Statistica Sinica **15**(2005), 325-357

A BAYESIAN MULTIRESOLUTION HAZARD MODEL WITH APPLICATION TO AN AIDS REPORTING DELAY STUDY

Peter Bouman, Vanja Dukic and Xiao-Li Meng

Northwestern University, University of Chicago and Harvard University

Abstract: Thanks to advances in MCMC methodology, Bayesian curve estimation has become an increasingly popular subject both in practice and in theoretical research. Prior specification for curves is a more challenging task than for scalar or multivariate parameters. Besides using fully parametric curves, common strategies include using a stochastic process or discretizing the curve, each with its own advantages and pitfalls. In this paper we adopt the second strategy, primarily for its practicality for general users, in the context of hazard (and survival) curve estimation. We adapt a multiresolution modeling approach from the engineering literature, which provides a resolution-invariant prior for hazard increments, with their *a priori* dependence conveniently specified via tuning a few hyperparameters. We also investigate a hierarchical mixing strategy to combat a pitfall of the multiresolution approach: that nearby cells may exhibit lower dependence than cells that are far apart due to the fact that the multiresolution approach is based on a binary tree construction and not the usual Euclidean topology. Our investigations include both simulated and textbook data, as well as comparisons to the first strategy based on a Beta process prior, and to the second strategy based on a discretized correlated Gamma process prior. The paper concludes with a detailed application of the proposed method to an AIDS reporting delay estimation for New York City, from data provided by the Centers for Disease Control and Prevention (CDC).

Key words and phrases: AIDS, Bayesian multiresolution models, propotional hazard model, reporting delay, semiparametric hazard models.

1. Introduction

The literature on Bayesian hazard estimation is extensive, with a number of competing methods for placing prior structure on the hazard function $h(t)$. A simple method is to model the hazard in discrete time, placing an i.i.d. Gamma prior on the collection of discrete-time hazard increments. In many analyses, however, it may be desirable to allow *a priori* dependence among the hazard increments for "borrowing of strength" among adjacent intervals. In this paper we investigate a semiparametric discrete-time model that chooses a level of dependence in a data-driven fashion via a hierarchical Bayes model involving hyperpriors on hazard "curvature" or smoothness parameters. We adapt

the multiresolution approach that has grown popular in the recent Bayesian function estimation literature, and which is surveyed in Kolaczyk (1999). Recent examples include Nowak and Kolaczyk (2000) using multiresolution methods for Poisson intensity estimation in an astrophysical application, Huang and Cressie (2001) applying multiscale graphical models in spatial statistics, and Huang, Cressie and Gabrosek (2002) formulating dynamic spatio-temporal models. The multiresolution method invokes a Haar-based wavelet, and has been extensively applied, because of its tractability, to density estimation in a non-Bayesian context (cf., Kolaczyk and Nowak (2003), Willett and Nowak (2003) and Antoniadis, Grégoire and Nason (1999); the latter used a class of wavelet (multiresolution) methods directly for the estimation of hazard curves, and can be considered a non-Bayesian counterpart to ours. For more applications, see http://math.bu.edu/people/kolaczyk/multigranular.html.) Our model uses a multiresolution prior on the discrete-time hazard increments, yielding a "resolution-invariant" prior specification for each hazard increment that does not vary regardless of whether the increment can be further divided into finer increments. The model setup also allows for an adaptive choice of the resolution level through suitable model selection criteria.

In the context of Bayesian hazard estimation, an important competing prior for the hazard function is the nonparametric Beta process prior on the cumulative hazard function (Hjort (1990)) with Beta-distributed independent increments. Both the theoretical properties of the corresponding posterior, and the needed computation methods for sampling from it, are in general complex (e.g., Kim and Lee (2003) and Lee and Kim (2002)). For practical purposes, therefore, a simpler semiparametric approach may be preferable, especially if it delivers comparable or sometime even superior inferential performance, as we demonstrate in this paper using the standard mean squared error (MSE) criterion. We also explore the impact of prior assumptions about hazard increment dependence for a small leukemia survival dataset, and investigate the information induced by the "curvature" hyperparameter in a simulation of Weibull and Gamma hazard estimation. By mixing over a second hyperparameter, we provide a practical method of dealing with a potential pitfall of the multiresolution approach: increments that are close to each other may be given lower *a priori* correlations than those that are further apart.

The literature on Bayesian hazard estimation using correlated hazard increments is also quite vast. Notable examples are papers on correlated Gamma increments for the hazard function; principal sources include Arjas and Gasbarra (1994), who place an autoregressive Gamma prior process on the hazard increments, and Nieto-Barajas and Walker (2002), who use two latent Markov processes to induce prior correlation among the hazard increments. Among other work in this area, we mention Sinha (1993), Sinha and Dey (1997), Sahu, Dey,

Aslanidou and Sinha (1997), Aslanidou, Dey and Sinha (1998), Sinha (1998) and Craiu and Duchesne (2004), whose model is compared to ours in Section 3.

An additional feature of our semiparametric model is that it is straightforward to use Bayesian missing data imputation to adjust for a relatively wide range of censoring mechanisms. This is demonstrated by our detailed application to an AIDS reporting delay study using the CDC AIDS Public Information Dataset, which suffers from both interval-censoring and truncation. Without appropriate adjustment for such missing data, hazard estimation can be seriously biased.

The paper is organized as follows: Section 2 presents the multiresolution hazard model, some of its theoretical properties, and MCMC procedures for fitting the model. Section 3 provides empirical investigation of the impact of prior assumptions on hazard inference, as well as a comparison of the multiresolution model to two other Bayesian prior processes on the cumulative hazard. Section 4 presents the AIDS reporting delay analysis with model selection criteria, while Section 5 discusses some future work.

2. Multiresolution Model Formulation

The goal in this section is to describe a strategy for estimating population survival $S(t)$, given possibly censored failure times and covariate data for N subjects. We adopt a Bayesian proportional-hazards model that allows for general censoring in the common population survival curve $S_{\text{base}}(t)$. We choose a fixed and ordered set of time horizons t_j and seek estimates of $S_{\text{base}}(t_j)$ at the chosen t_j. Note that $S_{\text{base}}(t)$ is an underlying baseline survival curve, and that individual patients will have different survival functions depending on their covariates and trial arm membership. For modeling convenience, however, we follow the well-established approach of working with the hazard $H_{\text{base}}(t) = -\ln(S_{\text{base}}(t))$ and its discrete increments $d_j \equiv H_{\text{base}}(t_j) - H_{\text{base}}(t_{j-1})$. The estimates for these quantities can easily be transformed into survival function estimates using the fact that $d_j = \int_{t_{j-1}}^{t_j} h_{\text{base}}(s)ds$, where the function $h_{\text{base}}(s)$ is the baseline hazard rate at time s, so that $S_{\text{base}}(t_j) = e^{-d_j} S_{\text{base}}(t_{j-1})$, with $S_{\text{base}}(t_0) \equiv S_{\text{base}}(0) = 1$.

2.1. Multiresolution prior for baseline hazard increments

A standard Cox proportional-hazards model estimates covariate effects for survival, but treats the baseline hazard function as a (high-dimensional) nuisance parameter. The model here estimates, besides covariate effects, a baseline survival $S_{\text{base}}(t_j)$ at a fixed set of time points $0 < t_1 < \cdots < t_J$. The spacing between the t_j need not be equal and in fact should be chosen according to the needs of the analysis and relevant assumptions about the underlying hazard function. As for most wavelet-based methods, we set $J = 2^M$, $M > 0$, as described

below. In general, the number of bins J should be chosen as a fraction of N, the total sample size available, so that there are multiple observations per bin.

Following the standard practice of survival analysis, we will estimate $H(t_j) \equiv H_{\text{base}}(t_j)$, $j = 1, \ldots, J$, and its discrete-time increments d_j. For times t such that $t_{j-1} < t < t_j$, we use a piecewise-constant hazard rate assumption to linearly interpolate the baseline cumulative hazard at time t. For times $t > t_J$, $S(t)$ is not defined under our model, and observed failure times after t_J are regarded as right-censored at t_J. Thus, the number of bins $J = 2^M$ and bin widths $t_j - t_{j-1}$ should be chosen so that t_J is close to the maximum observed failure time $t_{\max}$, and the amount of information lost by this approximation is minimized.

Given the assumed number J of time intervals, we can parametrize the hazard increments d_j in a way that makes it easy to express our prior beliefs about the curvature of the underlying hazard curve. We start with the highest resolution level where $H_{M,i-1} \equiv d_i$ for $i = 1, \ldots, J$ are the J hazard increments we are trying to estimate. We then aggregate via the dyadic summands $H_{m-1,p} = H_{m,2p} + H_{m,2p+1}$, from $m = M$ down to $m = 0$, where m indexes the level of resolution and $p = 0, \ldots, 2^m - 1$ indexes the position within each level. Note that $H_{0,0} \equiv H(t_J)$, the total hazard at the final time horizon t_J. Moving from the top of this "binary tree", we let $R_{m,p} \equiv H_{m,2p}/H_{m-1,p}$ and parametrize the hazard increments $d_1, \ldots, d_J$ by $H_{0,0}$ and the "splits" $R_{1,0}, \ldots, R_{M,2^{M-1}-1}$ (hereafter $R_{m,p}$).

Algebraically, our parametrization can be more conveniently expressed via matrix notation $\log \boldsymbol{d} = \boldsymbol{\Pi}\tilde{\boldsymbol{R}}$, whose definition is most clearly illustrated by the following example with $M = 3$ and thus $J = 2^M = 8$:

$$\begin{pmatrix} \log(d_1) \\ \log(d_2) \\ \log(d_3) \\ \log(d_4) \\ \log(d_5) \\ \log(d_6) \\ \log(d_7) \\ \log(d_8) \end{pmatrix} = \begin{pmatrix} 1&1&0&1&0&0&0&1&0&0&0&0&0&0&0&0 \\ 1&1&0&1&0&0&0&0&1&0&0&0&0&0&0&0 \\ 1&1&0&0&1&0&0&0&0&1&0&0&0&0&0&0 \\ 1&1&0&0&1&0&0&0&0&0&1&0&0&0&0&0 \\ 1&0&1&0&0&1&0&0&0&0&0&1&0&0&0&0 \\ 1&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0 \\ 1&0&1&0&0&0&1&0&0&0&0&0&0&1&0&0 \\ 1&0&1&0&0&0&1&0&0&0&0&0&0&0&1&0 \\ 1&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1 \end{pmatrix} \begin{pmatrix} \log(H_{0,0}) \\ \log(R_{1,0}) \\ \log(1-R_{1,0}) \\ \log(R_{2,0}) \\ \log(1-R_{2,0}) \\ \log(R_{2,1}) \\ \log(1-R_{2,1}) \\ \log(R_{3,0}) \\ \log(1-R_{3,0}) \\ \log(R_{3,1}) \\ \log(1-R_{3,1}) \\ \log(R_{3,2}) \\ \log(1-R_{3,2}) \\ \log(R_{3,3}) \\ \log(1-R_{3,3}) \end{pmatrix}.$$

The matrix expression makes clear the splitting nature of the multiresolution process, with each column of 2^m unit terms splitting into two subsequent columns, each with 2^{m-1} ones ($m = 1, \ldots, M$), continuing until we reach 2^M columns with only a single unit element.

Inspired by the development in Nowak and Kolaczyk (2000), we place independent symmetric Beta priors on the R_j's and a Gamma prior on $H_{0,0}$, multiplying the shape parameter of the Beta priors at each higher level of the resolution by a hyperparameter k. For example, the priors for $H_{0,0}$ and $R_{m,p}$ given $M = 3$ are

$$\begin{aligned} H_{0,0} &\sim \mathcal{G}a(a, \lambda) \\ R_{1,0} &\sim \mathcal{B}e(ka, ka) \\ R_{2,0}, R_{2,1} &\sim i.i.d.\ \mathcal{B}e(k^2a, k^2a) \\ R_{3,0}, \ldots, R_{3,3} &\sim i.i.d.\ \mathcal{B}e(k^3a, k^3a). \end{aligned} \tag{1}$$

This formulation is one way to ensure prior self-consistency under aggregation of hazard increments, as will be explained in the next few paragraphs. As we prove in Section 2.6 when $k = 0.5$, the baseline hazard increments $\{d_1, \ldots, d_J\}$ are *a priori* uncorrelated and, in fact, independently Gamma-distributed. Adopting k lesser or greater than 0.5 yields, respectively, negative or positive prior correlation among the d_j's. Positive prior correlation corresponds to a smoothing of the baseline hazard function, in which the inference for a particular hazard increment borrows strength from those of its neighbors. This may be especially appropriate in cases of heavy censoring. As explained in the next section, we place hyperpriors on k and a to allow their posteriors to incorporate information from the complete-data likelihood, as well as a hyperprior on λ to control the inference about the total cumulative hazard $H_{0,0}$.

Since the maximal resolution level M is an artifact of our model, it is desirable that our prior specification for $H_{0,0}$ and for any increment $H_{m-1,p}$ at *fixed* depth m (from the top of the tree) does not depend on our choice of M. For $H_{0,0}$, this is trivially true because M does not enter the $\mathcal{G}a(a, \lambda)$ prior specification for $H_{0,0}$. For $H_{m-1,p}$, this is true because changing from M to $M - 1$ in our choice is identical to discarding the "bottom-level" $R_{M,p}$. As a consequence of the identity $H_{M-1,p} = H_{M,2p} + H_{M,2p+1}$, the joint prior for $H_{M-1,p}, p = 0, \ldots, 2^{M-1} - 1$, remains the same regardless of whether the total resolution depth is M or $M - 1$. By induction on M, as a generic index, we see that the joint prior for $H_{m-1,p}, p = 0, \ldots, 2^{m-1} - 1$, will only depend on the depth m from the top of the tree, not the maximal depth of the tree M. Consequently, the multiresolution prior is self-consistent, or resolution-invariant, to aggregation. This property was also observed by Kolaczyk (1999) for a closely related prior on Poisson intensities.

2.2. Log-linear complete-data proportional hazards

This subsection considers the complete-data likelihood L for failure times T_i observed between times 0 and t_J. Because we only estimate the baseline hazard function in this time range, failures observed after t_J will be considered right-censored, and will only contribute to the estimation of $H_{\text{base}}(t_J)$ or, equivalently, $S_{\text{base}}(t_J)$. Depending on the applications and convenience of modeling and computation, we can take either the actual failure times T_i, or the "interval" failure times $(t_{j-1}, t_j]$ as our complete data. The distinction between these two types can be viewed as the choice of the basic "continuous" time unit we adopt for our model, because any "actual" time, no matter how precisely recorded, is in fact always an interval observation. In both cases, we use the proportional-hazards assumption (Cox (1972)) to specify the conditional hazard in the presence of covariates, so that $h(t|X, \beta) = \exp(X^T\beta)h_{\text{base}}(t)$.

For $T_i \in [0, t_J]$ observed continuously, the contribution to the likelihood is

$$L(\beta \mid T_i, X_i) = f(T_i|X_i, \beta) = \exp(X_i'\beta)h_{\text{base}}(T_i)S_{\text{base}}(T_i)^{\exp(X_i'\beta)}, \tag{2}$$

using the identity $f(T) = h(T)S(T)$. When we only observe that $T_i > t_J$ (a right-censoring time), the contribution becomes

$$L(\beta \mid T_i, X_i) = S(t_J|X_i, \beta) = S_{\text{base}}(t_J)^{\exp(X_i'\beta)}. \tag{3}$$

When we observe that $T_i \in (t_{j-1}, t_j]$, with conditional survival probability $p_j \equiv S(t_j)/S(t_{j-1})$, the contribution is

$$\left(\prod_{l=1}^{j-1} p_l\right)^{\exp(X_i'\beta)} \left(1 - p_j^{\exp(X_i'\beta)}\right) \tag{4}$$

if $j \leq J$, and

$$\left(\prod_{l=1}^{J} p_l\right)^{\exp(X_i'\beta)} \tag{5}$$

otherwise. Note that (4) and (5) represent individual contributions to the so-called "discrete proportional-hazards likelihood", as in for example, Tu, Meng and Pagano (1993).

2.3. Missing data strategy and implementation

The complete-data models above imply the existence of an imputation model to handle censoring under the missing at random (MAR) assumption. What we actually impute depends on what we regard as complete data, i.e., we may need

to impute either a continuous T_i or the interval observation $(t_{j-1}, t_j]$. The most common form of failure time censoring is right-censoring, or loss to followup. The AIDS dataset in Section 4 also displays interval censoring and truncation. In most cases, it is assumed that we know the interval $(t_s, t_e]$ in which the censored failure time occurred; in the special case of right-censoring, where t_e is ∞, we may integrate the missing data out of the likelihood and simply use the expressions in (3) or (5). When the complete data is discrete, for any discrete-time bin $(t_{j-1}, t_j]$ within the range $(t_s, t_e]$, the probability that the actual (unobserved) failure time t_{fail} falls within that bin is

$$\frac{1}{P(t_s < t_{\text{fail}} < t_e | X_i, \beta, S_{\text{base}}(t))} P(t_{j-1} < t_{\text{fail}} < t_j | X_i, \beta, S_{\text{base}}(t)),$$

creating a multinomial distribution on the discrete-time bins containing the censored t_{fail}. Note that this distribution is conditional upon the covariates of the censored patient and the current estimates of covariate effects and baseline survival curve. With this imputation strategy, of course, failure times are assumed to be missing at random (MAR) (Little and Rubin (1987)), so that censoring is not informative. Or putting it differently, we hope we have included enough covariates to adequately capture the real essence of the censoring mechanism. In practice, missing failure times are imputed at each step of the Gibbs sampler, conditional upon current draws of all the other parameters. When imputed values of the missing data items are not of primary interest in the analysis, they can be marginalized out of the parameter inference. Section 4 gives a detailed description of the imputation strategy for the AIDS reporting delay example.

2.4. Hyperpriors for hyperparameters a, k and λ

As detailed in Section 2.1, the hyperparameter k governs the relationship among priors for the $R_{m,p}$, and plays a critical role in determining the prior correlations among the baseline hazard increments. Although we can choose a fixed k in our specification of prior, it is often desirable to "estimate" or adaptively choose k from the data by putting a hyperprior on k, which we choose to be the exponential distribution with mean μ_k. This leads to the full conditional prior for k, conditioning on all other parameters of the model,

$$\pi(k|k^-) \propto \exp(-\frac{k}{\mu_k}) \prod_{m=1}^{M} \prod_{p=0}^{2^{m-1}-1} \left[\frac{\Gamma(2ak^m)}{\Gamma^2(ak^m)} \left(R_{m,p}(1-R_{m,p})\right)^{ak^m-1} \right], \qquad (6)$$

where the notation θ^- denotes all parameters/variables except for θ itself. Expression (6) shows that the information in the data for k will come from the

joint posterior of a and all the "splits" $R_{m,p}$. Both the model with 'fixed k' and the hierarchical Bayes model will be investigated in the context of the leukemia example in Section 3.1.

For the "basic" shape parameter a, we adopt a zero-truncated Poisson (ZTP) hyperprior. We use this discrete prior for computational convenience, with the understanding that integer shape parameters are adequate for most practical purposes. Put differently, the information in the data about a is typically not strong enough to meaningfully distinguish the "true" a from its closest integer approximation. The density for ZTP with mean μ_a is $e^{-\mu_a}{\mu_a}^a/[a!(1-e^{-\mu_a})]$, and thus full conditional distribution for a is proportional to

$$\frac{\mu_a^a}{\Gamma(a+1)}\frac{H_{0,0}^a}{\Gamma(a)}\prod_{m=1}^{M}\prod_{p=0}^{2^{m-1}-1}\left[\frac{\Gamma(2ak^m)}{\Gamma^2(ak^m-1)}\left(R_{m,p}(1-R_{m,p})\right)^{ak^m}\right].$$

This distribution can be sampled using any general method for sampling from univariate discrete distributions (e.g., the inverse CDF method).

Finally, to model our prior uncertainty about the scale parameter λ for the Gamma prior in (1) on the total cumulative hazard $H_{0,0} = H(t_J)$, we place an exponential prior on λ with mean μ_λ, creating the full conditional distribution

$$\pi(\lambda|\lambda^-) \propto \exp(-\frac{\lambda}{\mu_\lambda})\frac{\exp(-\frac{H_{0,0}}{\lambda})}{\lambda^a}.$$

The next section discusses how to generate samples from these distributions.

2.5. Markov chain Monte Carlo Bayesian model estimation

In the case of the continuous-time likelihood in (2) and (3), for patient i we observe the failure or censoring time T_i and a censoring indicator δ_i (0 for censoring, 1 for observed failure). The joint parameter posterior distribution is then proportional to

$$\pi(a)\pi(\lambda)\pi(k)\pi(H_{0,0}|\lambda,a)\prod_{m=1}^{M}\prod_{p=0}^{2^{m-1}-1}\pi(R_{m,p}|k,a)\prod_{i=1}^{N}L(T_i|\delta_i,H_{0,0},R_{m,p}),$$

where $\pi(\cdot)$ denotes the prior or hyperprior. For general types of censoring, as in the AIDS reporting delay example, it typically is necessary to include the missing data as part of our sampling variables. However, in the case of continuously observed, right-censored failure times, it is more efficient to integrate out the missing data analytically and allow for right-censoring in the likelihood expression. That is, for the continuous failure time data with simple right-censoring, the ith patient's contribution to the overall likelihood is

$$[h_{\text{base}}(T_i)]^{\delta_i}\exp(-H(\min(T_i,t_J))).$$

Patients with failure time $T_i > t_J$ are considered to suffer administrative censoring ($\delta_i = 0$) at time t_J, and contribute $\exp(-H(t_J))$ to the total likelihood. Note that here we present only the likelihood contribution for the baseline hazard; covariate effects are introduced as for the standard proportional hazards model (Cox (1972)).

We use the Gibbs sampler (Geman and Geman (1984)) to sample from the joint posterior for the missing failure times and model parameters by cycling the following steps, where the variables being conditional upon in each full conditional are listed only if they are not conditionally independent given the rest.

1. Impute missing failure times for continuous complete-data failure time observed only for an interval $[t_s, t_e]$, as in Section 2.3.
2. Draw $H_{0,0}$ from the full conditional $\pi(H_{0,0}|\lambda, R_{m,p}, a) = \mathcal{G}a((a + \sum_{i=1}^{N} \delta_i), 1/[(1/\lambda) + \sum_{i=1}^{N} F(T_i)])$, with mean $\mu = (a + \sum_{i=1}^{N}\delta_i)/[(1/\lambda) + \sum_{i=1}^{N}F(T_i)]$, where $F(T_i) = H(\min(T_i, t_J))/H(t_J)$, a function of the T_i and $R_{m,p}$. Note that $F(T_i)$ does not depend on $H_{0,0} = H(t_J)$, as it has been factored out of the cumulative hazard $H(\min(T_i, t_J))$.
3. Draw the $R_{m,p}$ (in any order) from $\pi(R_{m,p}|k, H_{0,0}, a)$ with log full conditional (for given m and p)

$$\left(ak^m - 1 + \sum_{i=1}^{N} \delta_i \Pi_{i,r}\right) \log(R_{m,p}) + \left(ak^m - 1 + \sum_{i=1}^{N} \delta_i \Pi_{i,r'}\right) \log(1 - R_{m,p}) - \sum_{i=1}^{N} H(\min(T_i, t_J)), \tag{7}$$

   where r and r' are the columns of $\mathbf{\Pi}$ corresponding to $\log(R_{m,p})$ and $\log(1 - R_{m,p})$ respectively. Note that this conditional distribution is *not* Beta, as the terms $H(\min(T_i, t_J))$ involve the $R_{m,p}$ as well as $H_{0,0}$ when $T_i < t_J$.
4. Draw k from $\pi(k|R_{m,p})$ as described in Section 2.4.
5. Draw λ from $\pi(\lambda|H_{0,0})$ as described in Section 2.4.
6. Draw a from $\pi(a|H_{0,0}, R_{m,p})$ as described in Section 2.4.

The full conditional in (7) can be shown to be log-concave and is therefore sampled using the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992). The full conditional distributions for λ and k are in general *not* log-concave, and are sampled from by an extension of ARS known as adaptive rejection Metropolis sampling (ARMS), described in Gilks, Best and Tan (1995).

In the case of an "interval" complete-data failure time as in (4) and (5) in Section 2.2, steps 4–6 are unaltered, but in steps 2 and 3 we sample from the

two conditional posterior distributions

$$\pi(H_{0,0}|\lambda, a, R_{m,p}, T_i) \propto \pi(H_{0,0}|\lambda, a) \prod_{i=1}^{N} L(H_{0,0}, R_{m,p}|T_i),$$

$$\pi(R_{m,p}|k, H_{0,0}, a, T_i) \propto \pi(R_{m,p}|k, a) \prod_{i=1}^{N} L(H_{0,0}, R_{m,p}|T_i).$$

The imputation in step 1 is then for interval failure time data rather than continuous-time data.

Even in the complete-data case, it is difficult to formally demonstrate convergence properties of this sampler. Although there are general results available to determine the geometric ergodicity of Gibbs samplers (see, for example, Robert and Casella (1999)), the time-inhomogeneous transition kernel employed by the ARMS algorithm does not allow these results to be directly applied, as the proposal density can change with each step of the sampler. In the place of a formal analysis, we therefore present empirical diagnostics in Section 4.5 to demonstrate practical convergence of marginal quantities of interest.

2.6. Theoretical properties of the multiresolution prior

Our first result is to analytically determine the *prior* covariance $cov(d_i, d_j) = E(d_i d_j) - E(d_i)E(d_j)$, under the specification given in (1), for any i and j. Assume that after $H_{0,0}$, d_i and d_j share the first $L-1\,(L \geq 1)$ levels of "splits". Then, by the tree construction, the remaining levels of "splits" will all be different, and therefore we can write, with $B_{m,p}$ being either $R_{m,p}$ or $(1 - R_{m,p})$,

$$d_i = H_{0,0} \left(\prod_{l=1}^{L-1} B_{l,p_l} \right) B_{L,p_L} \prod_{l=L+1}^{M} B_{l,p_{i,l}},$$

$$d_j = H_{0,0} \left(\prod_{l=1}^{L-1} B_{l,p_l} \right) (1 - B_{L,p_L}) \prod_{l=L+1}^{M} B_{l,p_{j,l}}.$$

Under (1), one can verify that $E(H_{0,0}) = \lambda a$; $E(H_{0,0}^2) = \lambda^2 a(a+1)$; $E(B_{m,p}) = 1/2$; $E(B_{m,p}^2) = (1/4)((2k^m a+2)/(2k^m a+1))$; $E(B_{m,p}(1-B_{m,p})) = (1/4)(2k^m a/(2k^m a+1))$. It follows then that

$$\begin{aligned}
&E(d_i d_j) - E(d_i)E(d_j) \\
&= \prod_{l=L+1}^{M} E(B_{l,p_{i,l}})E(B_{l,p_{j,l}}) \left[E(H_{0,0}^2) \prod_{l=1}^{L-1} E(B_{l,p_l}^2) E(B_{L,p_L}(1 - B_{L,p_L})) - \frac{\lambda^2 a^2}{4^L} \right] \\
&= \frac{\lambda^2 a(a+1)}{4^M} \left(\frac{2ka+2}{2ka+1} \right) \cdots \left(\frac{2k^{L-1}a+2}{2k^{L-1}a+1} \right) \left(\frac{2k^L a}{2k^L a+1} \right) - \frac{\lambda^2 a^2}{4^M}
\end{aligned}$$

$$= \frac{\lambda^2 a^2}{4^M}\left[(2ka+2k)\left(\frac{2k^2a+2k}{2ka+1}\right)\cdots\left(\frac{2k^La+2k}{2k^{L-1}a+1}\right)\left(\frac{1}{2k^La+1}\right)-1\right]$$

$$= \frac{\lambda^2 a^2}{4^M}\left[\left(\prod_{l=1}^{L}\frac{(2k^la+2k)}{(2k^la+1)}\right)-1\right]. \tag{8}$$

Now note that for $k = 0.5$, the terms $(2k^la + 2k)/(2k^la + 1)$ are all identically 1, implying zero prior correlation among the hazard increments. When $k > 0.5$, these terms are all greater than 1, and the increments will be positively correlated *a priori*. Correspondingly, for $k < 0.5$ the hazard increments will have negative prior correlation.

To show that for $k = 0.5$ the baseline hazard increments are in fact independent, consider the cumulative baseline hazard parameter $H_{0,0} \sim \mathcal{G}a(a, \lambda)$, as in Section 2.1, and the "top-level" ratio parameter $R_{1,0} = H_{1,0}/H_{0,0} \sim \mathcal{B}e(a/2, a/2)$. Then given that $H_{1,0} = H_{0,0}R_{1,0}$ and $H_{1,1} = H_{0,0}(1-R_{1,0})$, a change-of-variables calculation gives that $H_{1,0}$ and $H_{1,1}$ are *a priori* i.i.d. $\mathcal{G}a(a/2, \lambda)$. Following this argument recursively, the baseline hazard increments d_j will be i.i.d. $\mathcal{G}a(a/2^M, \lambda)$ in the prior.

For general k, the situation is more complex, although it is still possible to analyze the marginal prior distribution of the d_j's. Given that $d_j \equiv H_{0,0}\prod_{m=1}^{M} B_{m,p_m}$, these hazard increments are, in the prior, an independent product of a Gamma variable and M Betas. Consequently, it is easy to establish

$$E(d_j^n) = \frac{\lambda^n \langle a\rangle_n \prod_{m=1}^{M}\langle ak^m\rangle_n}{\prod_{m=1}^{M}\langle 2ak^m\rangle_n},$$

where $\langle a\rangle_n \equiv \Gamma(a+n)/\Gamma(a) = a(a+1)\ldots(a+n-1)$. This formula provides a very convenient tool for studying the marginal distribution of d_j. For example, it shows that the mean of d_j is $\lambda a/2^M$, which is free of k, as it should be. We can also easily simulate d_j and estimate its density with a Gaussian kernel density estimator, as shown in Figure 1, where $M = 3, \lambda = 1$, a follows a ZTP hyperprior with mean μ_a, and k has an exponential hyperprior with mean μ_k. We see, for example, that for a fixed μ_k, as μ_a increases, the right tail of the density for the d_j also increases. For given μ_a, however, the patterns are much less clear and therefore more analysis is required to determine the effect of differing μ_k on the distribution of the d_j.

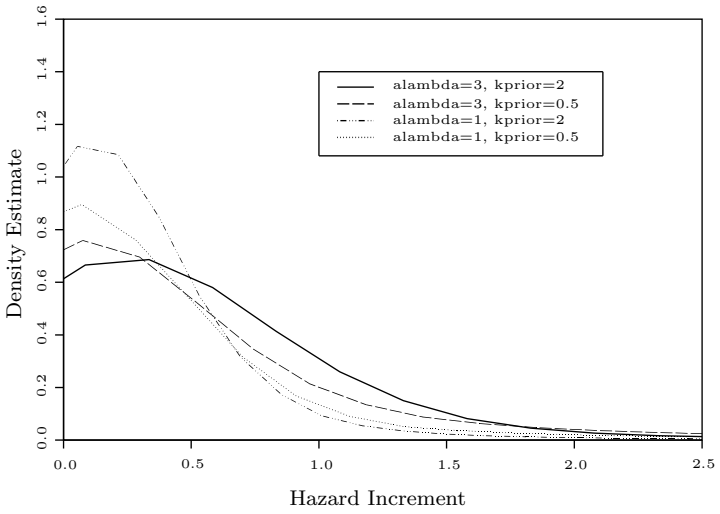

Figure 1. Density estimates for hazard increment priors under varying hyperpriors for a and k; *alambda* is the rate hyperparameter for a, while *kprior* is the mean of the hyperprior for k.

2.7. Combating a pitfall

From expression (8), we see that when $k > 0.5$, the prior covariance, and hence correlation, of d_i and d_j is an increasing function of L (since all $(2k^l a + 2k)/\,(2k^l a + 1) > 1$). However, since $L - 1$ is the number of common "splits" d_i and d_j share, and the splits are arranged according to a binary tree structure, L is not a monotone function of $|i - j|$. This implies that nearby increments can have smaller correlation than those that are further apart. For example, with $M = 3$, d_4 and d_5 have $L = 1$, but d_4 and d_1 have $L = 2$. Consequently, $\rho(d_4, d_5) = 0.07 < \rho(d_4, d_1) = 0.13$, as in Table 1, where $a = 1$.

Table 1. Correlations for differing L distances, 4-128 bin models with hyperprior on k and $a = 1$; ρ_i is the correlation for increments with $L = i$.

4 bins	$\rho_2 = 0.22$, $\rho_1 = 0.12$
8 bins	$\rho_3 = 0.18$, $\rho_2 = 0.13$, $\rho_1 = 0.07$
16 bins	$\rho_4 = 0.13$, $\rho_3 = 0.10$, $\rho_2 = 0.07$, $\rho_1 = 0.04$
32 bins	$\rho_5 = 0.09$, $\rho_4 = 0.07$, $\rho_3 = 0.05$, $\rho_2 = 0.04$, $\rho_1 = 0.02$
64 bins	$\rho_6 = 0.06$, $\rho_5 = 0.05$, $\rho_4 = 0.04$, $\rho_3 = 0.03$, $\rho_2 = 0.02$, $\rho_1 = 0.01$
128 bins	$\rho_7 = 0.03$, $\rho_6 = 0.03$, $\rho_5 = 0.02$, $\rho_4 = 0.02$, $\rho_3 = 0.01$, $\rho_2 = 0.01$, $\rho_1 = 0.01$

Although it is not a mathematical paradox to allow increments further from each other to have stronger dependence than those that are closer, from the perspective of prior specification, it is desirable to avoid a model that is generally against our intuition or actual prior belief.

Table 1 (where $a = 1$ and k has an exponential hyperprior with mean $\mu_k = 2$, inducing prior correlation) shows that the difference between the correlations for different L can be quite substantial, especially for $M \leq 4$, and thereby the aforementioned "reversal" can be of practical importance. Whereas this problem is inherent in the binary tree construction, we can significantly reduce its effect by mixing over a hyperprior of a. Table 2 is a counterpart of Table 1, but with a given a ZTP hyperprior with $\mu_a = 4$ and k given the same exponential hyperprior with $\mu_k = 2$. It is evident that the mixing has almost completely equalized all the correlations. Although this is still less desirable than having the correlations as monotone decreasing functions of $|i - j|$, the mixing at least prevents the correlation from increasing noticeably with $|i-j|$. We also see from both Table 2 and Table 1, that the *a priori* correlations decrease as the resolution level increases, and therefore the mixing strategy becomes less important when M is large. For this reason, the analysis in Section 4 is performed with $a = 1$.

Table 2. Correlations for differing L distances, 4-128 bin models with hyperpriors on k and a; ρ_i is the correlation for increments with $L = i$.

4 bins	$\rho_2 = 0.28$, $\rho_1 = 0.27$
8 bins	$\rho_3 = 0.17$, $\rho_2 = 0.16$, $\rho_1 = 0.16$
16 bins	$\rho_4 = 0.10$, $\rho_3 = 0.09$, $\rho_2 = 0.09$, $\rho_1 = 0.09$
32 bins	$\rho_5 = 0.06$, $\rho_4 = 0.05$, $\rho_3 = 0.05$, $\rho_2 = 0.05$, $\rho_1 = 0.04$
64 bins	$\rho_6 = 0.03$, $\rho_5 = 0.03$, $\rho_4 = 0.03$, $\rho_3 = 0.03$, $\rho_2 = 0.03$, $\rho_1 = 0.03$
128 bins	$\rho_7 = 0.02$, $\rho_6 = 0.02$, $\rho_5 = 0.02$, $\rho_4 = 0.02$, $\rho_3 = 0.02$, $\rho_2 = 0.02$, $\rho_1 = 0.02$

Other multiresolution applications (e.g., Kolaczyk (1999) and van Dyk, Connors, Esch, Freeman, Kang, Karovska, Kashyap, Siemiginowska and Zezas (2004)) use a "cycle-spinning" algorithm to handle the correlation problem, effectively assuming that the intensity being estimated is periodic. Because the hazard function is an aperiodic function of time and we cannot join the ends of the function together, mixing over hyperparameters appears to be a more satisfactory approach for our purposes.

3. Empirical Investigations Using Simulated and Textbook Datasets

3.1. Effects of hyperparameters and hyperpriors

In this subsection we investigate the impact of different choices of fixed k and a on the posterior hazard estimates for relatively small sample sizes. We use the well-known example of the leukemia dataset analyzed in Cox (1972) to illustrate the effects of different choices of k and a, comparing the inferences to those from the hierarchical Bayes models with hyperpriors on k and a. Table 3 gives leukemia remission times in weeks for two groups under either a drug or placebo;

with 41 patients, 12 of the drug-group patients have right-censored remission times. For this example, we assume that the remission times are continuously observed and use the ungrouped data likelihood as given in Section 2.2.

Table 3. Cox's Leukemia Data (in weeks).

Group	Observed Remission Times (* = Right-Censored)
Drug Group	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 34*, 35*
Placebo Group	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

We have already demonstrated in Section 2.6 that the prior choices of k and a influence prior correlations among the hazard increments d_j. We now wish to check the effects of induced correlation in our multiresolution prior upon the posterior inference for this dataset.

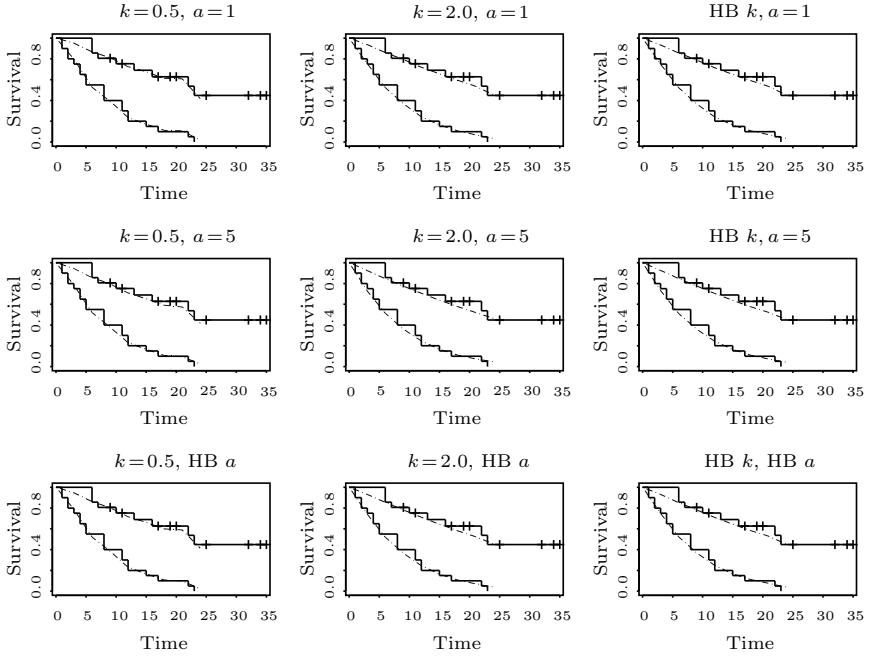

Figure 2. Posterior Inference of Survival Curves for Cox' Leukemia Data, under Differing Values of k and a (Solid Lines are Kaplan-Meier plots, dashed lines are multiresolution estimates; upper line is drug group, lower line control).

Figure 2 shows mean posterior survival estimates for three choices of k (0.5, 2.0 and hierarchical Bayes), with three choices of a (1.0, 5.0 and hierarchical Bayes). For hierarchical Bayes, we used a ZTP hyperprior with rate parameter $\mu_a = 4$ on a and an exponential hyperprior with mean $\mu_k = 2$ on k, creating prior correlation as in the previous section. We placed a hierarchical Bayes exponential hyperprior on λ with mean 100 in all nine analyses. Kaplan-Meier

plots of the data are shown in solid lines; dashed lines give survival estimates from the multiresolution model with eight equal time bins on the interval [0,24]. The hierarchical Bayes model for k and a gives posterior (2.5%, 50%, 97.5%) quantiles of (0.82, 2.07, 3.38) for k and a 95th sample percentile of 6.0 for a, indicating substantial posterior uncertainty about these hyperparameters given the data. As Figure 2 shows, there is very minor impact on the posterior inference from the choice of hyperparameters.

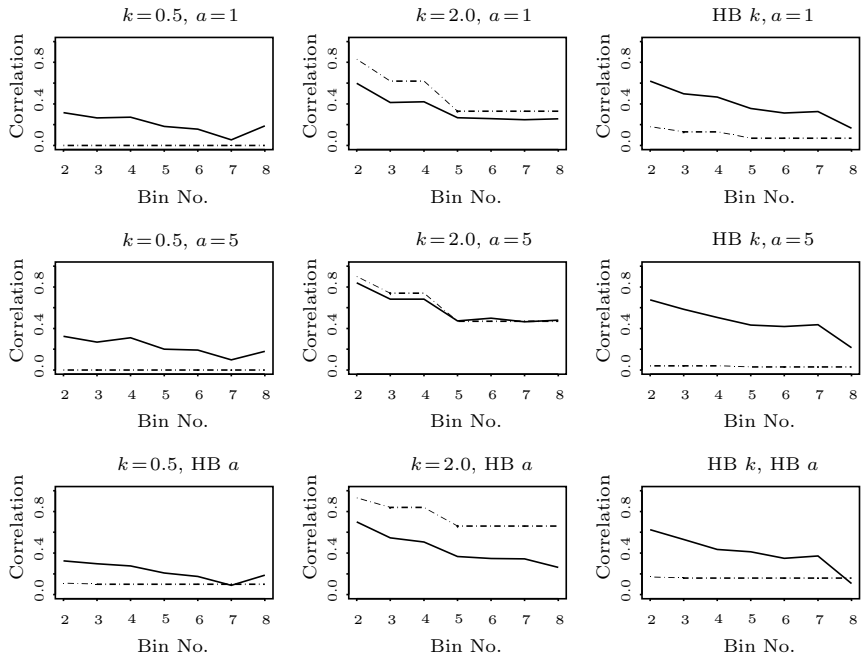

Figure 3. Prior (dashed) and posterior (solid) hazard increment correlations $\mathrm{Corr}(d_1, d_i)$, $i = 2 \ldots 8$, Leukemia example.

Figure 3 displays both prior and posterior correlations for hazard increment 1 (that is, $\mathrm{Corr}(d_1, d_i)$ for $i = 2, \ldots, 8$) for the nine models, showing the positive prior correlations for fixed $k = 2.0$, the zero prior correlation for $k = 0.5$ and the correlations induced by the hierarchical Bayes hyperprior on k. Dashed lines indicate prior correlations for the first hazard increment; solid lines indicate posterior correlations. Choice of k clearly has a strong impact on both prior and posterior correlations between bin 1 and bins 2-8; the choice of a has a weaker impact. However, a more interesting finding is that with the mixture of values of a given by the hierarchical Bayes model, the prior correlation between hazard increments remains fairly constant across time bins.

This lack of sensitivity, although not desirable from the perspective of promoting the multiresolution approach, is largely a piece of good news for the

process of making inferences. We report this investigation to show that in the case of even a relatively small dataset such as the leukemia data, a very informative prior on the hazard increments and their correlation would be required to have an impact on the resulting inference; similar results were obtained in Nieto-Barajas and Walker (2002).

3.2. Curvature information in the posterior of k

To examine what information is contained in the marginal posterior distribution for the parameter k, we compared the curvature, as measured by a "roughness penalty", of two discretized cumulative hazard functions to the respective modes of the full conditional of k given $H_{0,0}$ and $R_{m,p}$. For a range of shape parameters, we divided Weibull and Gamma cumulative hazards into 4, 8, 16, 32, 64 and 128 equally spaced discrete hazard increments on the time interval $[0, 1]$. For the Weibull distribution with unit scale parameter and shape parameter α, the cumulative hazard at time t is t^α. For the Gamma distribution with unit scale and shape α, the cumulative hazard involves an incomplete Gamma integral: $H_\alpha(t) = -\log(1 - \int_0^t (1/\Gamma(\alpha)) s^{\alpha-1} \exp(-s) ds)$. We chose the Weibull cumulative hazard on the interval $[0, 1]$ for the invariance of the cumulative hazard $H(1) \equiv 1^\alpha \equiv 1$ for all shape parameters; by contrast, the Gamma cumulative hazard exhibits very strong variation in $H(1)$ for the shape parameters considered.

For each discretized cumulative hazard at a given shape and level of aggregation, we generated 5,000 simulated observations from each distribution and then fit the six multiresolution models described above for the baseline hazard; we then determined the posterior distribution for k for each of six levels of resolution and multiple Weibull and Gamma shape parameters (Weibull shapes $\alpha = 1, 2, 2.5, 3, 3.5, 5, 7, 9$ and Gamma shapes $\alpha = 3, 4, 5, 6, 7, 8, 9, 10$). As a measure of smoothness or curvature for the cumulative hazard function, we used $\int (d^2 H(t)/dt^2)^2 dt$ from the smoothing spline literature (see e.g., Hastie and Tibshirani (1990)). The curvature of the Weibull cumulative hazard can be determined analytically for given α; the Gamma hazard curvature was computed using numerical integration. Note that this curvature is a global measure for a particular curve, while k governs local relationships among the $R_{m,p}$.

Figures 4 and 5 show posterior plots, for 4, 8, 16, 32, 64 and 128 bins, of the k posterior for differing shape values for both Weibull and Gamma distributions. Solid lines show the posterior medians for k as a function of curvature; dashed lines give posterior 2.5% and 97.5% quantiles. Interestingly, the posterior shows three trends: lower variance with an increasing number of bins (or levels M to the model), decreasing k with increasing integrated squared second derivative, and *increasing* k with an increasing number of bins. A larger number of bins

(particularly 128 time bins) gives an increasingly small variation in k with the curvature of the underlying cumulative hazard function. Intuitively, for a given number of bins, as the curvature increases, the correlation among the hazard increments becomes less and less, thereby driving the posterior distribution of k more closely around $k = 0.5$. We conjecture that as the curvature approaches infinity, all curves in the plots will reach $k = 0.5$ as their asymptotes, albeit it is clear that the rate at which this occurs is a monotone function of M, that is, the larger the M, the larger the curvature it takes to reach the asymptote.

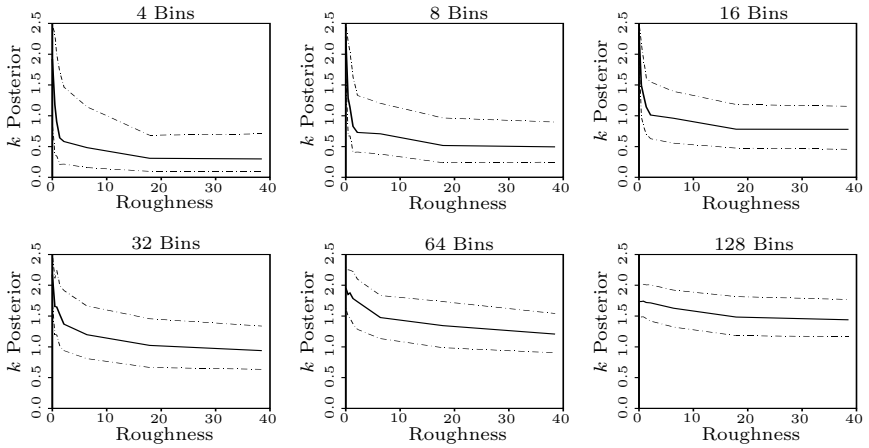

Figure 4. Posterior (2.5%, 50%, 97.5%) percentiles for k versus cumulative hazard roughness, Weibull simulation.

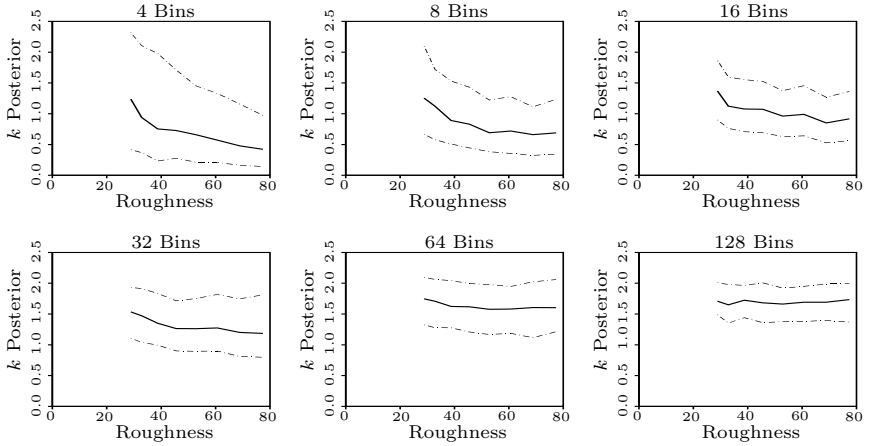

Figure 5. Posterior (2.5%, 50%, 97.5%) percentiles for k versus cumulative hazard roughness, Gamma simulation.

3.3. Comparison with estimation using other Bayesian priors

In this section we compare the multiresolution model of the baseline hazard to two classes of stochastic process priors for the cumulative hazard function $H(t)$. Beta processes were developed by Hjort (1990), who derived the posterior for a beta process prior for right-censored survival data without ties. Very recent work by Kim and Lee (2003) and Lee and Kim (2002) gives an efficient algorithm for sampling from a beta process posterior, as well as closed-form expressions for the posterior under left-truncation and right-censoring in the proportional-hazards model. Nieto-Barajas and Walker (2002) develop a correlated gamma-process prior on piecewise-constant hazard rates (hereafter the "NBW model"), using two latent processes to connect adjacent hazard rates. This model is further developed by Craiu and Duchesne (2004), who add data augmentation techniques to address competing-risks data with masked failure causes. Here we use a simulation to compare the performance of our semiparametric multiresolution model to that of these other two. Specifically, we evaluate the relative efficiency of our approach by the bias and variance we incur when estimating survival at a sequence of time horizons t_j. (A nice introduction to prior processes for the cumulative distribution function (CDF) and the cumulative hazard can be found in Ghosh and Ramamoorthi (2003).)

For the first model comparison, we use a beta process centered on a unit exponential cumulative hazard, with prior parameters $A_0(t) \equiv 1, c(t) \equiv 1$, where $A_0(t)$ is the prior mean of the exponential cumulative hazard $H(t)$ and $c(t)$ is a scaling parameter as a function of time t. Using the notation that $Y_n(t)$ is the number of patients alive and under observation at time t, for our chosen prior the posterior parameters can be calculated as

$$c_N(t) = 1 + Y_n(t), A_N(t) = \sum_{i=1}^{m+1} \frac{t_i - t_{i-1}}{N+2-i} + \sum_{t_j \le t} \frac{1}{N-j+2},$$

where we assume that there are no ties in the observations, and $t_1, \ldots, t_m$ are the $m = N_n(t)$ failure times before time t and $t_0 = 0, t_{m+1} = t$. The mean of this posterior is $A_N(t)$; the expression for the posterior variance at time t is

$$\sum_{t_j \le t} \frac{N-j+1}{(N-j+2)^2(N-j+3)} + \sum_{i=1}^{m+1} \frac{t_i - t_{i-1}}{(N+2-i)(N+3-i)}.$$

For our second comparison, we use a simplified version of the NBW model which Craiu and Duchesne (2004) used as the starting point for their competing-risks model. Following their notation, we divide the observation time $[0, t_{\max}]$ into J intervals of constant hazard rate d_j. These parameters are accompanied

by a latent process $u_j, j = 1, \ldots, J-1$ taking values on the non-negative integers, so that the total parameter set has the Markovian dependency graph

$$d_1 \to u_1 \to d_2 \to \ldots \to u_{J-1} \to d_J.$$

The latent process u_j allows correlation to be introduced between adjacent d_j. The parameters are given the prior structure $d_1 \sim \text{Ga}(\alpha_1, \beta_1)$, $u_j|d_j \sim \text{Poisson}(c_j d_j)$, $d_{j+1}|u_j \sim \text{Ga}(\alpha_{j+1} + u_j,\ \beta_{j+1} + c_j)$, $j = 1, \ldots, J-1$. For simplicity, we take $\alpha_j \equiv \alpha = 0.001, \beta_j \equiv \beta = 0.001, c_j \equiv C$. Values of C equal to 0, 1 and 5 are used in the simulation below to represent increasing levels of autocorrelation in the hazard rate process. (Note that when $C = 0$, the d_j are *a priori* independent; as C increases, so does prior correlation.) While we take the c_j to be constant, these hyperparameters may also be given Exponential hyperpriors to allow dependencies to be chosen in a data-driven fashion; see Nieto-Barajas and Walker (2002) for further details. Under this setup, the conditional posterior distributions for the tth Gibbs sampler iteration become

$$d_1^{(t)}|d_1^- \sim \mathcal{G}a(\alpha_1 + u_1^{(t-1)}, \beta_1 + C + e_1),$$

$$d_j^{(t)}|d_j^- \sim \mathcal{G}a(\alpha_j + u_{j-1}^{(t-1)} + u_j^{(t-1)} + n_j, \beta_1 + 2C + e_j),$$

where n_j counts the number of failures observed in the jth interval and e_j is the total unit exposure, or the sum over all units of time under observation during the jth time interval.

For our simulation, 100 uncensored observations were generated from a Weibull distribution with shape parameter 5 and $H(t) = t^5$. The NBW model and the multiresolution model were fitted with $J = 2^M = 16$ bins, bin width $\Delta t = 0.075$ and maximum time horizon $t_J = 1.2$. We then computed posterior mean estimates and variances for each estimator at 48 equally-spaced time points on the interval [0.025,1.200]. The NBW and multiresolution estimates for other time points were derived using the piecewise-constant hazards assumption in computing posterior means and variances.

Figure 6 plots the ratios of the posterior variance, squared bias and MSE as a function of time for the multiresolution model over the beta process prior model (first panel) and over the NBW model with three values of C (remaining panels) in estimating the Weibull cumulative hazard between times 0.6 and 1.2. (Cumulative hazard estimates earlier than time 0.6 are too small to give stable ratio estimates.) Analysis shows that while the multiresolution model often has slightly higher variance than the beta process model, the squared bias, which drives the MSE for much of the time interval, is slightly lower, giving a slightly lower MSE between times 0.6 and 1.0. (However, in the region from time $1.1-1.2$, the ratios become very large and numerically unstable due to the fact that the

bias and variance of the beta process estimator happen to be very near zero.) At the same time, the multiresolution model depends upon the assumption of a piecewise-constant hazard rate, which corresponds to linear interpolation for the cumulative hazard function.

For $C = 0$, the NBW model is very comparable in performance to the multiresolution model. For larger values ($C = 1$ and $C = 5$), however, the induced prior correlation creates so much bias that the MSE becomes much worse than for the multiresolution model. We emphasize that the variance ratio between the multiresolution and the NBW model is high (Figure 6, bottom panels) because of the NBW model's overwhelming prior assumptions (under large C) that overly constrain its estimator's variability. Figure 7 displays the posterior mean estimates of the cumulative hazard for all four pairs of models, along with the true cumulative hazard function and the Nelson-Aalen estimator $\hat{\Lambda}(t) = \sum_{t_j \leq t} 1/(N - j + 1)$. For this realization, the posterior estimates display the slight downward bias from the true hazard of all three Bayesian models; for the NBW model with large C the bias may become overwhelming.

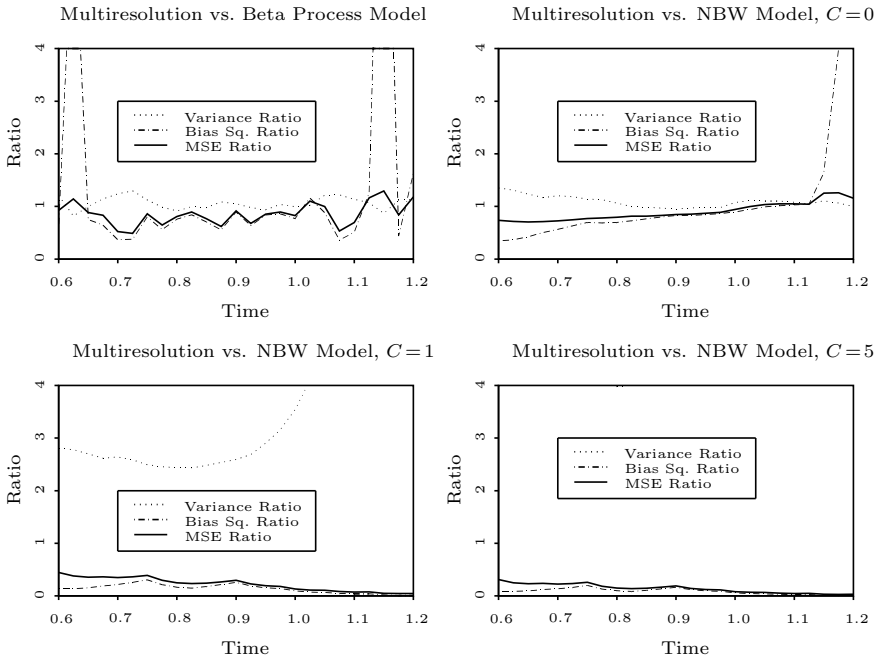

Figure 6. Ratios of variances, squared Biases and MSEs, Beta process and 16-Bin multiresolution estimators for the Weibull example, evaluated at 48 equally-spaced time points (ratios numerically unstable below t=0.6 due to low bias of beta process estimator). The variance ratio in the last panel is outside of range due to the overly strong C and consequently low variability for the NBW model.

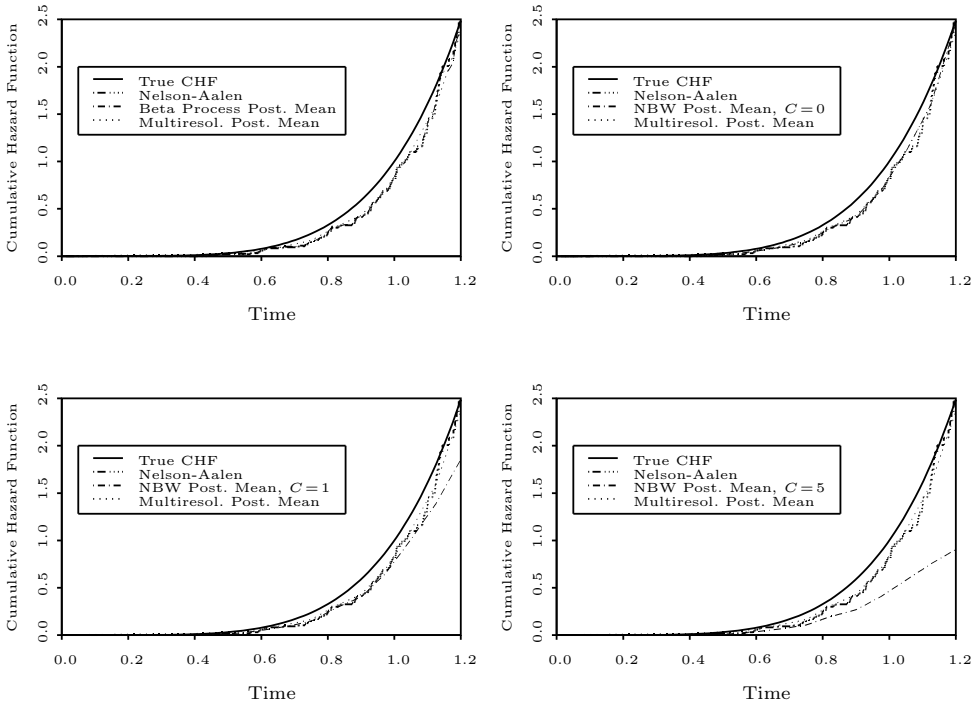

Figure 7. Comparison of Weibull cumulative hazard and Nelson-Aalen estimate with Mean posterior estimates for Beta process prior, 16-bin multiresolution prior, and NBW prior (based on one simulated sample of 100 uncensored observations). Comparison is performed at 48 equally-spaced time points.

Incidentally, we can show that the mean of the beta process estimator is close to the Nelson-Aalen estimator by computing

$$A'(t) - \hat{\Lambda}(t) = \sum_{i=1}^{m+1} \frac{t_i - t_{i-1}}{N+2-i} + \sum_{t_j \le t} \frac{1}{N-j+2} - \sum_{t_j \le t} \frac{1}{N-j+1}$$
$$= \sum_{i=1}^{m+1} \frac{t_i - t_{i-1}}{N+2-i} - \sum_{t_j \le t} \frac{1}{(N-j+1)(N-j+2)} \le \frac{t}{N+1-N_n(t)},$$

where the difference is small, bounded by $0.5/97 \approx 0.005$ at $t = 0.5$ and by $1.0/40 = 0.025$ at time $t = 1.0$.

4. Application: Estimating AIDS Reporting Delay

Our case study in hazard estimation involves the administrative delay between AIDS diagnosis and case report receipt at the Centers for Disease Control

and Prevention (CDC). In the AIDS Public Information Dataset (APIDS) both month and year of each diagnosis and month and year of report receipt are recorded, giving a reporting delay with two interval-censored endpoints. In any calculations of AIDS prevalence in larger studies of the epidemic, knowledge of this reporting delay is necessary to adjust for its biasing deflation of counts of recent AIDS diagnoses. In an analogous analysis of United States cancer trends between 1981 and 1998, Clegg, Feuer, Midthune, Fay and Hankey (2002) point out that reporting delay in the NIH Surveillance, Epidemiology and End Result (SEER) cancer registries can give rise to significant downward biases in cancer incidence trends unless properly adjusted for. The results in the following section illustrate how such a reporting delay distribution can be estimated in a Bayesian setting using our multiresolution prior, and build (in part) upon the work detailed in Tu, Meng and Pagano (1993), in which a model for the reporting delay is used to de-bias estimates of AIDS survival.

4.1. Data source and description

The CDC APIDS consists of 816,849 case reports, describing AIDS diagnoses made between 1981 and 2001 and received by the CDC as of the end of December 2001. Of these cases, we selected a study group by the following criteria: case meets the 1987 CDC AIDS case definition and was reported in the New York City metropolitan area, patient was a white male age 40-44 (the median age) and AIDS was contracted via sexual contact with other men. There were 2,528 cases that met this definition and reported both month and year of AIDS diagnosis and case report. A histogram of $\log_2$(1+reporting delay times in months) is given in Figure 8. The discretization at the lower end of the distribution is due to the delay being interval-censored to an integer number of months. Unadjusted for interval-censoring and truncation, the histogram displays a modal delay time of two months, with a maximal time of 146 months (12 years, 2 months), giving a strongly right-skewed distribution. We analyze the reporting delay by quarter instead of month, choosing for convenience to estimate its distribution at 32 quarterly intervals for eight years.

As we have noted, the CDC APIDS censors both endpoints of the reporting delay by not recording either the date of diagnosis or date of report receipt given month and year. This double interval censoring leads to an uncertainty of up to two months in the delay time, and we adjust for it via binomial imputation in a Bayesian missing data analysis. Additionally, the fact that APIDS contains only complete reports received as of December 2001 implies that some diagnoses made during the 1981-2001 study period may not have reached the CDC by the file closure date, biasing our analysis. To adjust for this truncation, we again impute missing reporting delays using the negative binomial model, as described below,

for each case report for which (*file closure date-diagnosis date*) < *maximal delay time*. For the conditional distributions in this imputation, we make the approximating assumption that the maximal reporting delay is 33 quarters (99 months or 8.25 years), retaining 2,522 (99.8%) of the cases in our study sample. Our estimates of the reporting delay distribution are therefore a slightly truncated approximation to the full curve.

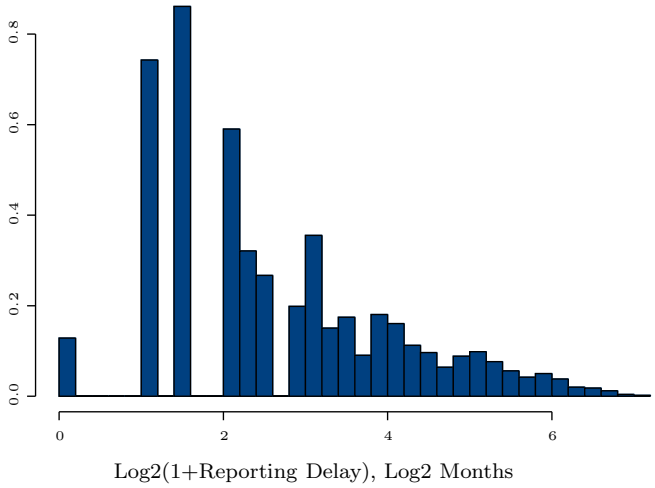

Figure 8. Variation in AIDS case reporting delay.

4.2. Example of interval-censoring in AIDS data

To provide a concrete (albeit pedagogical) example of the missing data imputation strategy for the CDC APIDS analysis, imagine that our dataset has only two years of case reports (January 1999-December 2000) to analyze, where each observation gives month and year of AIDS diagnosis and month and year of AIDS case report receipt at the CDC, and the file is closed at the end of December 2000. Our goal for this imaginary dataset is to analyze the hazard for the reporting delay time–the time taken for an AIDS diagnosis by a doctor or hospital to reach the CDC. As in Section 4.1, we analyze the reporting delay by quarter, and assume that the maximum possible reporting delay is one year or four quarters.

Suppose, for example, that a given patient is diagnosed with AIDS in March 1999 and his case report reaches the CDC in June 1999, for a putative reporting delay of 3 months. Because of the uncertainty in the exact date of diagnosis and report receipt, the underlying reporting delay could actually range from two months (for diagnosis at the end of March and report receipt at the beginning of June) to four months (for diagnosis at the beginning of March and report receipt

at the end of June), meaning that we are unsure whether the report receipt falls in the first or second quarter after diagnosis. For such cases, we use a binomial imputation model. Let $3Q$ be the putative number of months of reporting delay for n case reports with covariates X; then the number of cases imputed to a delay of Q quarters is

$$Y \sim \text{Binom}\Big(n, \frac{S(3Q-1 \text{ months}|X) - S(3Q \text{ months}|X)}{S(3Q-1 \text{ months}|X) - S(3Q+1 \text{ months}|X)}\Big),$$

and the number imputed to $Q+1$ quarters is $n-Y$. To compute the conditional survival probabilities for delays of a number of months not evenly divisible by three, we use the piecewise-constant hazards assumption for the baseline hazard and survival curve.

4.3. Example of truncation in AIDS data

Now suppose (in our pedagogical example) that a cell of patients is diagnosed with AIDS in March 2000 and their case reports reach the CDC in July 2000. Since the time between these patients' diagnoses and the file closure in December 2000 is at most ten months, less than the maximal time of one year, we must account for the possibility, conditional upon current estimates of covariate effects and the baseline reporting hazard, that the case report of someone diagnosed in March 2000 would have reached the CDC after December 2000 and therefore have not been included in the data we have at hand.

For patients diagnosed in a month and year for which the file closure date creates the possibility of truncated case report counts, we assume the diagnosis took place at the middle of the month, and employ the following negative binomial imputation strategy to "re-inflate" the report total. Let $t_{\text{lead}} = t_{\text{close}} - t_{\text{diag}}$ be the "lead time" between case diagnosis and file closure in quarters or fractions of quarters, and $p_{\text{report}} = 1 - S(t_{\text{lead}}|X)$ be the probability of a diagnosis at t_{diag} reaching the CDC before file closure, conditional upon covariates X. Then the number of additional imputed cases m, estimating the number of truncated reports with the same diagnosis time t_{diag} and covariates X will be $Y-1$, where Y is a geometric random variable with probability parameter p_{report}; if there are k such patients with parameter $p_{\text{report}} = 1 - S(t_{\text{lead}}|X)$, then $m = Y - k$, where Y is negative binomial with parameters (p_{report}, k). Furthermore, the distribution of the m imputed reporting delay times will be multinomial, with probabilities $(S(t_{\text{close}}|X) - S(t_{\text{diag}} + Q_j|X), S(t_{\text{diag}} + Q_j|X) - S(t_{\text{diag}} + Q_{j+1}|X), \ldots, S(t_{\text{diag}} + Q_{n-1}|X) - S(t_{\text{diag}} + Q_n|X))$, where $Q_j, \ldots, Q_n$ are the integer numbers of quarters after diagnosis time, up to the maximal delay time, that follow t_{close}. Again, cumulative hazards and survival probabilities for fractional numbers of quarters

are computed using the piecewise-constant hazards assumption. These m imputed reporting delay times are then added to the sufficient statistics used by the Gibbs sampler in the next MCMC parameter draws.

4.4. Hazard estimation

Five models were estimated for the AIDS reporting delay data, using the prior structures indicated in Table 4 and the discrete-time proportional-hazards likelihood from Section 2.2. For all five models, we used a value of $a = 1$ for the Gamma prior shape parameter for $H_{0,0}$, and placed hierarchical Bayes hyperpriors on λ and k as explained in Section 4.1. While the likelihood function involved estimates of the reporting delay distribution at 32 quarterly intervals, the series of models considered estimated the delay hazard at resolution of 1, 2, 4, 8 and 16 quarters, using the piecewise-constant hazard assumption to interpolate hazard values at intermediate times. As Table 4 shows, Models 2−5 are restrictions of Model 1 which fix "lower-level" subsets of the "split" parameters $R_{m,p}$ at 0.5, effectively estimating the hazard function at lower resolution. Pointwise median, 2.5% and 97.5% curves for Models 1−5 are given in Figure 9. Hazard uncertainty generally increases with increasing time as expected. Models 1-3 describe an initially high reporting hazard, decreasing significantly during the first two years, while Models 4-5 incorrectly force the hazard the hazard for the first two years to be constant.

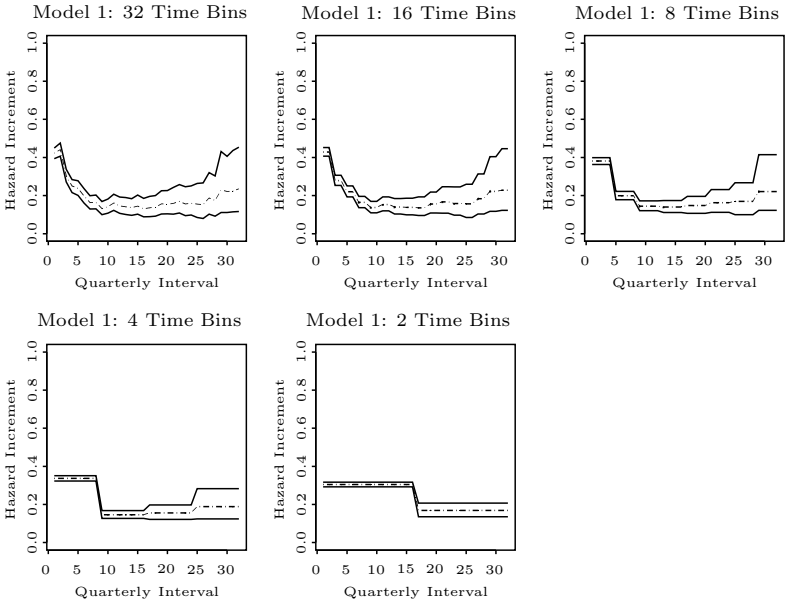

Figure 9. Posterior pointwise 95% credible intervals for hazard function, Models 1−5, AIDS dataset.

Table 4. Comparing prior structures for AIDS reporting delay example.

Model	No. of Bins	Bin Width, Quarters	Levels M	p_D	DIC
1	32	1	5	15.1	10695
2	16	2	4	10.3	10703
3	8	4	3	6.2	10765
4	4	8	2	3.5	10876
5	2	16	1	1.9	11012

We also performed a sensitivity analysis to choice of scale in the baseline hazard function, checking for independence of our inferences from the number of time bins $J = 2^M$. An insightful discussion of model construction and selection through invariance considerations may be found in Gelman (1996). Our model can be usefully constrained by requiring invariance to aggregation: our posterior inference about d_1, for example, should be comparable to our inference about $d_1 + d_2$ when the number of bins J is doubled and the width of each bin is halved. Inferences for proportional-hazards covariates should also not be significantly affected by the choice of time scale. We note that in comparison to the 32-bin model, for example, using the 16-bin model with the piecewise-constant hazards assumption is equivalent to assuming that $R_{m,p} \equiv 0.5$ for $m = 5$, because of the piecewise-constant hazard assumption, or to discarding the highest resolution.

Section 2.1 described the prior self-consistency properties of the multiresolution model. To check posterior self-consistency, we graphed the effects of aggregation in Models 1−5. Figure 10 shows posterior inferences for the baseline cumulative hazard increments for four levels of aggregation: 16 bins, 8 bins, 4 bins and 2 bins. In the first row, for example, we compare the posterior inference for the 16-bin model to that obtained if we added together adjacent bins of the 32-bin model posterior; the pointwise median and 2.5%/97.5% posterior quantiles for the 16-bin model are shown with bold lines, while those for the 32-bin model are shown with regular line width. The graph shows near posterior invariance under aggregation from 32 to 16 bins, equivalent to marginalizing out $R_{m,p}, m = 5$ from the 32-bin model. Likewise, the second, third and fourth rows display the results of aggregating "higher-resolution" models to 8 bins, 4 bins and 2 bins. Notably, the inference for the first time bin in the 4-bin and 2-bin models does not seem to be invariant under aggregation. We speculate that because 8-bin, 16-bin and 32-bin models reveal that the baseline hazard declines during the first two years, the 2-bin and 4-bin models, which assume a constant hazard over the first two years after diagnosis, begin to provide an incorrect inference about the hazard function during this period. This method of checking posterior self-consistency provides one way to restrict the choice of bins. Section 4.6 discusses a more formal way of choosing the level of model resolution.

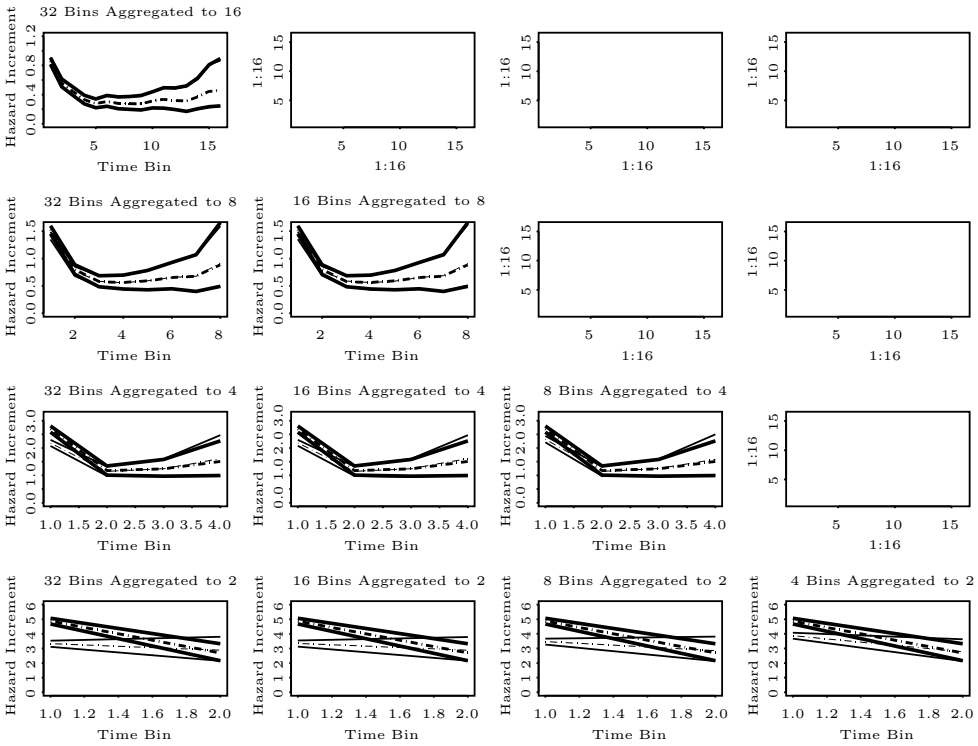

Figure 10. Results of posterior aggregation of hazard increments, Models 1−5, AIDS dataset.

4.5. MCMC convergence assessment

Section 2.5 notes that we used a combination of diagnostic criteria to empirically monitor convergence of the Monte Carlo sampler to the parameter posterior distribution for each model. Our primary tool is the multivariate Brooks-Gelman criterion (Brooks and Gelman (1998)), which compares an estimate $\hat{V}$ of the posterior parameter covariance matrix with an estimate W of the within-chain parameter covariance matrix, defining a scale reduction factor $R_p = \max_a(a'\hat{V}a/a'Wa)$. Convergence is diagnosed when R_p declines to 1 over a large number of iterations for m chains starting from overdispersed locations in parameter space. Once convergence has been diagnosed, the first half of each chain can be discarded as burn-in, and the second halves of the chains pooled together as a sample from the posterior.

For each of our five models, we ran five chains from random starting positions for 100,000 iterations, thinning the chains by a factor of 10 to reduce sample autocorrelation to yield five samples of size 10,000 from posterior parameter space.

Figure 11 displays the Brooks-Gelman plots for, respectively, the 2, 4, 8, 16 and 32-bin models, indicating good convergence for the hazard increments d_j in each case. (Note that $R_{\max}$ is the maximum Gelman-Rubin convergence diagnostic (Gelman and Rubin (1992)) across all model parameters. It was shown in Brooks and Gelman (1998) that $R_{\max} \leq R_p$.) Figure 12 shows the autocorrelation functions (ACFs) for each of the 32 d_j for the 32-bin model for one of five chains, displaying autocorrelation very close to 0 for these parameters after thinning the chain by a factor of 10. We are therefore reasonably confident that the sampler has converged for all five models, with nearly uncorrelated draws.

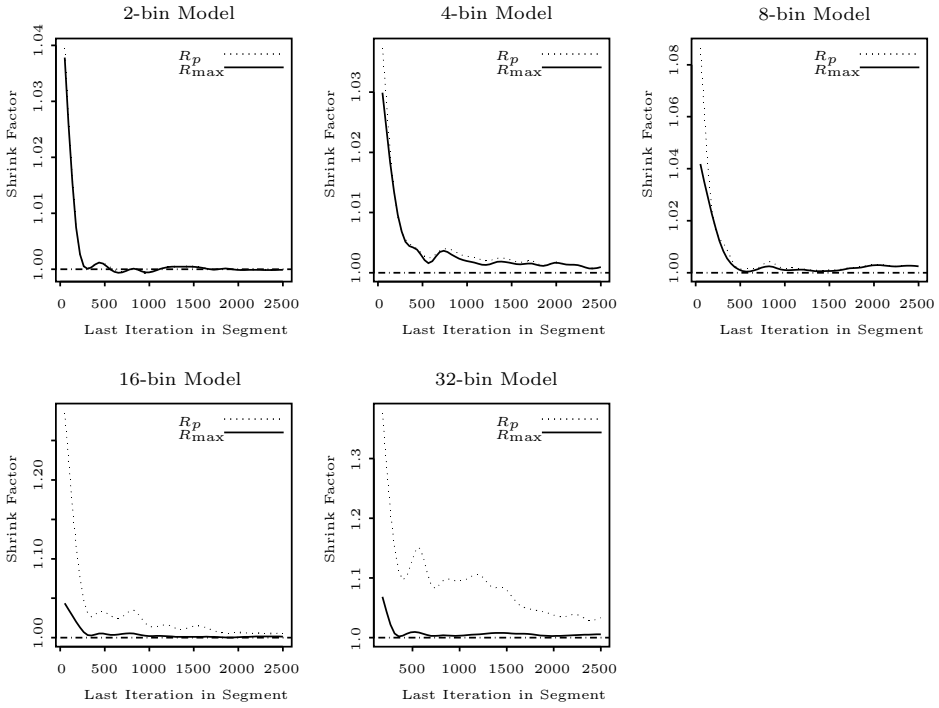

Figure 11. Brooks-Gelman convergence diagnostics.

4.6. Model comparison criteria for number of time bins J

In this section, we investigate the use of model selection criteria to address the question of how many time "bins" to include in the multiresolution model. The DIC explored in Spiegelhalter et al. (2002) begins with an estimation of the effective number of model parameters p_D using the formula

$$p_D\{y, \Theta, \tilde{\theta}(y)\} = E_{\theta|y}[-2\log(p(y|\theta))] + 2\log[p(y|\tilde{\theta}(y))], \tag{9}$$

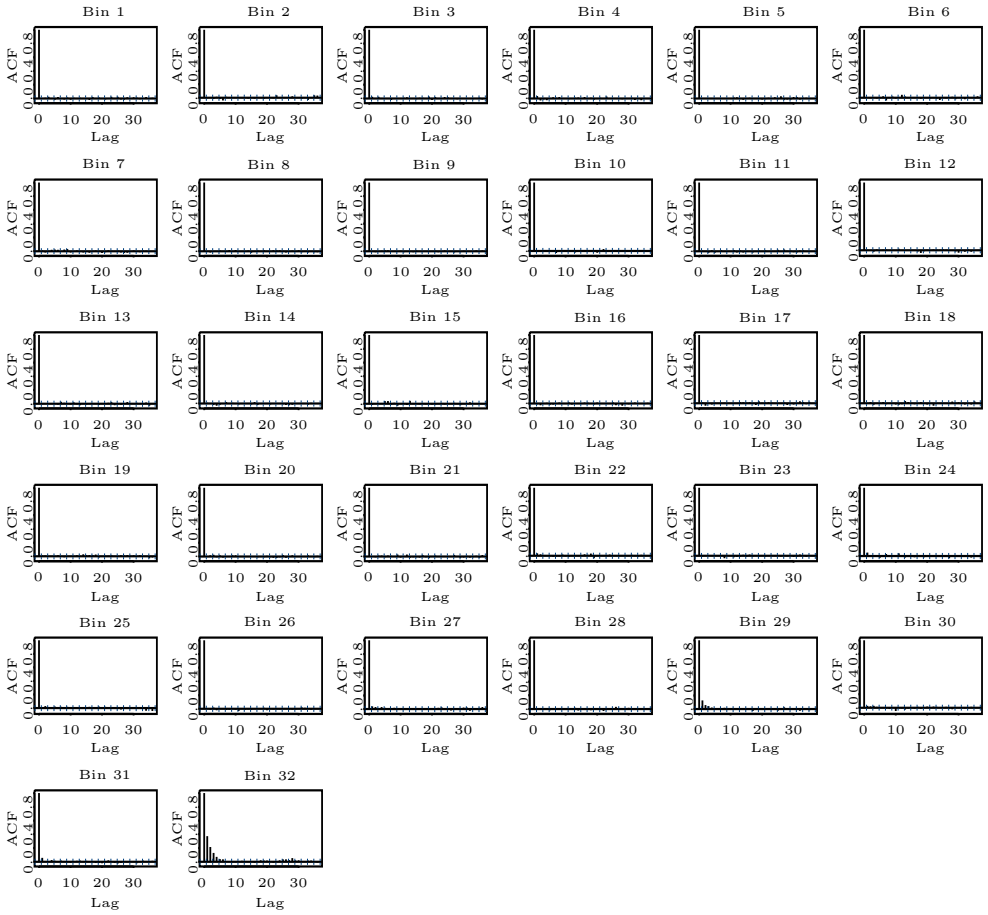

Figure 12. Autocorrelation functions for draws of d_j parameters from one chain, 32 bin-model, AIDS dataset.

where Θ is the set of model parameters which appear directly in the complete data likelihood and $\tilde{\theta}(y)$, usually the posterior mean $E(\theta|y)$, is the model estimate of the parameters chosen. This expression intuitively may be thought of as a reduction in uncertainty due to estimation, and it is argued in Spiegelhalter and others (2002) that parallels between p_D and classical measures of model complexity make it a good candidate for the effective number of parameters.

Writing $D(\theta) = -2\log(p(y|\theta))$, we can use MCMC output to estimate $p_D = \overline{D(\theta)} - D(\overline{\theta})$, the empirical averages corresponding to (9) above. Regarding $D(\overline{\theta})$ as a measure of deviance for the data at the posterior mean, Spiegelhalter and others (2002) propose $2p_D + D(\overline{\theta})$ as a model selection criterion, penalizing the decreasing deviance by twice the effective number of model parameters. In the case of the AIDS reporting delay analysis, we analyzed the distribution of failure times at the quarterly level; some of the reporting delays were interval-

censored or truncated by file closure at the end of December 2001. The complete data included missing data imputed through binomial and negative binomial imputation models; the DIC was evaluated as explained above based on observed and imputed data for each draw from the posterior sample.

Table 4 displays our calculations for the DIC for Models 1−5 in the AIDS example. Interestingly, we can see that p_D does not fall close to the "true" number of increments for Models 1-3, implying less than a "full" reduction in uncertainty due to the model fit. The DIC indicates that the 16-bin and 32-bin models are possibly comparable fits to the data, as suggested by the hazard posterior plots. Models with 8, 4 and 2 time bins apparently provide a significantly worse fit to the data in return for their decreased complexity, possibly because of the bias associated with the large bin size, a problem that was also revealed partially by the self-consistency checking in Section 4.4. It seems that combining the adjacent pairs of bins in the 32-bin model (whose baseline hazard fit is shown in Figure 9) and the hazard smoothing it implies does not lead to significant problems with the model. So among the five models investigated here, we can settle for either 32 bins or 16 bins. Figure 13 gives the corresponding reporting fraction (that is, the cumulative distribution function) by quarter for these models, showing that in fact the "survival" curves for Models 1 and 2 are not readily distinguishable.

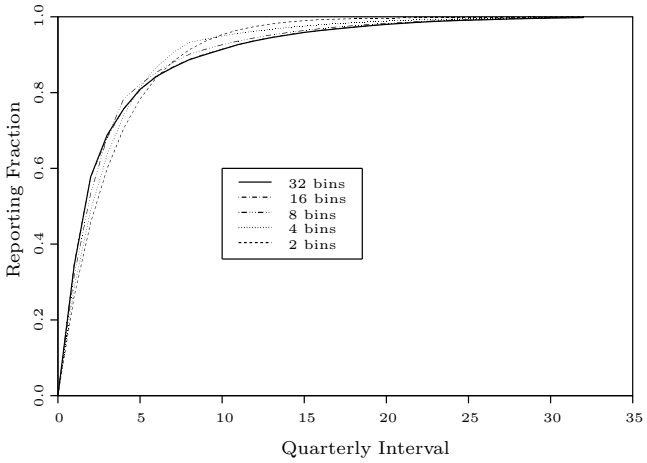

Figure 13. Results of posterior aggregation of hazard increments, Models 1−5, AIDS dataset (Bold lines indicated unaggregated model, thin lines the aggregated model.)

5. Summary and Further Work

We believe that the relative simplicity and flexibility of our hazard model, along with its ability to account for a variety of censoring mechanisms and other

interesting properties, make it a competitive method when compared to the greater difficulty of implementation for non-parametric methods like the Beta process prior. In many applied studies, a discrete-interval estimate of the baseline hazard, using the piecewise-constant hazards approach, may be completely sufficient for the needs of the analysis. We have shown in our applied example the possibility of using model-selection methodology to choose an appropriate level of resolution in a data-driven manner.

While this paper deals only with the hazard function for a single group of individuals, the model (as demonstrated in the description of the likelihood in Section 2.2) can easily incorporate the effects of covariates using standard proportional- or additive-hazards assumptions. Bouman, Meng, Dignam and Dukic (2005), for example, use this model to study the variation in disease-free survival of breast cancer patients treated with either placebo or tamoxifen in a five-year multicenter clinical trial.

Although we believe our method is useful and convenient, clearly it is not perfect and more research, both theoretical and methodological, is needed. For example, additional work will be required to reveal the exact nature of the information that the posterior for k and a conveys about the baseline hazard function. Following the general iterative process of model building, as outlined in the first chapter of Box and Tiao (1973), we would also like to develop posterior predictive model checking procedures to suggest further model refinements. Furthermore, we regard this model as an initial (discrete-time) attempt to construct a prior on baseline hazard curves while incorporating assumptions about the underlying curvature of these hazards. We hope in future research to be able to pursue extensions that come closer to the goal of placing a prior on the space of smooth curves itself, while at the same time largely maintaining the conceptual and computational simplicity of the multiresolution approach.

Acknowledgement

This paper is dedicated to Professor George Tiao, a pioneer in Bayesian analysis, on the occasion of his 70th birthday. The authors thank Radu Craiu and Per Mykland for helpful discussion. Xiao-Li Meng's work was partially supported by NSF grants DMS 0072510 and DMS 0204552.

References

Antoniadis, A., Grégoire, G. and Nason, G. (1999). Density and hazard rate estimation for right-censored data by using wavelet methods. *J. Roy. Statist. Soc. Ser. B* **61**, 63-84.

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4**, 505-524.

Aslanidou, H., Dey, D. K. and Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods. *Canad. J. Statist.* **26**, 33-48.

Bouman, P., Meng, X.-L., Dignam, J. and Dukic, V. (2005). A multiresolution hazard model for multicenter survival studies: application to tamoxifen treatment in early stage breast cancer. *J. Amer Statist. Assoc.* Forth coming.

Box G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley, New York.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7**, 434-455.

Clegg L., Feuer, E. J., Midthune, D. N., Fay, M. P. and Hankey, B. F. (2002). Impact of reporting delay and reporting error on cancer incidence rates and trends. *J. Nat. Cancer Inst.* **94**, 1537-1545.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Craiu, R. V. and Duchesne, T. (2004). Using EM and data augmentation for the competing risks model. In *Applied Bayesian Modelling and Causal Inference from Incomplete-Data Perspectives*. (Edited by A. Gelman and X.-L. Meng), 239-251. Wiley, London.

Gelman, A. (1996). Bayesian model-building by pure thought: some principles and examples. *Statist. Sinica* **6**, 215-232.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457-511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721-741.

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

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling. *Appl. Statist.* **44**, 455-472.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259-1294.

Huang, H.-C. and Cressie, N. (2001). Multiscale graphical modeling in space: applications to command and control. In *Spatial Statistics: Methodological Aspects and Applications* (Edited by M. Moore), 83-113. Springer-Verlag, New York.

Huang, H.-C., Cressie, N. and Gabrosek, J. (2002). Fast, resolution-consistent spatial prediction of global processes from satellite data. *J. Comput. Graph. Statist.* **11**, 63-88.

Kim, Y. and Lee, J. (2003). Bayesian analysis of proportional hazard models. *Ann. Statist.* **31**, 493-511.

Kolaczyk, E. D. (1999). Bayesian multiscale models for Poisson processes. *J. Amer. Statist. Assoc.* **94**, 920-933.

Kolaczyk, E. D. and Nowak, R. D. (2005). Multiscale generalized linear models for nonparametric function estimation. *Biometrika*. Forth coming.

Lee, J. and Kim, Y. (2002). *A new algorithm to generate beta processes*. Technical report. Department of Statistics, Pennsylvania State University.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Nieto-Barajas, L. E. and Walker, S. G. (2002). Markov beta and gamma processes for modeling hazard rates. *Scand. J. Statist.* **29**, 413-424.

Nowak, R. D. and Kolaczyk, E. D. (2000). A statistical multiscale framework for Poisson inverse problems *IEEE Trans. Inform. