

Bayesian Inferences in the Cox Model for Order Restricted Hypotheses

David B. Dunson^{1,*} and Amy H. Herring²

¹Biostatistics Branch,
National Institute of Environmental Health Sciences
MD A3-03, P.O. Box 12233
Research Triangle Park, NC 27709, U.S.A.

² Department of Biostatistics
The University of North Carolina
Chapel Hill, NC

* dunson1@niehs.nih.gov

SUMMARY. In studying the relationship between an ordered categorical predictor and an event time, it is standard practice to include dichotomous indicators of the different levels of the predictor in a Cox model. One can then use a multiple degree of freedom score or partial likelihood ratio test for hypothesis testing. Often, interest focuses on comparing the null hypothesis of no difference to an order restricted alternative, such as a monotone increase across levels of a predictor. This article proposes a Bayesian approach for addressing hypotheses of this type. We reparameterize the Cox model in terms of a cumulative product of parameters having conjugate prior densities, consisting of mixtures of point masses at one and truncated gamma densities. Due to the structure of the model, posterior computation can proceed via a simple and efficient Gibbs sampling algorithm. Posterior probabilities for the global null hypothesis and sub-hypotheses comparing the hazards for specific groups can be calculated directly from the output of a single Gibbs chain. The approach allows for level sets across which a predictor has no effect. Generalizations to multiple predictors are described, and the method is applied to a study of emergency medical treatment for stroke.

KEY WORDS: Categorical covariates; Gibbs sampler; Isotonic Regression; Monotonicity; Multiple comparisons; Proportional hazards; Survival analysis.

1. Introduction

In studies collecting event time data, researchers often have prior knowledge regarding the direction of the effect for certain predictors. For example, in studying time to critical neurological assessment for patients with stroke-like symptoms admitted to the emergency room, a possible predictor is the number of major stroke symptoms reported, ranging from 0 to 4 (Evenson, 2001; Schroeder, 2000). Certainly, it is reasonable to assume that the time to assessment does not lengthen with increasing numbers of symptoms, an assumption potentially leading to improvements in efficiency. Motivated by this application, we consider methods for order restricted inference in survival analysis problems involving multiple predictors.

Various tests have been proposed for comparing the null hypothesis of homogeneity in survival distributions for different groups to a stochastically ordered alternative. Under the Cox proportional hazards model (Cox, 1972), Sen (1984) proposed a score test which is practically equivalent to the test of Silvapulle and Silvapulle (1995), Silvapulle (1994) proposed a Wald type test, and Singh and Wright (1996; 1998) considered score and pseudo-likelihood ratio tests. These procedures focus on testing and, in general, do not produce estimates of the regression coefficients under the restriction. Although restricted maximum partial likelihood estimates can potentially be produced using constrained optimization software (Silvapulle, 1994), such software can be unreliable and standard errors and confidence intervals are not available.

Ideally, a single methodology could be used to test overall homogeneity and differences in specific groups, while also producing constrained point and interval estimates of the parameters as well as functions of the parameters such as survival curves. Motivated by the difficulty of addressing these issues simultaneously using classical methods, we propose a Bayesian approach, which places order restrictions on the parameters by choosing a prior density with constrained support.

In previous research, Gelfand, Smith and Lee (1992) proposed a Gibbs sampler for posterior computation in certain constrained parameter problems. For nonparametric Bayesian estimation of two survival curves under stochastic ordering, Arjas and Gasbarra (1996) proposed a coupled Metropolis-Hastings algorithm, and Gelfand and Kottas (2001) developed an alternative prior specification and computational algorithm. These approaches have focused on constrained estimation and cannot be used directly for hypothesis testing, since zero prior probability is assigned to the null hypothesis of no ordering.

Focusing on the Cox model, we express the regression function characterizing the change in the hazard across levels of a categorical covariate as a cumulative product of hazard ratio parameters. These parameters are assigned conditionally-conjugate prior distributions,

consisting of mixtures of point masses at one and truncated gamma densities. By choosing the support to be $(0, 1]$ or $[1, \infty)$ one can ensure non-increasing or non-decreasing hazards, respectively. Hypothesis testing, parameter estimation under the monotonicity constraint, and estimation of threshold effects can be implemented using the output of a single Gibbs sampling chain.

A related prior formulation and computational algorithm was proposed by Geweke (1996) for variable selection in linear regression, though he did not consider order constraints. By utilizing blocking, our Gibbs algorithm avoids the computational pitfall associated with mixed discrete and continuous priors highlighted by George and McCulloch (1993). For references on alternative Bayesian approaches for variable selection in survival analysis, in the absence of parameter constraints, refer to Ibrahim and Chen (2000), Ibrahim, Chen and MacEachern (1999), Volinsky and Raftery (2000), and Sinha, Chen and Ghosh (1999).

Section 2 proposes the model and prior specification. Section 3 outlines the approach for posterior computation. Section 4 applies the method to the stroke data set, and Section 5 discusses the results. Conjugacy results are included in an Appendix.

2. Order Restricted Cox Model

2.1 Proportional hazards regression with an ordinal predictor

Let $\lambda(t; \mathbf{w}_i)$ denote the hazard at time t conditional on predictors \mathbf{w}_i . We focus initially on the case where there is a single k level ordered categorical predictor, $w_i \in \{1, \dots, k\}$, and interest focuses on comparing the null hypothesis:

$$H_0 : \lambda(t; w_i = 1) = \lambda(t; w_i = 2) = \dots = \lambda(t; w_i = k) \quad \text{for all } t \in \mathfrak{R}^+, \quad (1)$$

either to the alternative hypothesis of non-decreasing hazards:

$$H_1^+ : \lambda(t; w_i = 1) \leq \lambda(t; w_i = 2) \leq \dots \leq \lambda(t; w_i = k), \quad \text{with at least one strict inequality,} \quad (2)$$

or to the alternative hypothesis of non-increasing hazards:

$$H_1^- : \lambda(t; w_i = 1) \geq \lambda(t; w_i = 2) \geq \dots \geq \lambda(t; w_i = k), \text{ with at least one strict inequality.} \quad (3)$$

