that the chains mixed well and there were no apparent problems with convergence to the stationary distributions. The MCMC diagnostic software CODA available in Splus and R was also used to check convergence. The Gelman-Rubin 50% and 97.5% shrink factors (based on three dispersed starting values) were found to be 1.00 and 1.01, respectively for β, which indicate good mixing (as also evident from the trace plots in Fig. 1).

From Table 3, it is apparent that age has a significant negative effect on the mean survival time. More precisely, for unit increase in standardized age (i.e., days/365.25), the expected survival time decreases by a factor of 1.09 exp(0.9 /10.1) for patients with ovarian cancer. This type of straightforward interpretation is possible only for these kind of AFT models.

Table 3 Posterior summaries of the regression parameter (β) and the shape parameter (α)

	mean	sd	2.5%L	97.5%U	median
β	−0.922	0.243	−1.460	−0.487	−0.911
α	2.470	0.640	1.330	3.760	2.440

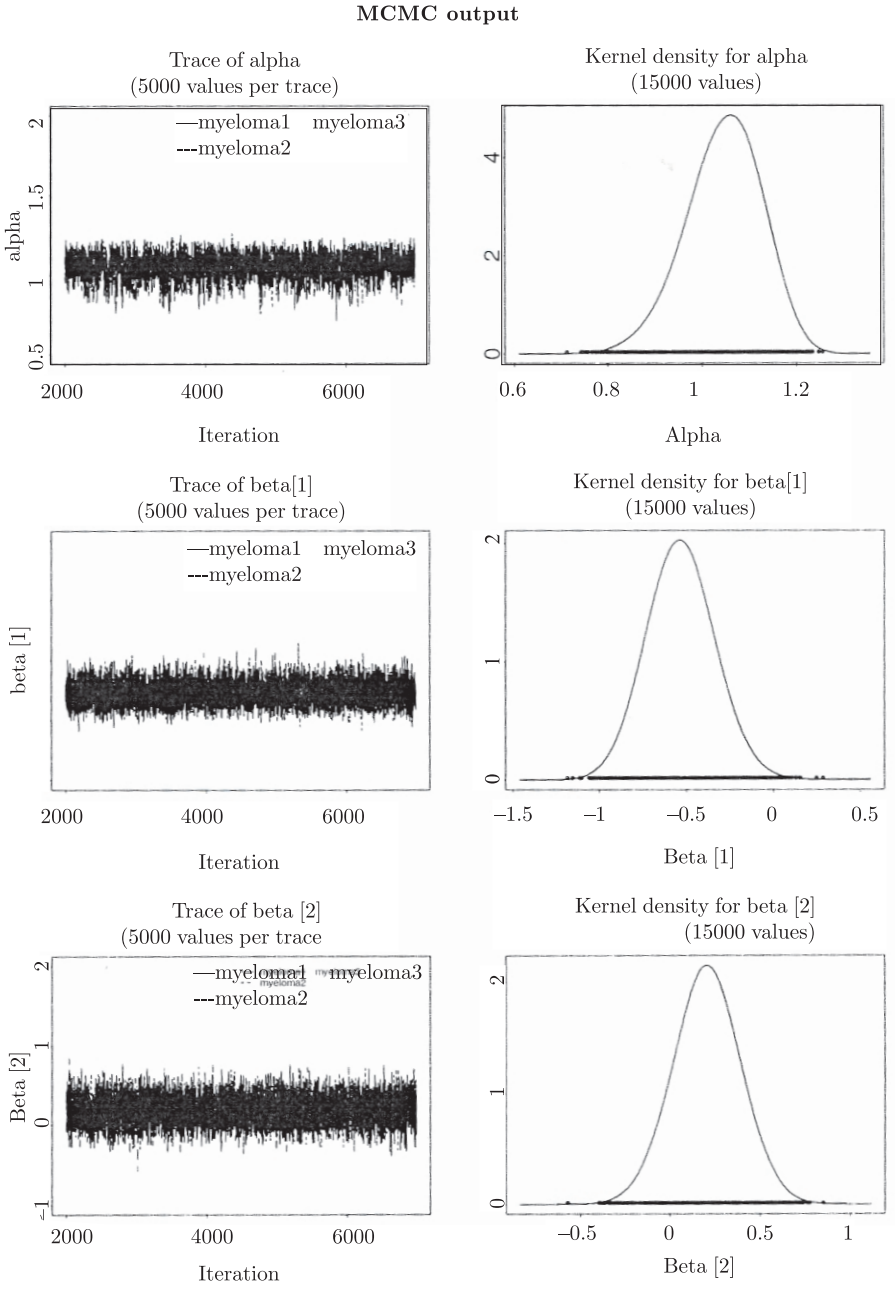

Fig. 2.

