Sankhyā: The Indian Journal of Statistics
Special Issue on Bayesian Analysis
1998, Volume spl, Series , Pt. 1, pp. 48-64

PRIOR DISTIBUTIONS ANS BAYESIAN COMPUTATION FOR PROPORTIONAL HAZARDS MODELS

By JOSEPH G. IBRAHIM*
Harvard School of Public Health
and
Dana-Farber Cancer Institute, Boston
and
MING-HUI CHEN**
Worcester Polytechnic Institute, Worcester

SUMMARY. In this paper, we propose a class of informative prior distributions for Cox's proportional hazards model. A novel construction of the prior is developed based on the notion of the availability of historical data. In many situations, especially in clinical trials, the investigator has historical data from past studies which are similar to the current study. We take a semi-parametric approach in that a non-parametric prior is specified for the hazard rate and a fully parametric prior is specified for the regression coefficients. The prior specifications focus on the observables in that the elicitation is based on a prior prediction y_0 for the response vector and a quantity a_0 quantifying the uncertainty in y_0. Then, y_0 and a_0 are used to specify a prior for the regression coefficients in a semi-automatic fashion. One of the main applications of our proposed priors is for model selection. Efficient computational methods are proposed for sampling from the posterior distribution and computing posterior model probabilities. A real data set is used to demonstrate our methodology.

1. Introduction

Historical data are often available in applied research settings where the investigator has access to previous studies measuring the same response and covariates as the current study. For example, in many cancer and AIDS clinical trials, current studies often use treatments that are very similar or slight modifications of treatments used in previous studies. In carcinogenicity studies,

---

AMS (1991) *subject classification.* 62C10
Key words and phrases. Cox model; Gibbs sampling; Historical Data; Nonparametric prior; Posterior distribution; Variable Selection.
∗ Research supported by NIH grants # CA 70101-01 and # CA 74015-01.
∗∗ Research supported by NSF grant No. DMS-9702172 and NIH grant #CA 74015-01.

large historical databases exist for the control animals from previous experiments. It is natural to incorporate the historical data into the current study by quantifying it with a suitable prior distribution on the model parameters. Our proposed methodology can be applied to each of these situations as well as in other applications that involve historical data.

From a Bayesian perspective, historical data from past similar studies can be very helpful in interpreting the results of the current study. However, very few methods exist for the formal incorporation of previous data to construct the prior distribution. There is some literature addressing this issue for the linear model and generalized linear models. See for example, Ibrahim and Laud (1994), Laud and Ibrahim (1995), Ibrahim, Ryan, and Chen (1998), Chen, Ibrahim, and Yiannoutsos (1999), and Bedrick, Christensen, and Johnson (1996). There is limited literature, however, on prior elicitation for the semi-parametric Cox model. Ibrahim, Chen, and MacEachern (1996) propose a class of semi-parametric informative priors for the Cox model and examine the problems of prior elicitation and variable subset selection. The priors they propose are quite different than the ones we propose here.

In this paper, we develop a class of semi-parametric informative prior distributions for the Cox model (Cox, 1972). We specify a non-parametric prior for the baseline hazard rate and a parametric prior for the regression coefficients. Our priors are best suited for situations where historical data is available. We also develop novel Markov chain Monte Carlo (MCMC) techniques for sampling from the posterior distribution of the parameters. Our prior specification focuses on the observables in that the elicitation is primarily based on a prior prediction y_0 for the response vector and a quantity a_0 quantifying the uncertainty in y_0. Then, y_0 and a_0 are used to specify the prior in a semi-automatic fashion. This type of elicitation scheme is quite attractive in variable subset selection, for example, since all that is needed is a one time specification of the prior prediction y_0 and a_0. This proves to be a useful approach for this problem since specifying meaningful prior distributions for the parameters in each model is a difficult task requiring contextual interpretations of a large number of parameters. We illustrate our priors in the model selection context in Section 5.

Semi-parametric Bayesian analyses of proportional hazards models are becoming computationally feasible due to modern technology and recent advances in computing techniques such as the Gibbs sampler (Gelfand and Smith, 1990) and other MCMC methods. Sinha and Dey (1997) give a nice overview of Bayesian semiparametric methods for the Cox model. The potential advantage of using Bayesian methods to jointly model the baseline hazard and the regression coefficients is that one can accurately compute posterior quantities of interest using MCMC simulation techniques. However, there still remains the chore of specifying meaningful prior distributions and doing intensive computations. We propose a methodology for this here and give the technical details in Sections 2 - 4.

The remainder of the article is organized as follows. In Section 2.2 we give the approximate likelihood based on the semi-parametric model. In Section 2.3, we propose the prior distributions for the model parameters, and in Section 2.4, we discuss elicitation of prior parameters. In Section 3, we discuss novel methods for sampling from the posterior distribution. In Sections 4.1 and 4.2, we discuss methods for computing prior and posterior model probabilities for variable subset selection. We demonstrate the priors with a real example in Section 5. We conclude the article with a discussion section.

2. **The Method**

2.1 *Model and notation.* A proportional hazards model is defined by a hazard function of the form

$$h(t,x) = h_b(t)\exp(x'\beta) \ , \qquad \ldots (2.1)$$

where $h_b(t)$ denotes the baseline hazard function at time t, x denotes the $p \times 1$ covariate vector for an arbitrary individual in the population, and β denotes a $p \times 1$ vector of regression coefficients. The likelihood function for a set of right censored data on n individuals for the current study in a proportional hazards model based on (2.1) is given by

$$L\left(\beta, h_b(t)\right) = \prod_{i=1}^{n} \left[h_b(t_i)\exp(\eta_i)\right]^{\nu_i} \left(S_b(t_i)^{\exp(\eta_i)}\right), \qquad \ldots (2.2)$$

where $\eta_i = x_i'\beta, t_i$ is an observed event time or censoring time for the i^{th} individual and ν_i is the indicator variable taking on the value 1 if t_i is an event time, and 0 if it is a censoring time. Moreover, x_i is a $p \times 1$ vector of covariates for the i^{th} individual, and X denotes the $n \times p$ covariate matrix of rank p. Further, $S_b(\cdot)$ denotes the baseline survivor function, which, since we consider continuous survival distributions, is related to $h_b(\cdot)$ by $S_b(t) = \exp\left(-\int_0^t h_b(u)\ du\right)$.

2.2 *The Approximate likelihood function.* We first construct a finite partition of the time axis. Let $0 \le s_0 < s_1 < \ldots < s_M$ denote this partition with $s_M > t_i$ for all $i = 1, \ldots, M$. Further, let

$$\delta_i = h_b(s_i) - h_b(s_{i-1})$$

denote the increment in the baseline hazard in the interval $(s_{i-1}, s_i]$, $i = 1, \ldots, M$, and let $\Delta = (\delta_1, \ldots, \delta_M)$. We follow Ibrahim, Chen, and MacEachern (1996) for constructing the approximate likelihood function of (β, Δ). For an arbitrary individual in the population, the cumulative distribution function for the proportional hazards model at time s is given by

$$\begin{aligned} F(s) &= 1 - \exp\left\{-\exp\{\eta\} \int_0^s h_b(t)\ dt\right\} \\ &\simeq 1 - \exp\left\{-\exp\{\eta\} \ \left((s-s_0)h_b(s_0) + \sum_{i=1}^M \delta_i (s - s_{i-1})^+\right)\right\}, \end{aligned} \quad \ldots (2.3)$$

where $(t)^+ = t$ if $t > 0$, 0 otherwise, and $\eta = x'\beta$. We assume here that $h_b(s_0) = 0$, and $F(s) = 1$ for $s > s_M$, so that (2.3) is slightly simplified. This first approximation arises since the specification of Δ does not specify the entire hazard rate, but only the δ_i. For purposes of approximation, we take the increment in the hazard rate, δ_i, to occur immediately after s_{i-1}. Let p_i denote the probability of a failure in the interval $(s_{i-1}, s_i], i = 1, \ldots, M$. Using the fact that $h_b(s_0) = 0$, we have

$$\begin{aligned} p_i &= F(s_i) - F(s_{i-1}) \\ &\simeq \exp\left\{-\exp\{\eta\} \sum_{j=1}^{i-1} \delta_j (s_{i-1} - s_{j-1})\right\} \\ &\quad \left[1 - \exp\left\{-\exp\{\eta\}(s_i - s_{i-1}) \sum_{j=1}^{i} \delta_j\right\}\right] . \end{aligned}$$

Thus, in the i^{th} interval $(s_{i-1}, s_i]$, the contribution to the likelihood function for an exact observation (i.e., a failure) is p_i and $1 - F(s_i)$ for a right censored observation. Let d_i be the number of failures and c_i be the number of right censored observations in the i^{th} interval, respectively, $i = 1, \ldots, M$. For ease of exposition, we order the observations so that in the i^{th} interval the first d_i are failures and the remaining c_i are right censored, $i = 1, \ldots, M$. Let x_{ik} denote the vector of covariates for the k^{th} individual in the i^{th} interval and define

$$\begin{aligned} u_{ik}(\beta) &= \exp\{x'_{ik}\beta\}\ , \\ a_i &= \sum_{j=i+1}^M \sum_{k=1}^{d_j} u_{jk}(\beta)(s_{j-1} - s_{i-1})\ , \\ b_i &= \sum_{j=i}^M \sum_{k=d_j+1}^{d_j+c_j} u_{jk}(\beta)(s_j - s_{i-1})\ , \\ T_i(\Delta) &= (s_i - s_{i-1})\ \sum_{j=1}^i \delta_j\ . \end{aligned} \quad \ldots (2.4)$$

Let $D = (n, t, X, \nu)$ denote the data for the current study. The likelihood function given the data D for the current study over all M intervals is given by

$$L(\beta, \Delta \mid D) = \left\{\prod_{i=1}^M \exp\left\{-\delta_i(a_i + b_i)\right\}\right\} \left\{\prod_{i=1}^M \prod_{k=1}^{d_i}\ (1 - \exp\{-u_{ik}(\beta) T_i(\Delta)\})\right\}\ . \quad \ldots (2.5)$$

We note that this likelihood involves a second approximation. Instead of conditioning on exact event times, we condition on the intervals in which events occur, and thus we approximate continuous right censored data by interval censored data.

2.3 *The prior distributions.* For ease of exposition, we assume that we have one previous study, as the extension to multiple previous studies is straightforward. Let $D_0 = (n_0, y_0, X_0, \nu_0)$ denote the data from the previous study, where n_0 denotes the sample size of the previous study, y_0 denotes a right censored vector of survival times with censoring indicators ν_0, and X_0 denotes the $n \times p$ matrix of covariates. We shall refer to D_0 as the historical data throughout.

In general, for most problems, there are no firm guidelines on the method of prior elicitation. Typically, one tries to balance sound theoretical ideas with practical and computationally feasible ones. The issue of how to incorporate D_0 into the current study has no obvious solution since it depends in large part of how similar the two studies are. In most clinical trials no two studies will ever be identical. In many cancer clinical trials for example, the patient populations typically differ from study to study even when the same regimen is used to treat the same cancer. In addition, other factors such as institutional and measurement effects may make the two studies heterogeneous. Due to these differences, an analysis which combines the data from both studies may not be desirable. In this case, it may be more appropriate to "weight" the data from the previous study so as to control its impact on the current study. Thus, it is desirable for the investigators to have a prior distribution that summarizes the historical data in an efficient and useful manner and allows them to tune or weight D_0 as they see fit in order to control its impact on the current study.

Towards this goal, let $\pi_0(\beta, \Delta)$ denote the joint prior distribution for (β, Δ) used in the previous study. Thus, $\pi_0(\beta, \Delta)$ can be viewed as the "initial prior" for the parameters before D_0 was observed. For the current study, we propose joint prior distribution for (β, Δ) of the form

$$\pi(\beta, \Delta \mid D_0, a_0) \ \propto \ L(\beta, \Delta \mid D_0)^{a_0} \ \pi_0(\beta, \Delta) \ , \qquad \ldots (2.6)$$

where $L(\beta, \Delta \mid D_0)$ is the likelihood function of (β, Δ) based on the data from previous study (i.e., D_0), and thus, $L(\beta, \Delta \mid D_0)$ is (2.5) with D replaced by $D_0 = (n_0, y_0, X_0, \nu_0)$. The parameter a_0 is a scalar prior parameter that controls the influence of the historical data on the prior distribution and can be interpreted as a precision parameter. It is reasonable to restrict the range of a_0 to be between 0 and 1, and thus we take $0 \le a_0 \le 1$. Setting $a_0 = 1$ in (2.6) corresponds to the formal update of $\pi_0(\beta, \Delta)$ using Bayes theorem. That is, with $a_0 = 1$, (2.6) corresponds to the posterior distribution of (β, Δ) from the previous study. When $a_0 = 0$, no historical data are incorporated into the prior, and in this case, $\pi(\beta, \Delta \mid D_0, a_0 = 0) = \pi_0(\beta, \Delta)$. The prior in (2.6) can be viewed as a generalization of the usual Bayesian update of $\pi_0(\beta, \Delta)$. The parameter a_0 allows the investigator to control the influence of the historical data on the current study. As mentioned earlier, such control is desirable in situations where there is heterogeneity between the previous and current study, or the sample sizes of the two studies are far apart. In these cases, one would not want to equally weight the historical data and the current data.