Under a Cox proportional hazards regression model, we have

$$\lambda(t; w_i = l) = \lambda_0(t) \exp(\beta_{l-1}), \quad \text{for } l = 1, \dots, k, \quad (4)$$

where $\lambda_0(t)$ is the unknown baseline hazard function and $\beta_0 = 0$. Letting $\beta_l = \sum_{h=1}^l \log \gamma_h$ for $l = 1, \dots, k-1$, where $\gamma_h = \lambda(t; w_i = h+1)/\lambda(t; w_i = h)$ is a hazard ratio, we have

$$\lambda(t; w_i = l) = \lambda_0(t) \prod_{h=1}^{l-1} \gamma_h, \quad \text{for } l = 1, \dots, k. \quad (5)$$

It is straightforward to show that hypotheses H_0 , H_1^+ and H_1^- can be expressed as:

$$H_0 : \gamma_h = 1, \quad H_1^+ : \gamma_h \geq 1, \quad \text{and} \quad H_1^- : 0 < \gamma_h \leq 1, \quad (6)$$

for $h = 1, \dots, k-1$, with at least one strict inequality needed for the alternative hypotheses to hold. In the next subsection, we propose a prior density for $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{k-1})'$, which allocates probability to the null and alternative hypotheses and has a convenient conjugate structure which simplifies computation.

2.2 Priors for non-decreasing and non-increasing hazards

For the case in which the hazard is assumed to be non-decreasing with increases in an ordinal predictor, we propose the following prior density for $\boldsymbol{\gamma}$:

$$\pi(\boldsymbol{\gamma}) = \prod_{h=1}^{k-1} \text{I}_1\text{-}\mathcal{G}_{[1,\infty)}(\gamma_h; \pi_{0h}, a_h, b_h), \quad (7)$$

where $\text{I}_1\text{-}\mathcal{G}_{[1,\infty)}(\cdot; \pi, a, b)$ denotes the density consisting of a mixture of a point mass at one (with probability π) and a $\mathcal{G}(a, b)$ (gamma) density truncated below by one,

$$\text{I}_1\text{-}\mathcal{G}_{[1,\infty)}(z; \pi, a, b) = 1(z = 1)\pi + 1(z > 1)(1 - \pi) \frac{\mathcal{G}(z; a, b)}{\int_1^\infty \mathcal{G}(u; a, b) du}.$$

Prior (7) allocates probability $\pi_0 = \prod_{h=1}^{k-1} \pi_{0h}$ to the null hypothesis H_0 and probability π_{0h} to the sub-hypothesis $H_{0h} : \lambda(t; w_i = h) = \lambda(t; w_i = h + 1)$. In addition, the alternative hypothesis H_1^+ is assigned prior probability $1 - \pi_0$, and the sub-hypothesis $H_{1h}^+ : \lambda(t; w_i = h) < \lambda(t; w_i = h + 1)$ is assigned prior probability $1 - \pi_{0h}$. Density (7) has support on the space of non-decreasing hazard functions, but does allow flat regions over which increases in the predictor have no impact on the hazard. In the non-increasing case, truncate the $\mathcal{G}(a, b)$ component density above by one instead of below by one, and replace $I_1\text{-}\mathcal{G}_{[1, \infty)}(z; \pi, a, b)$ with $I_1\text{-}\mathcal{G}_{(0, 1]}(z; \pi, a, b)$.

In addition to computational advantages to be discussed in Section 3, prior (7) is conceptually appealing in that it allows for thresholds in which a predictor has no effect at lower levels, and can be used to adjust for multiple comparisons in considering sub-hypotheses of ordering between specific groups (e.g., low dose group relative to control). For example, one could set $\pi_0 = 0.5$ to assign equal prior probability to H_0 and H_1^+ , and then let $\pi_{0h} = \pi_0^{1/(k-1)}$. It follows that the prior probability of H_{0h} , the hypothesis corresponding to no change in the hazard attributable to w_i increasing from h to $h + 1$, will increase as the number of categories (and hence the number of comparisons) increases. In this way, the prior will automatically adjust for multiple comparisons in a similar manner to the approach considered by Westfall et al. (1997).

2.3 Generalization to multiple predictors

Under *a priori* independence assumptions, it is straightforward to generalize the approach to accommodate multiple predictors having a variety of restrictions on their regression coefficients. In particular, to generalize expression (5) to accommodate a vector of categorical predictors, $\mathbf{w}_i = (w_{i1}, \dots, w_{iq})'$, we let

$$\lambda(t; \mathbf{w}_i) = \lambda_0(t) \prod_{h=1}^q \prod_{l=1}^{w_{ih}-1} \gamma_{hl}, \quad (8)$$

where γ_{hl} is the multiplicative change in the hazard attributable to increasing w_{ih} from l to

$l + 1$, k_h is the number of categories of w_{ih} , $\boldsymbol{\gamma}_h = (\gamma_{h1}, \dots, \gamma_{h,k_h-1})'$ for $h = 1, \dots, q$, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \dots, \boldsymbol{\gamma}'_q)'$. To specify a prior for $\boldsymbol{\gamma}$, we let $\pi(\boldsymbol{\gamma}) = \prod_h \prod_l \pi(\gamma_{hl})$, where the $\pi(\gamma_{hl})$ are chosen to be one-inflated truncated gamma densities, as described in subsection 2.2. To avoid restricting the regression coefficients for a given predictor, simply choose the support of the corresponding prior densities to be \mathfrak{R}^+ instead of $(0, 1]$ or $[1, \infty)$.

To further generalize the procedure to accommodate a vector of continuous predictors, $\mathbf{z}_i = (z_{i1}, \dots, z_{ir})'$, we let

$$\lambda(t; \mathbf{w}_i, \mathbf{z}_i) = \lambda_0(t) \exp(\mathbf{z}'_i \boldsymbol{\alpha}) \prod_{h=1}^q \prod_{l=1}^{w_{ih}-1} \gamma_{hl} = \lambda_0(t) \exp\left(\mathbf{z}'_i \boldsymbol{\alpha} + \sum_{h=1}^q \sum_{l=1}^{w_{ih}-1} \log \gamma_{hl}\right), \quad (9)$$

where $\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\gamma})$ and $\pi(\boldsymbol{\alpha})$ is a $N(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}\boldsymbol{\alpha})$ density (as is the standard choice in Bayesian analyses of the Cox model, cf., Ibrahim, Chen and Sinha, 2001). Potentially, one could also place constraints on the regression coefficients, $\boldsymbol{\alpha}$, for the continuous predictors, but we do not consider that case here. Note that this structure assumes linearity of the effects of the continuous predictors on the multiplicative scale. Potentially, one could relax this linearity assumption by categorizing a predictor and then using a restricted prior for isotonic regression. Optimal or adaptive choice of categories is an open problem.

3. Posterior Computation and Inference

3.1 Counting Process Likelihood

For subject i ($i = 1, \dots, n$), let $N_i(t) = 1$ if the event occurs in $[0, t]$ and $N_i(t) = 0$ otherwise, and let $Y_i(t) = 1$ if the subject is at risk at time t and $Y_i(t) = 0$ otherwise. Focusing initially on the model with a single categorical predictor, the hazard at time t for subject i is

$$\lambda_i(t) = Y_i(t) \lambda_0(t) \prod_{h=1}^{w_i-1} \gamma_h, \quad (10)$$

which is the intensity process for $N_i(t)$. The counting process likelihood is proportional to

$$\prod_{i=1}^n \left\{ \prod_{t \geq 0} \lambda_i(t)^{dN_i(t)} \right\} \exp \left\{ - \int_{t \geq 0} \lambda_i(t) dt \right\}, \quad (11)$$

with $dN_i(t)$ denoting the increment of N_i over the infinitesimal interval $[t, t + dt)$. Since (11) follows a Poisson form, we can equivalently express the likelihood as a product of independent Poisson sampling densities, $dN_i(t) \sim \text{Poisson}(\lambda_i(t)dt)$, for $i = 1, \dots, n$. For background on the counting process formulation, refer to Andersen and Gill (1982) and Clayton (1991).

Following Clayton (1994), we express the counting process likelihood as

$$\prod_{i=1}^n \prod_{j=1}^J \text{Poisson}\left(dN_{ij}; Y_{ij} d\Lambda_{0j} \prod_{h=1}^{w_i-1} \gamma_h\right), \quad (12)$$

where t_1, \dots, t_J denote the unique failure times observed for a set of data, $Y_{ij} = 1$ if individual i is at risk at t_j and $Y_{ij} = 0$ otherwise, $dN_{ij} = 1$ if individual i fails at t_j and $dN_{ij} = 0$ otherwise, and $d\Lambda_{0j}$ is an increment on the cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. We complete a Bayesian specification of the model with a gamma process prior for the cumulative baseline hazard function (Kalbfleisch, 1978), $\Lambda_0(t) \sim \mathcal{GP}(R(t)c, c)$. This \mathcal{GP} prior implies the following prior for $d\mathbf{\Lambda}_0 = (d\Lambda_{01}, \dots, d\Lambda_{0J})'$:

$$\pi(\mathbf{\Lambda}_0) = \prod_{j=1}^J \mathcal{G}(d\Lambda_{0j}; R_j c, c), \quad (13)$$

where $R(t) = \int_0^t r(u) du$ is the prior guess for $\Lambda_0(t)$, $R_j = \int_{t_{j-1}}^{t_j} r(u) du$, and c controls the prior precision.

3.2 Conditional Posterior Distributions and Gibbs Sampling

Focusing initially on the non-decreasing hazards case, we can derive the conditional posterior densities for the elements of $\boldsymbol{\gamma}$ and $\mathbf{\Lambda}_0$ from likelihood (12) and the priors shown in expressions (7) and (13) by following standard algebraic routes. In particular, multiplying the prior density by the likelihood and factoring out terms not involving $d\Lambda_{0j}$, we have

$$\begin{aligned} \pi(d\Lambda_{0j} | d\mathbf{\Lambda}_{0(-j)}, \boldsymbol{\gamma}, \text{data}) &\propto d\Lambda_{0j}^{R_j c + \sum_i dN_{ij} - 1} \exp\left\{-d\Lambda_{0j}\left(c + \sum_{i=1}^n Y_{ij} \prod_{h=1}^{w_i-1} \gamma_h\right)\right\} \\ &= \mathcal{G}\left(d\Lambda_{0j}; R_j c + \sum_{i=1}^n dN_{ij}, c + \sum_{i=1}^n Y_{ij} \prod_{h=1}^{w_i-1} \gamma_h\right), \end{aligned} \quad (14)$$

where $d\mathbf{\Lambda}_{0(-j)}$ denotes the vector formed by excluding the j th element of $d\mathbf{\Lambda}_0$. In addition, following a similar approach for γ_h , we show in the Appendix that

$$\pi(\gamma_h | \boldsymbol{\gamma}_{(-j)}, d\mathbf{\Lambda}_0, \text{data}) = \text{I}_1\text{-}\mathcal{G}_{[1,\infty)}(\gamma_h; \tilde{\pi}_h, \tilde{a}_h, \tilde{b}_h), \quad (15)$$

where the conditional posterior probability of $\gamma_h = 1$ is

$$\tilde{\pi}_h = \frac{\pi_{0h} \exp \left\{ - \sum_i \sum_j 1(h < w_i) Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l \right\}}{\pi_{0h} \exp \left\{ - \sum_i \sum_j 1(h < w_i) Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l \right\} + (1 - \pi_{0h}) \frac{C(a_h, b_h)}{C(\tilde{a}_h, \tilde{b}_h)} \frac{1-F(1; \tilde{a}_h, \tilde{b}_h)}{1-F(1; a_h, b_h)}}, \quad (16)$$

where $C(a, b) = b^a / \Gamma(a)$, $F(\cdot; a, b)$ is the $\mathcal{G}(a, b)$ c.d.f.,

$$\tilde{a}_h = a_h + \sum_{i=1}^n \sum_{j=1}^J 1(h < w_i) dN_{ij} \quad \text{and} \quad \tilde{b}_h = b_h + \sum_{i=1}^n \sum_{j=1}^J 1(h < w_i) Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l. \quad (17)$$

Since the conditional posterior density (15) follows the same form as prior density (7), we have a conditionally conjugate structure. We can sample directly from (15) by setting $\gamma_h = 1$ with probability $\tilde{\pi}_h$ and otherwise sampling γ_h from $\mathcal{G}(\tilde{a}_h, \tilde{b}_h)$ truncated on the left by one. One can sample from the truncated gamma density by using the inverse c.d.f. method.

Samples from the joint posterior density of $\boldsymbol{\gamma}$ and $\mathbf{\Lambda}_0$ can be obtained using a Gibbs sampling algorithm, which alternately samples from (14) and (15) for a large number of iterations. Under mild regularity conditions, these samples will converge to a target distribution that is the joint posterior. This algorithm is easy to program, involving only simple calculation and sampling steps. In addition, in examples we have considered, the algorithm is efficient, having rapid convergence and low autocorrelation in the samples.

In the non-increasing hazards case, we simply replace (15) with

$$\pi(\gamma_h | \boldsymbol{\gamma}_{(-j)}, d\mathbf{\Lambda}_0, \text{data}) = \text{I}_1\text{-}\mathcal{G}_{(0,1)}(\gamma_h; \tilde{\pi}_h, \tilde{a}_h, \tilde{b}_h), \quad (18)$$

where the conditional posterior probability of $\gamma_h = 1$ is now

$$\tilde{\pi}_h = \frac{\pi_{0h} \exp \left\{ - \sum_i \sum_j 1(h < w_i) Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l \right\}}{\pi_{0h} \exp \left\{ - \sum_i \sum_j 1(h < w_i) Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l \right\} + (1 - \pi_{0h}) \frac{C(a_h, b_h)}{C(\tilde{a}_h, \tilde{b}_h)} \frac{F(1; \tilde{a}_h, \tilde{b}_h)}{F(1; a_h, b_h)}},$$

and the other parameters are defined as before. We can generalize the procedure to accommodate multiple covariates by (i) multiplying all terms in (14) - (18) that involve a product of γ 's by a term for the other covariate effects; and (ii) including steps in the Gibbs sampler for sampling the additional regression coefficients from their conjugate full conditional distributions. For categorical covariates, these conditionals will be conjugate, whether or not constraints are included, and for continuous covariates, adaptive rejection or Metropolis-Hastings steps can be used.

3.3 Posterior Probabilities and Hypothesis Testing

We can formally compare H_0 and H_1^+ (or H_1^-) using posterior probabilities, which can be calculated directly from the Gibbs sampling output. In particular, following a similar approach to that proposed by Carlin and Chib (1995), we can estimate the posterior probability of the global null hypothesis H_0 as follows:

$$\hat{\pi} = \frac{1}{S} \sum_{s=1}^S 1(\gamma_1^{(s)} = \gamma_2^{(s)} = \dots = \gamma_{k-1}^{(s)} = 1), \quad (19)$$

where S is the number of Gibbs iterates collected after apparent convergence, and $\gamma_h^{(s)}$ is the value of γ_h at iteration s , for $s = 1, \dots, S$. Following the common convention of using the posterior probability of H_0 as a Bayesian alternative to the p-value, one could conclude statistical significance if $\hat{\pi} < \alpha = 0.05$.

A major advantage of our approach, compared with the available frequentist score tests for testing H_0 , is that we can calculate posterior probabilities for comparing specific groups directly from the output of the same analysis used for testing the global null hypothesis. For example, we may be interested in comparing $H_{0h} : \gamma_h = 1$ to $H_{1h}^+ : \gamma_h > 1$ to judge the weight of evidence of an increase in the hazard for individuals with $w_i = h + 1$ compared to those with $w_i = h$. The posterior probability of H_{0h} can be estimated by the Rao-Blackwellized estimator $\hat{\pi}_h = \sum_{s=1}^S \tilde{\pi}_h^{(s)} / S$. This estimator makes more efficient use of the Gibbs iterates than the alternative approach of averaging model indicators sampled at each step.

4. Application

4.1 Data and Background

Treatment for acute stroke includes thrombolytic therapy, which can potentially improve neurological functioning for ischemic stroke patients if administered soon after symptom onset (within 3 hours) (NINDS, 1995; Marler et al., 1997; Barinaga, 1996; Evenson, 2001). Since treating patients quickly is critically important for their long term prognosis, minimizing the times from symptom onset to emergency room (ED) arrival, from ED arrival to diagnosis, and from diagnosis to treatment is of paramount concern.

Our interest focuses on factors predictive of the time of critical neurological assessment following admission to the ED for $n = 335$ patients with mild to moderate motor impairment. We hypothesize that an important predictor of the time to neurologic assessment is severity of clinical presentation, which is measured as a count of reported major stroke symptoms including headache, loss of motor skills or weakness, trouble talking or understanding, and vision problems. The goal of our analysis is to perform inferences on the impact of clinical presentation, gender, and race on time to neurological assessment, incorporating stochastic ordering constraints on the survival distributions across categories of clinical presentation.

4.2 Model and Prior Specification

We assume the following proportional hazards model for the rate of neurological assessment:

$$\lambda(t; \mathbf{w}_i) = \lambda_0(t) \gamma_1^{w_{i1}} \gamma_2^{w_{i2}} \gamma_3^{w_{i3}} \prod_{h=1}^{w_{i4}-1} \gamma_{4h}, \quad (20)$$

where $\mathbf{w}_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4})'$, w_{i1}, w_{i2}, w_{i3} are binary indicators of male gender, African American ethnicity, and Hispanic ethnicity, respectively, and $w_{i4} \in \{1, \dots, 5\}$ is the clinical presentation for individual i (expressed as the number of major symptoms + one). The reference group is white race, female gender with no major reported symptoms. The proportional hazards assumption was considered to be a reasonable approximation based on standard diagnostic plots.

We complete a Bayesian specification of the model with prior distributions for $d\Lambda_0$ and $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma'_4)'$. We will consider priors following the structure proposed in Sections 2 and 3, with a variety of strategies used to choose the hyperparameters including (i) diffuse but proper priors; (ii) subjectively-chosen informative priors; and (iii) hyperpriors that allow uncertainty in key hyperparameters.

Under strategy (i), we chose $R_j = r(t_j - t_{j-1})$, with $r = 0.1$ and $c = 0.0001$ to express ignorance about the baseline hazard function. In addition, we chose $\mathcal{G}(0.01, 0.01)$ priors for γ_1, γ_2 and γ_3 and $I_1\text{-}\mathcal{G}_{[1, \infty]}(\pi_{0h}, 0.01, 0.01)$ priors for $\gamma_{41}, \gamma_{42}, \gamma_{43}$ and γ_{44} . Finally, we let $\pi_{0h} = 0.5^{1/4} = 0.841$ in order to assign equal probability to H_0 and H_1 *a priori*. We assessed the sensitivity of the inferences to the choice of a_h and b_h by repeating the analysis for $a_h = b_h = 0.001$ and $a_h = b_h = 0.1$ instead of 0.01.

Under strategy (ii), we first chose an exponential survival function with constant hazard rate $r = 0.5$ as our best guess for the baseline, resulting in $R_j = 0.5(t_j - t_{j-1})$, and expressed our uncertainty in this guess by letting $c = 1.0$. For γ_1, γ_2 , and γ_3 , the hazard ratios characterizing the effect of male gender, African American ethnicity, and Hispanic ethnicity, respectively, we chose $\mathcal{G}(0.5, 0.5)$ priors to express our belief that these parameters fall relatively close to one, with values close to zero or much larger than 2-3 considered unlikely. Finally, for the γ_{4h} parameters, we let $\pi_{0h} = 0.25$ and $a_h = b_h = 1.5$ to express our belief that there is a good chance of a moderate increase in the hazard of neurological assessment associated with a unit increase in the number of symptoms.

Following a common strategy in the literature, we accommodated uncertainty in the choice of hyperparameters by choosing hyperprior densities for the precision in the gamma process prior (c) and for the hyperparameters in the priors for $\gamma_{41}, \gamma_{42}, \gamma_{43}, \gamma_{44}$ (π_{0h}, b_h):

$$\pi(c) = \mathcal{G}(c; 1, 1), \quad \pi(\pi_{0h}) = \mathcal{B}(\pi_{0h}; 1, 3), \quad \text{and} \quad \pi(b_h) = \mathcal{G}(b_h; 3, 2),$$

where $\mathcal{B}(\cdot)$ is the beta density, $a_h = b_h$, and π_{0h}, b_h are assumed constant. To generalize the

Gibbs sampler, we included Metropolis-Hastings and Gibbs steps for updating c , π_{0h} , and b_h . It is our expectation that the data are informative about c , and to a lesser extent π_{0h} and b_h .

For purposes of comparison, we also obtained unrestricted frequentist estimates of γ by maximization of the partial likelihood, and we obtained unrestricted Bayesian estimates by using $\mathcal{G}(0.01, 0.01)$ priors for the elements of γ instead of the restricted priors described above.

4.3 Analysis and Results

The Gibbs sampler described in subsection 3.2 was used for posterior computation, with 25,000 iterations collected, and the first 1,000 discarded as a burn-in. This burn-in interval appeared more than sufficient, since plots of the parameters showed very rapid convergence to a stationary distribution, even for the analysis involving estimation of the hyperparameters. The chains also had low autocorrelation and excellent mixing, suggesting good computational efficiency.

Table 1 shows the unconstrained Bayesian and maximum partial likelihood estimates. As expected because we are using diffuse priors in this initial analysis, the Bayesian and maximum partial likelihood estimates were similar, as were the standard errors. A four degree of freedom partial likelihood ratio test of homogeneity in the hazard function across levels of clinical presentation was non-significant ($p = 0.21$), as were pairwise comparisons between different levels of clinical presentation after adjustment for multiple comparisons.

Table 2 presents posterior means and standard deviations under the proposed order constrained Bayesian approach for the different strategies of prior elicitation outlined in subsection 4.2, and Figure 1 plots estimated posterior densities for $\gamma_{41}, \gamma_{42}, \gamma_{43}, \gamma_{44}$. There were no apparent systematic differences between the constrained estimates and the unconstrained estimates presented in Table 1, which suggests that the order constraint is supported by the

data. The estimates from the informative prior analysis were quite similar to the estimates from the diffuse prior analysis, except for γ_{44} , the hazard ratio characterizing the difference between individuals with three symptoms and those with four. This is not surprising, since there were only six subjects having four symptoms, and the diffuse prior analysis incorporated a conservative adjustment for multiple comparisons, which tends to shrink estimates towards one.

Interestingly, the analysis that used hyperprior distributions to account for uncertainty in the choice of hyperparameters actually had lower posterior standard deviations. To explain this apparently counter-intuitive result, we first note that the posterior mean of c is 42, and the 95% credible interval is [33,52], values much higher than we anticipated *a priori*. It appears that the data are highly informative about c , and the more flexible gamma process mixture is a more adequate characterization of the baseline hazard function than the gamma process. By assigning higher precision to the exponential cumulative hazard prior mean adaptively based on the data, we are buying efficiency in estimation of the cumulative baseline hazard, which in turn results in (apparently) improved efficiency in estimating the covariate effects.

In the analyses under priors (i), (ii), and (iii), the estimated posterior probabilities of H_0 were 0.02, < 0.01 , and < 0.01 , respectively, which provides strong evidence in favor of the alternative hypothesis that there is an overall decrease in the time to neurological assessment as the number of major symptoms increases. Table 3 presents the posterior probabilities of an increase in the rate of neurological assessment attributable to a unit increase in the number of symptoms. The analyses were consistent in showing some evidence of a faster rate of assessment for individuals with one major symptom compared to those with none, but no evidence of an increase in the hazard as the number of symptoms increased from 1 to 3. The diffuse prior analysis, which incorporated a conservative adjustment for multiple comparisons, showed no evidence of an increase in the hazard in going from 3 to 4 symptoms,

while the other analyses showed clear evidence. This sensitivity to the prior is as expected, since the number of patients with four symptoms was small.

5. Discussion

This article proposes a general Bayesian approach for inference under proportional hazards models with monotonicity constraints on the regression functions characterizing the change in hazards across levels of a categorical covariate. Under the proposed procedures, inferences on order-constrained hypotheses and sub-hypotheses are straightforward, as are point and interval estimation under the constraint. The method is both easy to apply using Gibbs sampling and straightforward to generalize to a variety of settings. For example, we can easily generalize the procedure to accommodate frailty models for multiple event time data (Clayton, 1991).

The proposed methodology is related to methods for Bayesian variable selection in linear regression, an area in which inferences are known to be sensitive to the choice of hyperparameters. In the application section, we considered three general strategies of prior elicitation, including diffuse but proper priors, subjectively-chosen informative priors, and an approach that used hyperprior distributions to account for uncertainty in choice of hyperparameters. The proposed hyperprior approach provides a more general and robust method of inference, and we found that the data are informative about the hyperparameters. The strategy of estimating the precision in the nonparametric gamma process prior for the cumulative baseline hazard improves flexibility (and potentially efficiency) without adding substantially to computation, and hence should be useful even when order restrictions are not appropriate.

We have focused on the Cox proportional hazards model, since it is by far the most widely-used and familiar model for event time analysis. In addition, the Cox model results in many simplifications in computation and interpretation, which our approach uses to full extent. An extremely interesting and challenging area of future research is the development

of general methods for assessing stochastic ordering in survival distributions with respect to several categorical and continuous predictors without requiring proportional hazards. Methods for more general order restrictions involving unknown peaks or changepoints are also of interest.

ACKNOWLEDGEMENTS

DASH II was funded by Glaxo-Wellcome. The authors thank Kelly Evenson and Wayne Rosamond for providing the data and Dr. Evenson for providing comments on the application. In addition, we thank Zhen Chen and Gregg Dinse for their critical reading of the manuscript, and an anonymous referee for insightful comments that greatly improved the paper.

REFERENCES

- Andersen, P.K. and Gill, R.D. (1982). Cox regression-model for counting-processes - A large sample study. *Annals of Statistics* **10**, 1100-1120.
- Arjas, E. and Gasbarra, D. (1996). Bayesian inference of survival probabilities under stochastic ordering constraints. *Journal of the American Statistical Association* **91**, 1101-1109.
- Barinaga, M. (1996). Finding new drugs to treat stroke. *Science* **272**, 664-666.
- Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **57**, 473-484.
- Clayton, D.G. (1991). A Monte-Carlo method for Bayesian-inference in frailty models. *Biometrics* **47**, 467-485.
- Clayton, D.G. (1994). Bayesian analysis of frailty models. Technical report, Medical Research Council Biostatistics Unit, Cambridge.

- Cox, D.R. (1972). Regression models and life-tables (with discussions). *Journal of the Royal Statistical Society: B* **34**, 187-202.
- Evenson, K.R., Rosamond, W.D., Vallee, J.A., and Morris, D.L. (2001). Concordance of Stroke Symptom Onset Time: The Second Delay in Accessing Stroke Healthcare (DASH II) Study. *Annals of Epidemiology* **11**, 202-207.
- Gelfand, A.E. and Kottas, A. (2001). Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of Statistical Mathematics* **53**, 865-876.
- Gelfand, A.E., Smith, A.F.M. and Lee, T.M. (1992). Bayesian-analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association* **87**, 523-532.
- George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.
- Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics* **5**, J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.). Oxford University Press, 609-620.
- Kalbfleisch, J.D. (1978). Non-parametric Bayesian analysis of survival data. *Journal of the Royal Statistical Society B* **40**, 214-221.
- Ibrahim, J.G. and Chen, M.-H. (2000). Bayesian methods for variable selection in the Cox model. In *Generalized Linear Models: A Bayesian Perspective*. eds. D.K. Dey, S.K. Ghosh and B.K. Mallick. New York: Marcel Dekker, Inc.
- Ibrahim, J.G., Chen, M.-H., and MacEachern, S.N. (1999). Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics* **27**, 701-717.

- Ibrahim, J.G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.
- Marler, J., Winters Jones, P., Emr, M., editors (1997). National Institute of Neurological Disorders and Treatment of Acute Stroke. Proceedings of a National Symposium on Rapid Identification and Treatment of Acute Stroke. Washington, DC: NIH Publication No. 97-4239.
- National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995). Tissue plasminogen activator for acute ischemic stroke. *New England Journal of Medicine* **333**, 1581-1587.
- Schroeder, E.B., Rosamond, W.D., Morris, D.L., Evenson, K.R., and Hinn, A.R. (2000). Determinants of use of emergency medical services in a population with stroke symptoms: The second delay in accessing stroke healthcare (DASH II) study. *Stroke* **31**, 2591-2596.
- Sen, P.K. (1984). Subhypotheses testing against restricted alternatives for the Cox regression model. *Journal of Statistical Planning and Inference* **10**, 31-42.
- Silvapulle, M.J. (1994). On tests against one-sided hypotheses in some generalized linear models. *Biometrics* **50**, 853-858.
- Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association* **90**, 342-349.
- Singh, B. and Wright, F.T. (1996). Testing order restricted hypotheses with proportional hazards. *Lifetime Data Analysis* **2**, 363-390.
- Singh, B. and Wright, F.T. (1998). Comparing survival times for treatments with those of a control under proportional hazards. *Lifetime Data Analysis* **4**, 265-279.

Sinha, D., Chen M.-H., and Ghosh, S.K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics* **55**, 585-590.

Volinsky, C.T. and Raftery, A.E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256-262.

Westfall, P.H., Johnson, W.O. and Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84**, 419-427.

APPENDIX

Derivation of Conditional Posterior Densities and Proof of Conjugacy

When there is a single categorical covariate, w_i , expressions (5) and (7) imply

$$\begin{aligned} \pi(\boldsymbol{\gamma} \mid d\boldsymbol{\Lambda}_0, \text{data}) &\propto \left[\prod_{h=1}^{k-1} \left\{ 1(\gamma_h = 1)\pi_{0h} + 1(\gamma_h > 1)(1 - \pi_{0h}) \frac{\mathcal{G}(\gamma_h; a_h, b_h)}{1 - F(1; a_h, b_h)} \right\} \right] \\ &\times \left\{ \prod_{h=1}^{k-1} \gamma_h^{\sum_i \sum_j 1(h < w_i) dN_{ij}} \right\} \exp \left(- \sum_i \sum_j Y_{ij} d\Lambda_{0j} \prod_{h=1}^{w_i-1} \gamma_h \right), \end{aligned}$$

which is proportional to the prior density multiplied by the counting process likelihood, with F denoting the gamma cdf. Holding all other parameters fixed, we can derive the full conditional density of γ_h from this expression,

$$\begin{aligned} \pi(\gamma_h \mid \boldsymbol{\gamma}_{(-h)}, d\boldsymbol{\Lambda}_0, \text{data}) &\propto 1(\gamma_h = 1)\pi_{0h} \exp \left(- \sum_i \sum_j Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l \right) + \frac{1(\gamma_h > 1)(1 - \pi_{0h})}{1 - F(1; a_h, b_h)} \\ &\times C(a_h, b_h) \gamma_h^{a_h-1} \exp(-\gamma_h b_h) \gamma_h^{\sum_i \sum_j 1(h < w_i) Y_{ij} dN_{ij}} \exp \left(- \gamma_h \sum_i \sum_j 1(h < w_i) Y_{ij} d\Lambda_{0j} \prod_{l:l \neq h}^{w_i-1} \gamma_l \right), \end{aligned}$$

where $C(a, b)$ is the constant in the $\mathcal{G}(\cdot; a, b)$ density and the second line in this term is equivalent to $C(a_h, b_h) \mathcal{G}(\gamma_h; \tilde{a}_h, \tilde{b}_h) / C(\tilde{a}_h, \tilde{b}_h)$, where \tilde{a}_h and \tilde{b}_h are defined in expression (17).

Dividing by the normalizing constant, the conditional posterior density of γ_h is

$$\pi(\gamma_h \mid \boldsymbol{\gamma}_{(-h)}, d\boldsymbol{\Lambda}_0, \text{data}) = 1(\gamma_h = 1)\tilde{\pi}_h + 1(\gamma_h > 1)(1 - \tilde{\pi}_h) \frac{\mathcal{G}(\gamma_h; \tilde{a}_h, \tilde{b}_h)}{1 - F(1; \tilde{a}_h, \tilde{b}_h)},$$

where $\tilde{\pi}_h$ is defined in (16). This density follows the same $I_1\text{-}\mathcal{G}_{[1, \infty)}$ form as the prior density, but with updated parameters that depend on the data. Therefore, the prior is conditionally conjugate. A similar approach can be used for non-increasing regression functions. In addition, due to the multiplicative structure of the prior and likelihood, the conditional densities follow the same form in problems with multiple predictors, with only simple modifications to $\tilde{\pi}_h$, \tilde{a}_h and \tilde{b}_h .

Table 1*Estimates of the hazard ratios in the unconstrained frequentist and Bayesian analyses.*

Predictor	Frequentist Estimate		Posterior Summary	
	$\hat{\gamma}$	$\widehat{se}(\hat{\gamma})^\ddagger$	$\hat{\gamma}$	$\widehat{se}(\hat{\gamma})$
male	0.868	0.106	0.875	0.108
black	0.866	0.121	0.872	0.121
hispanic	0.736	0.273	0.751	0.276
1 symptom [†]	1.181	0.196	1.202	0.199
2 symptoms	0.998	0.157	1.005	0.150
3 symptoms	1.037	0.269	1.051	0.268
4 symptoms	2.768	1.301	2.933	1.483

[‡] Calculated using delta method.[†] Multiplicative change due to increasing the number of symptoms by one.**Table 2**

Posterior summaries of the hazard ratios in the order constrained analysis for different strategies of prior elicitation: (i) diffuse but proper with conservative adjustment for multiple comparisons; (ii) subjectively-chosen informative priors; (iii) hyperpriors.

Predictor	Prior Elicitation Strategy		
	(i) Diffuse	(ii) Informative	(iii) Hyperprior
male	0.87 [†] _(0.11)	0.84 _(0.10)	0.83 _(0.09)
black	0.87 _(0.12)	0.87 _(0.12)	0.94 _(0.13)
hispanic	0.76 _(0.29)	0.71 _(0.25)	0.57 _(0.20)
1 symptom	1.22 [‡] _(0.16)	1.15 _(0.13)	1.14 _(0.10)
2 symptoms	1.07 _(0.10)	1.08 _(0.10)	1.13 _(0.11)
3 symptoms	1.16 _(0.20)	1.17 _(0.18)	1.19 _(0.18)
4 symptoms	1.74 _(1.06)	1.92 _(0.65)	1.73 _(0.56)

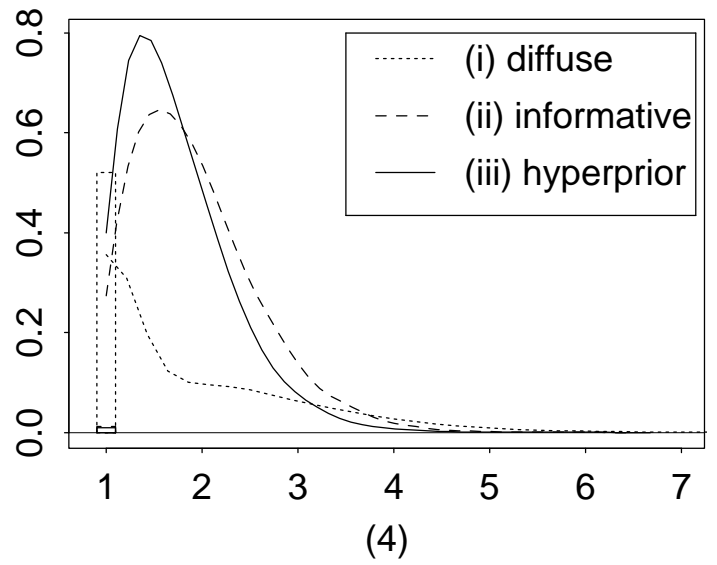
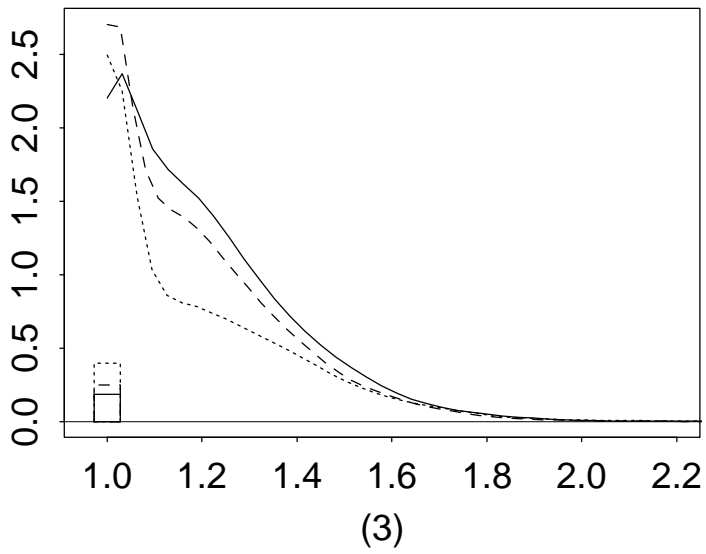
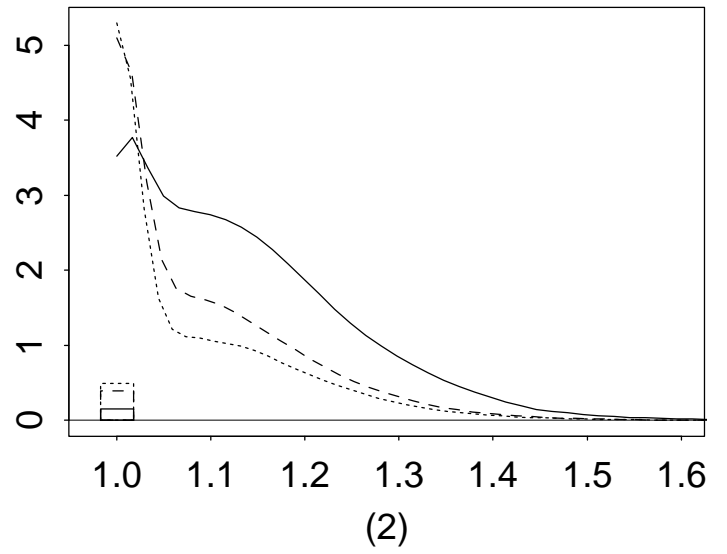
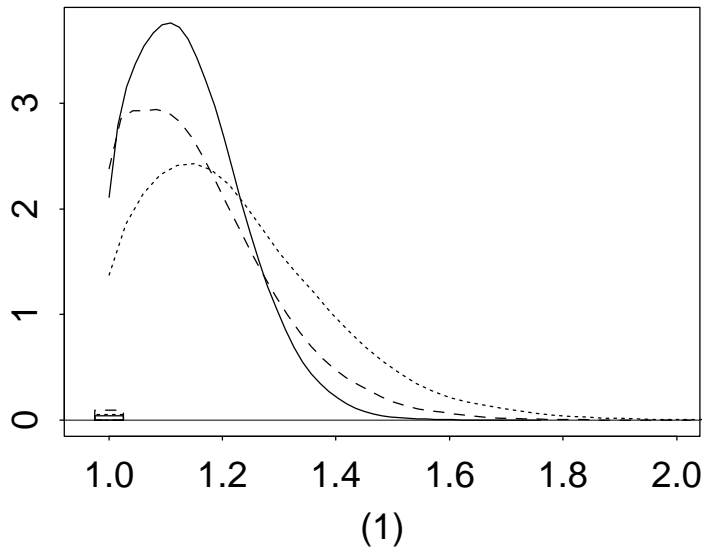
[†] Posterior mean_(sd) for the hazard ratio, γ [‡] Multiplicative change due to increasing the number of symptoms by one.

Table 3

Estimated posterior probabilities of an increase in the rate of neurological assessment attributable to a unit increase in the number of symptoms.

Increase	Prior Elicitation Strategy		
	(i) Diffuse [†]	(ii) Informative	(iii) Hyperprior
0 → 1 symptom	0.95	0.90	0.96
1 → 2 symptoms	0.51	0.61	0.75
2 → 3 symptoms	0.60	0.75	0.82
3 → 4 symptoms	0.48	0.99	0.99

[†] Inflated the prior probability of H_{0h} to account for multiple comparisons



Caption for Figure 1:

Estimated posterior densities for (1) γ_{41} , (2) γ_{42} , (3) γ_{43} , and (4) γ_{44} , the hazard ratios for unit increases in the number of symptoms, under different strategies of prior elicitation.