7. Analysis of Multiple Myeloma Data

We analyze a data set on multiple myeloma reported by Krall et al. (1975). This data has recently been analyzed by Jin et al. (2003) using a semiparametric AFT model. We compare our findings

Table 4 Posterior summaries of the regression parameter (β) and the shape parameter (α). Within ($\cdot$) corresponding frequentist estimates obtained by Jin et al. (2003) are presented.

	mean	sd	2.5%L	97.5%U	median
β_1	-0.534	0.175	-0.870	-0.190	-0.531
freq est.	(-0.532)	(0.146)	(-0.818)	(-0.246)	
β_2	0.202	0.168	-0.128	0.528	0.199
freq est.	(0.292)	(0.169)	(-0.039)	(0.623)	
α	1.040	0.074	0.885	1.170	1.050

to that of Jin et al. (2003). In this study $n = 65$ patients were treated with alkylating agents and survival time (in months) were monitored along with several prognostic variables.

In order to compare our results to that of Jin et al. (2003) we use the logarithm of the blood urea nitrogen measurement and the hemoglobin measurement, both measured at diagnosis as the covariates and standardize them to bring numerical stability. In other words, we fit a model such that $\log[E(T|\boldsymbol{Z})] = \boldsymbol{Z}^T\boldsymbol{\beta}$, where $\boldsymbol{Z} = (Z_1, Z_2)^T$ is the vector of standardized covariates and $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ the vector of regression coefficients to be estimated. We ran three independent parallel chains to draw MCMC samples from the posterior distribution of $(\alpha, \boldsymbol{\beta})$. Figure 2, shows the plot of trace and density estimates of $(\alpha, \boldsymbol{\beta})$ based 5000 samples from three parallel chains after a burn-in of 2000 samples. The Gelman-Rubin 50% and 97.5% shrink factors (based on three dispersed starting values) were found to be 1.00 and 1.01, respectively for $\boldsymbol{\beta}$, which indicate good mixing (as also evident from the trace plots in Fig. 2). It appears that there is no problem of convergence of the samples to stationary (posterior) distribution.

Based on the final 15000 samples we compute the posterior summaries of $(\alpha, \boldsymbol{\beta})$ and present these values in Table 4. First note that the posterior mean and median of $\boldsymbol{\beta}$ are quite similar to those obtained by Jin et al. (2003), which are presented in parenthesis. Also from the posterior MCMC samples of $\boldsymbol{\beta}$ we can easily obtain the posterior correlation coefficient of $\boldsymbol{\beta}$ to be $\text{Corr}(\beta_1, \beta_2|\text{data}) = -0.187$. This is

clearly an advantage of a Bayesian method over a frequentist method, as almost any posterior summary can be computed from the posterior samples generated by the MCMC method. Note that Jin et al. (2003) had to use a resampling method to compute the asymptotic variance covariance matrix of $\hat{\boldsymbol{\beta}}$. In contrast, our method produces the entire posterior density of $\boldsymbol{\beta}$ using a relatively simple code in BUGS (see Appendix B). Further, our interval estimates of $\boldsymbol{\beta}$ as presented in Table 4, are based on the exact (finite sample) posterior percentiles as compared to large sample normal percentiles of Jin et al. (2003), and hence our interval estimates seems to be wider (accounting for the true uncertainty) than those obtained by Jin et al. (2003).

From this study, it appears that the logarithm of the blood urea nitrogen measurement has a significant negative effect on the logarithm of the mean survival time, whereas hemoglobin does not seem to have a significant effect. We have not done a formal Bayesian testing (say using Bayes Factor). These conclusions are based on the concentration of posterior distribution around zero. Based on our study we may conclude that, for unit increase in the logarithm of blood urea nitrogen (in the standardized scale), the expected survival time decreases by a factor of 5.53 ($= \exp\ (0.534/0.312)$) for patients with multiple myeloma. This again illustrates the straightforward interpretation of the regression coefficients based on semiparametric AFT models.

Acknowledgements

The work of the seemd author has partially supported by the NSF Grant, DMS-0349111.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R.V. (2003). Posterior consistency for semiparametric regression problems. Bernoulli **9** 291–312.

Banfield, J. and Raftery, A. (1993). Model based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803-821.

Barron, A., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, **27** 536–561.

Campolieti, M. (2001). Bayesian semiparametric estimation of discrete duration models: an application of the Dirichlet process prior. *Journal of Applied Econometrics*, **16**, 1–22.

Christensen, R. and Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process. *Biometrika* **75**, 693–704.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc., Ser. B* **34** 187–202.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14**, 1–67.

Edmunson, J. H., Fleming, T. R., Decker, G. D., Malkasam, G. D., Kvols, L. K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal disease residual. *Cancer Treatment Reports*, **63**, 241–247.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577–588.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (Rizvi M., Rustagi, J. and Siegmund, D., Eds.) 287–302.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29** 1264–1280.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–158.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence of maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233–1263.

Ghosh, J. K. and Ramamoorthi, R. V. (1994). Consistency of Bayesian inference for survival analysis with or without censoring. In *Analysis of Censored Data* (Pune, 1994/1995), 95–103, IMS Lecture Notes Monogr. Ser., **27**, Inst. Math. Statist., Hayward, CA.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*, Springer-Verlag, New York.

Ghosh, J. K., Ramamoorthi, R. V. and Srikanth, K. R. (1999). Bayesian analysis of censored data. Special issue in memory of V. Susarla. *Statist. Probab. Lett.* **41**, 255–265.

Hanson, T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational Graphics and Statistics*, **13**, 341–361.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrika*, **52**, 271–320.

Ishwaran, H. (1996). Uniform rates of estimation in the semiparametric Weibull mixture model. *Annals of Statistics*, **24**, 1572–1585.

Ishwaran, H. (1998). Exponential posterior consistency via generalized Plya urn schemes in finite semiparametric mixtures. *Annals of Statistics*, **26**, 2157–2178.

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, **12**, 941–963.

Jin, Z. Lin, D.Y., Wei, L.J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, **90**, 341–353.

Johnson, W. and Christensen, R. (1989). Nonparametric Bayesian analysis of the accelerated failure time model. *Statistics and Probability Letters*, **8**, 179–184.

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29**, 666–686.

Koul, H., Sursula, V. and Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *Ann. Statist.* **9**, 1276–1288.

Krall, J. N., Utho., V. A. and Harley, J. B. (1975). A set-up procedure for selecting variables associated with survival. *Biometrics*, **31**, 49–57.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā Ser. B*, **60** 65–81.

Lai, T. L., Ying, Z. and Zheng, Z. (1992). Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. *J. Multivariate Anal.*, **52**, 259–279.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*, NSF- CBMS Regional Conference Series in Probability and Statistics **5**, Institute of Mathematical Statistics, Hayward, CA.

Lo, A. Y. (1994). On a class of Bayesian nonparametric estimates I: Density estimates. *Ann. Statist.* **12**, 351–357.

McCullagh P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd Edition), Chapman & Hall, London.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel-Dekker, New York.

Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc., Ser. B*, **64**, 79–100.

Reid, N. (1994). A conversation with Sir David Cox. *Statist. Sci.*, **9**, 439–455.

Robert, C. (1996). Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson and D. Spiegelhalter, eds.), Chapman and Hall, London, 441–464.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894–902.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahr. Verw. Gebiete* **4**, 10–26.

Spiegelhalter, D. J., Thomas, A. and Best, N. (1999). *WinBUGS Version 1.2 User Manual*, MRC Biostatistics Unit.

http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

Walker, S. and Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics*, **55**, 477–483

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, **11**, 1871–1879.

Zheng, Z. (1984). *Regression with Randomly Censored Data.* Ph.D Thesis, Columbia University, NY.

Zhou, M. (1992). Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. *Ann. Statist.*, **20**, 1002–1021.

Appendix A: Proof of the Main Result

Now we prove the main theorem by verifying the conditions of the following general result.

Theorem 2 *Let $Y_1, \ldots, Y_n$ be i.i.d. with density p lying in a class $\mathcal{P}$ of probability densities which is compact under the total variation metric. Let $\prod$ be a prior on $\mathcal{P}$ and p_0 be the true value of the density and let P_0 be the probability measure corresponding to p_0. If for all $\varepsilon > 0$,*

$$\prod \left\{ p \in \mathcal{P} : \int p_0 \log \frac{p_0}{p} < \varepsilon \right\} > 0$$

then the posterior distribution is consistent at p_0 under the total variation (or the L_1 distance a.s. $[P_0^\infty]$, that is, for all $\varepsilon > 0$,

$$\prod \left\{ p \in P : \int |p - p_0| < \varepsilon | Y_1, \ldots, Y_n \right\} \to 1 \qquad a.s.\ P_0^\infty$$

Proof. Observe that a compact set has finite metric entropy. Taking $\mathcal{P}$ itself as the sieve at each stage, (ii) of Theorem 2 of Ghosal et al. (1999) is satisfied. Alternatively, as $\mathcal{P}$ is compact under the variation distance, which gives a topology stronger than the weak one, the two topologies actually coincide on $\mathcal{P}$. Thus, total variation neighborhoods are also weak neighborhoods. As Schwartz's (1965) testing condition is satisfied for the weak topology (see Theorem 4.4.2 of Ghosh and Ramamoorthi, 2003), the result follows. □

To verify the conditions of Theorem 2, we shall need the following three lemmas.

Lemma 1. *The model is identifiable.*

Proof. This result follows from the discussion presented in Section 1 of Heckman and Singer (1984). □

Lemma 2. *Consider the product topology on $(\alpha, \boldsymbol{\beta}, H)$, where α and $\boldsymbol{\beta}$ are given the usual Euclidean topology and H the weak topology. On the densities $f_{\alpha,\boldsymbol{\beta},H}(x, \delta, \boldsymbol{z})$, put the total variation (or the L_1) distance defined as*

$$||f_{\alpha_1,\boldsymbol{\beta}_1,H_1} - f_{\alpha_2,\boldsymbol{\beta}_2,H_2}|| = \int\int |f_{\alpha_1,\beta_1,H_1}(x, \delta, \boldsymbol{z}) - f_{\alpha_2,\boldsymbol{\beta}_2,H_2}(x, \delta, \boldsymbol{z})| dx d\boldsymbol{z} \tag{6}$$

Then

$$||f_{\alpha_\nu,\beta_\nu,H_\nu} - f_{\alpha,\beta,H}|| \to 0$$

if and only if $(\alpha_\nu, \beta_\nu, H_\nu) \to (\alpha, \boldsymbol{\beta}, H)$. In other words, the variation topology on the densities is equivalent to the product topology on the indexing parameters.

Proof. For the "if" part, it suffices to show that the densities converge pointwise and then apply Scheffe's theorem. Fix $(x, \delta, \boldsymbol{z})$. We shall show the proof only for $\delta = 1$.

As μ has a fixed compact range, the integrand

$$\alpha x^{\alpha-1} (\mu \exp\{\boldsymbol{z}^T \boldsymbol{\beta}\})^{-\alpha} \exp[-(x/\mu \exp\{\boldsymbol{z}^T \boldsymbol{\beta}\})^\alpha]$$

in the definition of $f_{\alpha,\boldsymbol{\beta},H}$, as a family of functions of $(\alpha, \boldsymbol{\beta})$ indexed by μ is equicontinuous. For a given $\varepsilon > 0$, find ν large enough so that the integrands are ε-close for all μ. For such a ν, in the decomposition,

$$|f_{\alpha_\nu,\boldsymbol{\beta}_\nu,H_\nu} - f_{\alpha,\boldsymbol{\beta},H}| \leqslant |f_{\alpha_\nu,\boldsymbol{\beta}_\nu,H_\nu} - f_{\alpha,\boldsymbol{\beta},H_\nu}| + |f_{\alpha,\boldsymbol{\beta},H_\nu} - f_{\alpha,\boldsymbol{\beta},H}| \tag{7}$$

the first term on the RHS is clearly less than ε. The second term on the RHS converges to 0 as $\nu \to \infty$ since H_ν converges weakly to H and the integrand is a bounded continuous function of μ.

To prove the "only if" part, consider any subsequence of the original sequence. By the compactness of the domains of $\alpha, \boldsymbol{\beta}$ and μ, we can extract a further subsequence along which $\alpha_\nu \boldsymbol{\beta}_\nu$ and H_ν converge in the respective topologies to α^*, $\boldsymbol{\beta}^*$ and H^*, say. Thus by the "if" part, along that subsequence, $||f_{\alpha_\nu,\beta_\nu,H_\nu} - f_{\alpha^*,\boldsymbol{\beta}^*,H^*}|| \to 0$. Hence $f_{\alpha,\boldsymbol{\beta},H} = f_{\alpha^*,\boldsymbol{\beta}^*,H^*}$, and so by Proposition 1, $(\alpha, \beta, H) = (\alpha^*, \beta^*, H^*)$. Therefore the original sequence must converge to (α, β, H). □

Lemma 3. *For all $\varepsilon > 0$,*

$$\prod \left\{ (\alpha, \boldsymbol{\beta}, H) : \int \int f_{\alpha_0,\boldsymbol{\beta}_0,H_0} \log \frac{f_{a_0,\boldsymbol{\beta}_0,H_0}}{f_{\alpha,\boldsymbol{\beta},H}} dx d\boldsymbol{z} < \varepsilon \right\} > 0$$

Proof. The log likelihood ratio is given by

$$\Lambda(\alpha, \boldsymbol{\beta}, H) = \begin{cases} \log \dfrac{p_{\alpha,\boldsymbol{\beta},H}(x, \boldsymbol{z}, \delta)}{p_{\alpha_0,\boldsymbol{\beta}_0,H_0}(x, \boldsymbol{z}, \delta)}, & \text{if } \delta = 1 \\ \log \dfrac{s_{\alpha,\boldsymbol{\beta},H}(x, \boldsymbol{z}, \delta)}{s_{\alpha_0,\boldsymbol{\beta}_0,H_0}(x, \boldsymbol{z}, \delta)}, & \text{if } \delta = 0 \end{cases}$$

Thus, the Kullback-Leibler divergence breaks up into two terms corresponding to $\delta = 1$ and $\delta = 0$. We shall bound the contribution corresponding to $\delta = 1$, the other case being simpler. Note that $\boldsymbol{\beta}, \alpha, \boldsymbol{z}$ are all bounded. Thus the integrand defining $f_{\alpha,\boldsymbol{\beta},H}$ is bounded above and below by functions of the form $Cx^{k_1} e^{cx_2^k}$. Taking ratio and then logarithm, Λ in the tails (in x) is bounded by a multiple of a power of x. This is integrable with respect to the true density, and hence we can find a central region, outside which, the contribution to the integral $\int \int f_{\alpha_0,\boldsymbol{\beta}_0,H_0} \log \frac{f_{\alpha_0,\boldsymbol{\beta}_0,H_0}}{f_{a,\boldsymbol{\beta},H}} dx d\boldsymbol{z}$ is small. In the central region, the family indexed by $(x, \boldsymbol{z})$ is equicontinuous in the parameters. Rest of the proof follows as in that of Theorem 3 of Ghosal et al. (1999). □

Proof of Theorem 1. Lemma 2 shows that it suffices to consider neighborhoods with respect to the L_1-distance.Also, the space is compact. The condition of prior positivity has been verified in Lemma 3. The proof therefore follows. □

Appendix B: BUGS Code for AFT Model

```
model myeloma;
const
n=65, N=8, p=2, b0=-10, B0=10, M=1, a0=0.1, A0=10;
var
X[n], alpha, lambda[n], Cen[n], eta[N], Z[n, p], delta[p], beta[p],
latent[n], prob[N], mu[N], a[N];
data X, Cen, Z in "myeloma.dat";
inits in "myeloma.in";
{
for(i in 1:n){
X[i] ~ dweib(alpha, lambda[i])I(Cen[i], );
log(lambda[i]) <- eta[latent[i]] + inprod(Z[i, ], delta[]);
latent[i] ~ dcat(prob[]);}
```

```
for(j in 1:p){
delta[j] ~ dunif(b0, B0);
beta[j] <- -delta[j]/alpha; }
prob[1:N] ~ ddirch(a[]);
for(k in 1:N){
eta[k] ~ dunif(b0, B0);
mu[k] <- exp(-eta[k]/alpha);
a[k] <- M/N;}
alpha ~ dunif(a0, 0);
}
```

We have used the unix version of classic BUGS to fit our models. This code has not been tested on latest version of WinBUGS