To simplify the prior specification, we take

$$\pi_0(\beta, \Delta) = \pi_0(\beta \mid c_0)\ \pi_0(\Delta \mid \theta_0)\ ,$$

where c_0 and θ_0 are fixed hyperparameters. Specifically, we take a p dimensional multivariate normal density for $\pi_0(\beta \mid c_0)$ with mean 0 and covariance matrix $c_0 W_0$, where c_0 is a specified scalar and W_0 is a specified $p \times p$ diagonal matrix. Thus,

$$\pi_0(\beta \mid c_0) = (2\pi)^{-p/2}\ c_0^{-p/2} \mid W_0 \mid^{-1/2}\ \exp\left\{-\frac{1}{2c_0}\beta' W_0^{-1}\beta\right\}. \qquad \ldots (2.7)$$

We take $\pi_0(\Delta \mid \theta_0)$ to have a gamma density of the form

$$\pi_0(\Delta \mid \theta_0) \propto \prod_{i=1}^{M}\ \delta_i^{f_{0i}-1}\ \exp\{-\delta_i g_{0i}\}\ , \qquad \ldots (2.8)$$

where $\theta_0 = (f_{01}, g_{01}, \ldots, f_{0M}, g_{0M})$. Thus, $\pi_0(\Delta \mid \theta_0)$ consists of a product of M independent gamma densities, each with mean f_{0i}/g_{0i} and variance f_{0i}/g_{0i}^2, $i = 1, \ldots, M$. The motivation for (2.8) is that it is a discrete approximation to an underlying gamma process prior for $h_b(t)$. Such a prior appears to be reasonable for the context discussed here. Ibrahim, Chen, and MacEachern (1996) discuss the gamma process priors for the baseline hazard rate in detail. We note that in (2.6), (β, Δ) are not independent, and also the components of Δ are not independent a priori.

The prior specification is completed by specifying a prior for a_0. Since $0 \leq a_0 \leq 1$, a beta prior is reasonable for a_0, so that

$$\pi(a_0 \mid \alpha_0, \lambda_0) \propto a_0^{\alpha_0 - 1}(1 - a_0)^{\lambda_0 - 1}. \qquad \ldots (2.9)$$

Thus we propose a joint prior for (β, Δ, a_0) to be of the form

$$\pi(\beta, \Delta, a_0 \mid D_0)\ \propto\ L(\beta, \Delta \mid D_0)^{a_0}\ \pi_0(\beta \mid c_0)\ \pi_0(\Delta \mid \theta_0)\ \pi(a_0 \mid \alpha_0, \lambda_0). \qquad \ldots (2.10)$$

It can be shown that (2.10) is proper and is best suited for situations in which a historical data set is available. However, the prior can still be used if a previous experiment does not exist for the current study. In this case, D_0 can be elicited from expert opinion, case-specific information about each individual in the current study, or perhaps by a theoretical model giving predictions for the event times. The elicitation scheme is less automated than the one in which a previous study exists. In this situation, there are a number of ways one could elicit y_0, and this elicitation depends on the context of the problem. We do not attempt to give a general elicitation scheme for y_0 in this setting, but rather mention some general possibilities. In any case, the cleanest specification of D_0 is to use data from a previous study for which y_0 would be taken to be the vector of survival times from the previous study and X_0 is taken as the design matrix from the previous study.

2.4 *Choices of prior parameters.* The choices of the prior parameters play a crucial role in any Bayesian analysis. We first discuss elicitation of (α_0, λ_0). For the purposes of prior elicitation, it is easier to work with

$$\mu_0 = \alpha_0/(\alpha_0 + \lambda_0)$$

and

$$\sigma_0^2 = \mu_0(1 - \mu_0)(\alpha_0 + \lambda_0 + 1)^{-1}.$$

A uniform prior (i.e., $\alpha_0 = \lambda_0 = 1$), which corresponds to $(\mu_0, \sigma_0^2) = (1/2, 1/12)$ may be a suitable noninformative starting point, and facilitates a useful reference analysis for other choices. The investigator may choose μ_0 to be small (say $\mu_0 \leq .1$), if he/she wishes to have low prior weight on the historical data. If a large prior weight is desired, then $\mu_0 \geq .5$ may be desirable. It is reasonable to choose σ_0^2 in the range $\mu_0/1000 \leq \sigma_0^2 \leq \mu_0/10$. In any case, in an actual analysis, we recommend that several choices of (μ_0, σ_0^2) be used, including ones that give small and large weight to the historical data, and several sensitivity analyses conducted. We do not recommend doing an analysis based on one set of prior parameters. The choices recommended here can be used as starting points from which sensitivity analyses can be based.

It is reasonable to specify a noninformative prior for $\pi_0(\beta \mid c_0)$ since this is the initial prior for β and thus contains no information about the historical data D_0. The quantity $c_0 \geq 0$ is a scalar variance parameter which serves to control the impact of $\pi_0(\beta \mid c_0)$ on $\pi(\beta, \Delta, a_0 \mid D_0)$. To make $\pi_0(\beta \mid c_0)$ noninformative, we take large values of c_0 so that $\pi_0(\beta \mid c_0)$ is flat relative to $L(\beta, \Delta \mid D_0)^{a_0}$. Small values of c_0 will let $\pi_0(\beta \mid c_0)$ dominate (2.10). Thus, c_0 is an important tuning parameter that allows us, for example, to control the impact of the marginal distribution of the data for the calculation of posterior model probabilities in model selection problems.

The actual size of c_0 used will depend on the structure of the data set and the prior parameters for a_0. From the example in Section 5, reasonable choices of c_0 are $c_0 \geq 3$. In any case, we do not recommend an automatic one time specification for c_0, but rather we emphasize that several sensitivity analyses be conducted with several values of c_0 to examine the impact of $\pi_0(\beta \mid c_0)$ on posterior quantities of interest. The matrix W_0 plays a less crucial role than c_0. We take W_0 to be a diagonal matrix consisting of the sample variances of the covariates on the diagonal. The purpose of picking W_0 in this way is to properly adjust for the different scales of the measured covariates. If the covariates are all standardized or are measured on the same scale, then we take $W_0 = I$. In any case, W_0 plays a minimal role when c_0 is large.

For $\pi_0(\Delta \mid \theta_0)$, the values of θ_0 should be chosen so that $\pi_0(\Delta \mid \theta_0)$ is flat relative to $L(\beta, \Delta \mid D_0)^{a_0}$. Here, we choose f_{0i} to be proportional to the interval width, i.e., $f_{0i} \propto s_i - s_{i-1}$, and $g_{0i} \to 0$, for $i = 1, \ldots, M$. Choosing g_{0i} small creates a noninformative gamma prior, and choosing the shape parameters to be proportional to the width of each interval is useful in allowing the variance of each δ_i to depend on the interval length. We emphasize that we always recommend doing several sensitivity analyses using various choices of $(c_0, \theta_0, \mu_0, \sigma_0^2)$. We demonstrate such sensitivity analyses in the numerical example of Section 5.

3. **Sampling from the Joint Posterior Distribution of** (β, Δ, a_0)

In this section, we describe how to sample from the joint posterior distribution of (β, Δ). Samples from the joint posterior of (β, Δ) will enable us to compute any posterior summaries involving β or Δ. For notational convenience, we denote the distribution of a random vector X by $[X]$.

The posterior density of $[\beta, \Delta, a_0 \mid D]$ for the current study is given by

$$\begin{aligned} p(\beta, \Delta, a_0 \mid D) \quad \propto \quad & \left\{ \prod_{i=1}^{M} \exp\left(-\delta_i (a_i + b_i)\right) \right\} \\ & \left\{ \prod_{i=1}^{M} \prod_{k=1}^{d_i} \left(1 - \exp\left\{-u_{ik}(\beta) T_i(\Delta)\right\}\right) \right\} \\ \times \quad & L(\beta, \Delta \mid D_0)^{a_0} \; \pi_0(\beta \mid c_0) \; \pi_0(\Delta \mid \theta_0) \; \pi(a_0 \mid \alpha_0, \lambda_0), \end{aligned} \qquad \ldots (3.1)$$

where u_{ik}, a_i and b_i are given by (2.4), and $L(\beta, \Delta \mid D_0)$ is the likelihood function given by (2.5) with D replaced by D_0. In (3.1), we use the same intervals $(s_{i-1}, s_i]$ for both $L(\beta, \Delta \mid D)$ and $L(\beta, \Delta \mid D_0)$ so that Δ has the same meaning for the historical data and the current data. To specify the s_i's, (i) we combine $\{t_i, i = 1, 2, \ldots, n\}$ and $\{y_{0i}, i = 1, 2, \ldots, n_0\}$ together, denoted by $\{t_i^*, i = 1, 2, \ldots, n + n_0\}$; (ii) the intervals $(s_{i-1}, s_i]$ are chosen to have an equal number of failures or censored observations for the combined survival or censored time t_i^*'s, i.e., s_i is chosen to be the $(i/M)th$ quantile of the t_i^*'s, where M is the total number of the intervals. We take M so that in each interval $(s_{i-1}, s_i]$, there is at least one failure or censored observation from the y_{0i}'s and at least one failure or censored observation from the t_i's.

To obtain samples from the posterior distribution $[\beta, \Delta, a_0 \mid D]$ given by (3.1), we now describe a Gibbs sampling strategy for sampling from $[\beta \mid \Delta, a_0, D]$, $[\Delta \mid \beta, a_0, D]$, and $[a_0 \mid \beta, \Delta, D]$.

To sample β from its conditional posterior distribution, we first observe that the density function of $[\beta \mid \Delta, a_0, D]$ is log-concave in each component of β. Therefore, we may directly use the algorithm of Gilks and Wild (1992) to sample from this conditional posterior distribution. To show that the density function of $[\beta \mid \Delta, a_0, D]$ is log-concave in each component of β, it suffices to show that

$$\frac{\partial^2 \log p(\beta \mid \Delta, a_0, D)}{\partial \beta_r^2} \leq 0$$

for all $r = 1, \ldots, p$, where $p(\beta \mid \Delta, a_0, D)$ is the conditional posterior density of $[\beta \mid \Delta, a_0, D]$. We note that

$$\log(p(\beta \mid \Delta, a_0, D)) = \log(L(\beta, \Delta \mid D)) + a_0 \log(L(\beta, \Delta \mid D_0)) + \log(\pi_0(\beta \mid c_0)). \qquad \ldots (3.2)$$

Since $\pi_0(\beta \mid c_0)$ is a normal density, it is clearly log-concave. If $L(\beta, \Delta \mid D)$ is log-concave, then, clearly $L(\beta, \Delta \mid D_0)^{a_0}$ will be log-concave since $a_0 \geq 0$. Thus it suffices to show that $L(\beta, \Delta \mid D)$ is log-concave. To this end, letting $A_{ik}(\beta, \Delta) = u_{ik}(\beta) T_i(\Delta)$, $B_{ik}(\beta, \Delta) = 1 - \exp\{-A_{ik}(\beta, \Delta)\}$, and $C_{ik}(\beta, \Delta) = 1 - A_{ik}(\beta, \Delta) - \exp\{-A_{ik}(\beta, \Delta)\}$, we get

$$\begin{aligned} \frac{\partial^2 \log(L(\beta, \Delta \mid D))}{\partial \beta_r^2} &= \sum_{i=1}^{M} \sum_{k=1}^{d_i} \left\{ x_{ikr}^2 \ A_{ik}(\beta, \Delta) \ B_{ik}^{-2}(\beta, \Delta) \ \exp\{-A_{ik}(\beta, \Delta)\} C_{ik}(\beta, \Delta) \right\} & \ldots (3.3) \\ &- \sum_{i=1}^{M} \sum_{j=i+1}^{M} \sum_{k=1}^{d_j} \left\{ \delta_i \ x_{jkr}^2 \exp\{x_{jk}' \beta\} (s_{j-1} - s_{i-1}) \right\} & \ldots (3.4) \\ &- \sum_{i=1}^{M} \sum_{j=i}^{M} \sum_{k=d_j+1}^{d_j+c_j} \left\{ \delta_i \ x_{jkr}^2 \exp\{x_{jk}' \beta\} (s_j - s_{i-1}) \right\}. & \ldots (3.5) \end{aligned}$$

We see that the summands in (3.4) and (3.5) are negative. Thus, to show that $\frac{\partial^2 \log(L(\beta, \Delta|D))}{\partial \beta_r^2} \leq 0$, it suffices to show that $C_{ik}(\beta, \Delta)$ in (3.3) is negative, since all of the other terms in the summand of (3.3) are positive. We see that $C_{ik}(\beta, \Delta)$ is of the form $f(x) = 1 - e^x - e^{-e^x}$. Clearly, when $x > 0$, $f(x) < 0$. For $x \in (-\infty, 0)$, we see that $f(x)$ is a monotonic decreasing function and $\lim_{x \to -\infty} f(x) = 0$. Thus $f(x) < 0$ for all $x \in R^1$, and thus $C_{ik}(\beta, \Delta) \leq 0$ for all (β, Δ). Thus $L(\beta, \Delta, D)$ is log-concave in each component of β, which implies that $p(\beta \mid \Delta, a_0, D)$ is log-concave in β.

Now, we consider sampling from $[\Delta \mid \beta, a_0, D]$. Let $p(\Delta|\beta, a_0, D)$ denote the conditional posterior density of Δ. As for $p(\beta \mid \Delta, a_0, D)$, it can be shown that $p(\Delta \mid \beta, a_0, D)$ is log-concave in each component of Δ as long as $\pi_0(\Delta \mid \theta_0)$ is log-concave in each component of Δ. Thus, we may use the adaptive rejection algorithm of Gilks and Wild (1992) to sample Δ. To ensure the log-concavity of $\pi(\Delta|\theta_0)$, we need to take $f_{0i} \geq 1$.

Finally, we briefly discuss the generation of a_0. We note that the generation of a_0 does not depend on the data D from the current study. The conditional posterior density of a_0 can be written as

$$p(a_0 \mid \beta, \Delta, D) \ \propto \ L(\beta, \Delta \mid D_0)^{a_0} \ \pi(a_0 \mid \alpha_0, \lambda_0). \qquad \ldots (3.6)$$

In general, the conditional posterior density of a_0 is not log-concave. Therefore, we use an adaptive Metropolis algorithm to sample a_0 as proposed in Chen, Ibrahim, and Yiannoutsos (1999). The algorithm proceeds as follows. Let

$$a_0 = \frac{\exp(\xi)}{1+\exp(\xi)}, \qquad \ldots (3.7)$$

then the conditional posterior distribution $[\xi \mid \beta, \Delta, D]$ is

$$p(\xi \mid \beta, \Delta, D) \;\propto\; p(a_0 \mid \beta, \Delta, D)\, \frac{\exp(\xi)}{(1+\exp(\xi))^2}, \qquad \ldots (3.8)$$

where $p(a_0 \mid \beta, \Delta, D)$ is given by (3.6) and a_0 is evaluated at $a_0 = \exp(\xi)/(1+\exp(\xi))$. Instead of directly generating a_0 from (3.6), we first generate ξ from (3.8) and then use (3.7) to obtain a_0. To generate ξ, we use a normal proposal $N(\hat{\xi}, \hat{\tau}^2_{\hat{\xi}})$, where $\hat{\xi}$ is a maximizer of the logarithm of the right side of (3.8), which can be obtained by the Nelder-Mead algorithm implemented by O'Neill (1971). Also, $\hat{\tau}^2_{\hat{\xi}}$ is minus the inverse of the second derivative of $\log p(\xi \mid \beta, \Delta, D)$ evaluated at $\xi = \hat{\xi}$, given by

$$\hat{\tau}^{-2}_{\hat{\xi}} = -\left.\frac{d^2 \log p(\xi \mid \beta, \Delta, D)}{d\xi^2}\right|_{\xi=\hat{\xi}}.$$

The algorithm to generate ξ operates as follows:

Step 1. Let ξ be the current value.

Step 2. Generate a proposal value ξ^* from $N(\hat{\xi}, \hat{\tau}^2_{\hat{\xi}})$.

Step 3. A move from ξ to ξ^* is made with probability

$$\min\left\{ \frac{p(\xi^* \mid \beta, \Delta, D)\phi\left(\frac{\xi-\hat{\xi}}{\hat{\tau}_{\hat{\xi}}}\right)}{p(\xi \mid \beta, \Delta, D)\phi\left(\frac{\xi^*-\hat{\xi}}{\hat{\tau}_{\hat{\xi}}}\right)}, 1 \right\},$$

where ϕ is the standard normal probability density function.

After we obtain ξ, we compute a_0 by using (3.7).

4. Applications to Variable Selection

One of the main applications of our priors is for variable subset selection. In the next two subsections, we develop novel computational methods for the calculation of prior and posterior model probabilities. We first establish the necessary notation. Let p denote the number of covariates for the full model and let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m = 1, 2, \ldots, \mathcal{K}$, where $\mathcal{K}$ is the dimension of $\mathcal{M}$ and model $\mathcal{K}$ denotes the full model.

Also, let $\beta^{(\mathcal{K})} = (\beta_1, \ldots, \beta_p)'$ denote the regression coefficients for the full model and let $\beta^{(m)}$ denote a $k_m \times 1$ vector of regression coefficients for model m. We write $\beta^{(\mathcal{K})} = (\beta^{(m)'}, \beta^{(-m)'})'$, where $\beta^{(-m)}$ is $\beta^{(\mathcal{K})}$ with $\beta^{(m)}$ deleted. Also, let $D_0^{(m)} = (n_0, y_0, X_0^{(m)}, \nu_0)$ denote the historical data under model m, and let $L(\beta^{(m)}, \Delta \mid D_0^{(m)})$ denote the likelihood function of $(\beta^{(m)}, \Delta)$ for model m based on the historical data $D_0^{(m)}$.

4.1 *Prior on the model space.* Let

$$p_0^*(\beta^{(m)}, \Delta | D_0^{(m)}) = L(\beta^{(m)}, \Delta \mid D_0^{(m)})\ \pi_0(\beta^{(m)} \mid d_0)\ \pi_0(\Delta \mid \kappa_0)\ , \qquad \ldots (4.2)$$

where $\pi_0(\beta^{(m)} \mid d_0)$ is the same density as that in (2.7) with c_0 replaced by d_0 and $\pi_0(\Delta \mid \kappa_0)$ is the same density as that in (2.8) with θ_0 replaced by κ_0. We propose to take the prior probability of model m as

$$\begin{aligned} p(m) &\equiv p(m \mid D_0^{(m)}) \\ &= \frac{\int\int p_0^*(\beta^{(m)}, \Delta | D_0^{(m)})\ d\beta^{(m)}\ d\Delta}{\sum_{m \in \mathcal{M}} \int\int p_0^*(\beta^{(m)}, \Delta | D_0^{(m)})\ d\beta^{(m)}\ d\Delta}\ . \end{aligned} \qquad \ldots (4.2)$$

Since Δ is viewed as a nuisance parameter, we recommend taking $\kappa_0 = \theta_0$ to simplify the elicitation scheme. The prior parameter d_0 controls the impact of $\pi_0(\beta^{(m)} \mid d_0)$ on the prior model probability $p(m)$. This choice for $p(m)$ has several nice interpretations. First, $p(m)$ in (4.2) corresponds to the posterior probability of model m based on the data $D_0^{(m)}$ using a uniform prior for the previous study. That is, $p_0(m) = 2^{-p}$ for $m \in \mathcal{M}$, where $p_0(m)$ is the prior probability of model m before observing the historical data $D_0^{(m)}$. Therefore, $p(m) \propto p(m \mid D_0^{(m)})$, and thus $p(m)$ corresponds to the usual Bayesian update of $p_0(m)$ using $D_0^{(m)}$ as the data. Second, as $d_0 \to 0$, $p(m)$ reduces to a uniform prior on the model space. Therefore, as $d_0 \to 0$, the historical data $D_0^{(m)}$ have a minimal impact in determining $p(m)$. On the other hand, with a large value of d_0, $\pi_0(\beta^{(m)} \mid d_0)$ plays a minimal role in determining $p(m)$, and in this case, the historical data play a larger role in determining $p(m)$. Thus as $d_0 \to \infty$, $p(m)$ will be regulated by the historical data. The parameter d_0 plays the same role as c_0 and thus serves as a tuning parameter to control the impact of $D_0^{(m)}$ on the prior model probability $p(m)$. It is important to note that we use a scalar parameter c_0 in constructing the prior distribution $\pi(\beta^{(m)}, \Delta, a_0 \mid D_0^{(m)})$ given in (2.10), while we use a *different* scalar parameter d_0 in determining $p(m)$. This development provides us with great flexibility in specifying the prior distribution for $\beta^{(m)}$ as well as the prior model probabilities $p(m)$.

To compute $p(m)$ in (4.2), we adopt a Monte Carlo approach similar to Ibrahim, Chen and MacEachern (1996) to estimate all of the prior model probabilities using a single Gibbs sample from the full model. The details of this procedure are given in Chen, Ibrahim, and Yiannoutsos (1999).

Following the Monte Carlo method of Chen and Shao (1997a) the prior probability of model m can be estimated by

$$\hat{p}(m) \equiv \hat{p}(m|D_0^{(m)}) = \frac{\frac{1}{N_0}\sum_{l=1}^{N_0} \frac{p_0^*(\beta_{0(l)}^{(m)}, \Delta_{0(l)}|D_0^{(m)})\ w(\beta_{0(l)}^{(-m)}|\beta_{0(l)}^{(m)}, \Delta_{0(l)})}{p_0^*(\beta_{0(l)}^{(\mathcal{K})}, \Delta_{0(l)}|D_0^{(\mathcal{K})})}}{\frac{1}{N_0}\sum_{j=1}^{\mathcal{K}}\sum_{l=1}^{N_0} \frac{p_0^*(\beta_{0(l)}^{(j)}, \Delta_{0(l)}^{(j)}|D_0^{(j)})\ w(\beta_{0(l)}^{(-j)}|\beta_{0(l)}^{(j)}, \Delta_{0(l)}^{(j)})}{p_0^*(\beta_{0(l),\Delta_{0(l)}}^{(\mathcal{K})}|D_0^{(\mathcal{K})})}}, \qquad \ldots(4.3)$$

where $p_0^*(\beta^{(m)}, \Delta|D_0^{(m)})$ is given by (4.1), $\beta_{0(l)}^{(\mathcal{K})} = ({\beta^{(m)}}'_{0(l)}, {\beta^{(-m)}}'_{0(l)})'$, $l = 1, 2, \ldots, N_0$, are samples from

$$p_0(\beta^{(\mathcal{K})}, \Delta|D_0^{(\mathcal{K})}) \ \propto\ p_0^*(\beta^{(\mathcal{K})}, \Delta|D_0^{(\mathcal{K})}), \qquad \ldots(4.4)$$

and $w(\beta^{(-m)}|\beta^{(m)}, \Delta)$ is a *completely* known conditional density whose support is contained in or equal to the support of the conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$ with respect to the full model joint prior distribution (4.4). The implementation of Markov chain Monte Carlo sampling, the justification for $\hat{p}(m)$, and the procedure to construct a good $w(\beta^{(-m)}|\beta^{(m)}, \Delta)$ are given in Chen, Ibrahim, and Yiannoutsos (1999).

There are several advantages of the above Monte Carlo procedure. Firstly, we need only one random draw from $p_0(\beta^{(\mathcal{K})}, \Delta \mid D_0^{(\mathcal{K})})$, which greatly eases the computational burden. Secondly, it is more numerically stable since we calculate ratios of the densities in (4.3). Thirdly, in (4.3), $p_0(\beta^{(\mathcal{K})}, \Delta|D_0^{(\mathcal{K})})$ plays the role of a ratio importance sampling density (see Chen and Shao 1997b) which needs to be known only up to a normalizing constant since this common constant is cancelled out in the calculation.

4.2 *Computing posterior model probabilities.* The posterior probability of model m is given by

$$p(m|D^{(m)}) = \frac{p(D^{(m)} \mid m)\ p(m)}{\sum_{m\in\mathcal{M}} p(D^{(m)} \mid m)\ p(m)}\ , \qquad \ldots(4.5)$$

where

$$p(D^{(m)}|m) = \int\int\int L(\beta^{(m)}, \Delta \mid D^{(m)})\ \pi(\beta^{(m)}, \Delta, a_0 \mid D_0^{(m)})\ d\beta^{(m)}\ d\Delta\ da_0 \qquad \ldots(4.6)$$

denotes the marginal distribution of the data $D^{(m)}$ for the current study under model m, and $p(m)$ denotes the prior probability of model m in (4.2), which is estimated by (4.3). The marginal density $p(D^{(m)} \mid m)$ is precisely the normalizing constant of the joint posterior density of $[\beta^{(m)}, \Delta, a_0 \mid D]$.

Computing the posterior model probability $p(m|D^{(m)})$ given in (4.5) requires a different Monte Carlo method other than the one for computing the prior model probability $p(m)$ given in (4.2). Let $p(\beta^{(m)}, \Delta, a_0 \mid D^{(m)})$ denote the joint posterior density of $[\beta^{(m)}, \Delta, a_0 \mid D]$, and let $\pi(\beta^{(-m)} \mid D^{(\mathcal{K})})$ and $p(\beta^{(-m)} \mid D^{(\mathcal{K})})$ denote the respective marginal prior and posterior densities of $\beta^{(-m)}$ obtained from the full model. Then it can be shown that the posterior probability of model m is given by

$$p(m \mid D^{(m)}) = \frac{\left(p(\beta^{(-m)} = 0 \mid D^{(\mathcal{K})})/\pi(\beta^{(-m)} = 0 \mid D_0^{(\mathcal{K})})\right)\ p(m)}{\sum_{j=1}^{\mathcal{K}} \left(p(\beta^{(-j)} = 0 \mid D^{(\mathcal{K})})/\pi(\beta^{(-j)} = 0 \mid D_0^{(\mathcal{K})})\right)\ p(j)}\ , \quad \ldots (4.7)$$

$m = 1, \ldots, \mathcal{K}$, where $\pi(\beta^{(-m)} = 0 \mid D_0^{(\mathcal{K})})$ and $p(\beta^{(-m)} = 0 \mid D^{(\mathcal{K})})$ are the marginal prior and posterior densities of $\beta^{(-m)}$ evaluated at $\beta^{(-m)} = 0$. A derivation of the result in (4.7) is given in Chen, Ibrahim, and Yiannoutsos (1999). In (4.7), we use (4.3) to compute the prior model probability $p(m)$, and for notational convenience we let $p(\beta^{(-\mathcal{K})} = 0|D^{(\mathcal{K})}) = \pi(\beta^{(-\mathcal{K})} = 0 \mid D_0^{(\mathcal{K})}) = 1$. Due to the complexity of the prior and posterior distributions, the analytical forms of $\pi(\beta^{(-m)} \mid D_0^{(\mathcal{K})})$ and $p(\beta^{(-m)} \mid D^{(\mathcal{K})})$ are not available. However, we can adopt the importance-weighted marginal posterior density estimation (IWMDE) method of Chen (1994) to estimate these marginal prior and posterior densities. The IWMDE is a Monte Carlo method developed by Chen (1994), which is particularly suitable for estimating marginal posterior densities when the joint posterior density is known up to a normalizing constant. The IWMDE method requires using only two respective MCMC samples from the prior and posterior distributions for the full model, making the computation of complicated posterior model probabilities feasible. It directly follows from the IWMDE that a simulation consistent estimator of $p(\beta^{(-m)} = 0 \mid D^{(\mathcal{K})})$ is given by

$$\begin{aligned}\hat{p}(\beta^{(-m)} &= 0 \mid D^{(\mathcal{K})}) = \tfrac{1}{N} \sum_{l=1}^{N} w\left(\beta_{(l)}^{(-m)} \mid \beta_{(l)}^{(m)}, \Delta_{0(l)}, a_{0(l)}\right) \\ &\qquad \frac{p(\beta_{(l)}^{(m)}, \beta^{(-m)} = 0, \Delta_{0(l)}, a_{0(l)} \mid D^{(\mathcal{K})})}{p(\beta_{(l)}^{(\mathcal{K})}, \Delta_{0(l)}, a_{0(l)} \mid D^{(\mathcal{K})})},\end{aligned} \quad \ldots (4.8)$$

where $w(\beta^{(-m)} \mid \beta^{(m)}, \Delta, a_0)$ is a completely known conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$, Δ, and a_0, $\{(\beta_{(l)}^{(\mathcal{K})}, \Delta_{0(l)}, a_{0(l)}), l = 1, 2, \ldots, N\}$ is a sample from the joint posterior distribution $p(\beta^{(\mathcal{K})}, \Delta, a_0 \mid D^{(\mathcal{K})})$ of $[\beta^{(\mathcal{K})}, \Delta, a_0|D]$. To construct a good $w(\beta^{(-m)} \mid \beta^{(m)}, \Delta, a_0)$, we can use a similar procedure that is used to construct $w(\beta^{(-m)} \mid \beta^{(m)}, \Delta)$ in (4.3) for calculating the prior model probabilities. Similarly, we can obtain $\hat{\pi}(\beta^{(-m)} = 0 \mid D_0^{(\mathcal{K})})$, an estimate of $\pi(\beta^{(-m)} = 0 \mid D_0^{(\mathcal{K})})$, using a sample from the joint prior distribution $\pi(\beta^{(\mathcal{K})}, \Delta, a_0 \mid D_0^{(\mathcal{K})})$.

5. Myeloma Data

We consider two studies in multiple myeloma. Krall, Uthoff and Harley (1975) analyzed data from a study (historical data) on multiple myeloma in which researchers treated $n_0 = 65$ patients with alkylating agents. Of those patients, 48 died during the study and 17 survived. The response variable for these data was survival time in months from diagnosis. Several covariates were measured for these data at diagnosis. These are blood urea nitrogen, hemoglobin, platelet count, age, white blood cell count, fractures, percentage of the plasma cells in bone marrow, proteinuria, and serum calcium. The data from this study serve as the historical data for the analysis. A few years later, another multiple myeloma study (current study) using similar alkylating agents was undertaken by the Eastern Cooperative Oncology Group (ECOG). This study, labeled, E2479, had $n = 479$ patients with the same set of covariates being measured as the historical data. The results of E2479 are available in Kalish (1992). The two studies have similar patient populations, and thus serve as good examples in which to apply the proposed methodology. Here, y_0 consists of the 65 survival times from the historical study and $X_0^{(m)}$ is an $n_0 \times k_m$ matrix of covariates, where k_m denotes the number of covariates under model m.

Our main goal in this example is to illustrate the proposed prior elicitation and variable selection techniques. We also examine the sensitivity of the posterior probabilities to the choices of (μ_0, σ_0^2), c_0, and d_0. Initially, there were nine covariates for both studies. However, one covariate, proteinuria, had nearly 50% missing values and therefore was deleted from the analysis. Thus, a total of $n = 339$ observations were available from the current study, with 8 observations being right censored. Our analysis used $p = 8$ covariates. These are blood urea nitrogen (x_1), hemoglobin (x_2), platelet count (x_3) (1 if normal, 0 if abnormal), age (x_4), white blood cell count (x_5), fractures (x_6), percentage of the plasma cells in bone marrow (x_7), and serum calcium (x_8). To ease the computational burden, we standardized all of the variables. In fact, the standardization helped the numerical stability in the implementation of the adaptive rejection algorithm (Gilks and Wild, 1992) for sampling the regression coefficients from the posterior distribution.

We conduct sensitivity analyses with respect to i) c_0, ii) d_0, and iii) (μ_0, σ_0^2). These are shown in Tables 1, 2 , and 3, respectively. To compute the prior and posterior model probabilities, 50,000 Gibbs iterations were used to get convergence. We use $M = 28$, with the intervals chosen so that with the combined data sets from the historical and current data, at least one failure or censored observation falls in each interval. This technique for choosing M is reasonable and preserves the consistency in the interpretation of Δ for the two studies. In addition, we take $\theta_0 = \kappa_0$ and use $f_{0i} = s_i - s_{i-1}$ if $s_i - s_{i-1} \geq 1$ and $f_{0i} = 1.1$ if $s_i - s_{i-1} < 1$, and $g_{0i} = .001$. For the last interval, we take $g_{0i} = 10$ for $i = M$ since very little information in the data is available for this last interval. The above choices of f_{0i} and g_{0i} ensure the log-concavity of $\pi_0(\Delta \mid \theta_0)$, as this is required in sampling Δ from its conditional prior and posterior distributions (see Section 3). A stepwise variable selection procedure in SAS for the current study yields $(x_2, x_3, x_4, x_7, x_8)$ as the top model.

Table 1: POSTERIOR MODEL PROBABILITIES FOR $(\mu_0, \sigma_0^2) = (.5, .004)$, $d_0 = 3$ AND VARIOUS CHOICES OF c_0

c_0	m	$p(m)$	$p(D \mid m)$	$p(m\|D)$
3	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.015	.436	.769
10	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.015	.310	.679
30	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.015	.275	.657

Table 1 gives the model with the largest posterior probability using $(\mu_0, \sigma_0^2) = (.5, .004)$, (i.e., $\delta_0 = \lambda_0 = 30$) for several values of c_0. For each value of c_0 in Table 1, the model $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$, obtains the largest posterior probability, and thus model choice is not sensitive to these values. In addition, for $d_0 = 3$ and for any $c_0 \geq 3$, the $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ model obtains the largest posterior probability. Although not shown in Table 1, values of $c_0 < 3$ do not yield $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ as the top model. Thus, model choice may become sensitive to the choice of c_0 when $c_0 < 3$.

Table 2: THE POSTERIOR MODEL PROBABILITIES FOR $(\mu_0, \sigma_0^2) = (.5, .004)$, $c_0 = 3$ AND VARIOUS CHOICES OF d_0

d_0	m	$p(m)$	$p(D \mid m)$	$p(m\|D)$
5	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.011	.436	.750
10	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.005	.436	.694
30	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.001	.436	.540

From Table 2, we see how the prior model probability is affected as d_0 is changed. In each case, the true model obtains the largest posterior probability. Under the settings of Table 2, the $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ model obtains the largest prior probability when $d_0 \geq 3$. With values of $d_0 < 3$, however, model choice may be sensitive to the choice of d_0. For example, when $d_0 = .0001$

and $c_0 = 10$, the top model is $(x_1, x_2, x_4, x_5, x_7, x_8)$ with posterior probability of .42 and the second best model is $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ with posterior probability of .31. Finally, we mention that as both c_0 and d_0 become large, the $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ model obtains the largest posterior model probability. In addition, Tables 1 and 2 indicate a monotonic decrease in the posterior probability of model $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ as c_0 and d_0 are increased. This indicates that there is a moderate impact of the historical data on model choice.

Table 3: THE POSTERIOR MODEL PROBABILITIES FOR $c_0 = 10$, $d_0 = 10$ AND VARIOUS CHOICES OF (μ_0, σ_0^2)

(μ_0, σ_0^2)	m	$p(m)$	$p(D \mid m)$	$p(m\|D)$
$(.5, .008)$	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.005	.274	.504
$(.5, .004)$	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.005	.310	.558
$(.98, 3.7 \times 10^{-4})$	$(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$	.005	.321	.572

Table 3 shows a sensitivity analysis with respect to (μ_0, σ_0^2). Under these settings, model choice is not sensitive to the choice of (μ_0, σ_0^2). We see that in each case, $(x_1, x_2, x_3, x_4, x_5, x_7, x_8)$ obtains the largest posterior probability. In addition, there is a monotonic increase in the posterior model probability as more weight is given to the historical data.

Although not shown here, several different partitioning schemes for the choices of the intervals $(s_{i-1}, s_i]$, following a procedure similar to Ibrahim, Chen, and MacEachern (1996). The prior and posterior model probabilities do not appear to be too sensitive to the choices of the intervals $(s_{i-1}, s_i]$. This is a comforting feature of our approach since it allows the investigator some flexibility in choosing these intervals. As indicated by Ibrahim, Chen, and MacEachern (1996), sensitivity analyses on (f_{0i}, g_{0i}) show that as long as the shape parameter f_{0i} remains fixed, the results are not too sensitive to changes in the scale parameter g_{0i}. However, when we fix the scale parameter and vary the shape parameter, the results are sensitive.

6. **Discussion**

We have proposed a class of informative semi-parametric priors for proportional hazards models. The priors are based on the availability of historical data and are quite useful in model selection contexts, due to their semi-automatic elicitation scheme. Computational methods have been developed and have been implemented for these priors.

Acknowledgments. The authors wish to thank the editor and a referee for several suggestions which have improved the presentation of the paper.

References

BEDRICK, E. J., CHRISTENSEN, R., AND JOHNSON, W. (1996). A new perspective on priors for generalized linear models. *Jour. Amer. Statis. Assoc.*, **91**, 1450-1460.

COX, D. R. (1972). Regression models and life tables, (with discussion). *Jour. Royal. Statis. Soc., Series B,* **34**, 187-220.

CHEN, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Jour. Amer. Statis. Assoc.,***89**, 818-824.

CHEN, M. -H., IBRAHIM, J. G., AND YIANNOUTSOS, C, (1999). Prior elicitation, variable selection, and Bayesian computation for logistic regression model, *Jour. Royal Statis. Soc., Series B,***61**, to appear.

CHEN, M.-H. AND SHAO, Q.-M. (1997a). Estimating ratios of normalizing constants for densities with different dimensions," *Statistica Sinica,***7**, 607-630.

−−−− (1997b). On Monte Carlo methods for estimating ratios of normalizing constants, *The Ann. Statis.,***25**, 1563-1594.

GELFAND, A. E., AND SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Jour. Amer. Statis. Assoc.,***85**, 398-409.

GILKS, W.R., AND WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics,* **41**, 337-348.

IBRAHIM, J. G., CHEN, M. -H., AND MACEACHERN, S. N. (1996). Bayesian variable selection for proportional hazards model, *Technical Reportz*, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute.

IBRAHIM, J.G., AND LAUD, P.W. (1994). A predictive approach to the analysis of designed experiments,. *Jour. Amer. Statis. Assoc.,***89**, 309-319.

IBRAHIM, J. G., RYAN, L. M., AND CHEN, M-H, (1998). Use of historical controls to adjust for covariates in trend tests for Binary data, *Jour. Amer. Statis. Assoc.,***93**, to appear.

KALISH, L. A. (1992). Phase III multiple myeloma: evaluation of combination chemotherapy in previously untreated patients, *Technical Report # 726E,* Department of Biostatistics, Dana-Farber Cancer Institute.

KRALL, J. M., UTHOFF, V. A., AND HARLEY, J. B. (1975). A step-up procedure for selecting variables associated with survival, *Biometrics,* **31**, 49-57.

LAUD, P. W., AND IBRAHIM, J. G. (1995). Predictive model selection, *Jour. Royal Statis. Soc.,***Series B, 57**, 247-262.

O'NEILL, R. (1971). Algorithm AS47-Function minimization using a simplex procedure, *Applied Statistics,***20**, 338-345.

SINHA, D. AND DEY, D. K. (1997). Semiparametric Bayesian analysis of survival data, *Jour. Amer. Statis. Assoc.,***92**, 1195-1212.

JOSEPH G. IBRAHIM
DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
AND DANA-FARBER CANCER INSTITUTE
44 BINNEY ST., BOSTON
MA 02115
U.S.A.
e-mail : ibrahim@jimmy.harvard.edu

MING-HUI CHEN
DEPARTMENT OF MATHEMATICAL SCIENCES
WORCESTER POLYTECHNIC INSTITUTE
100 INSTITUTE ROAD
WORCESTER, MA 01609
U.S.A.
e-mail : mhchen@wpi.edu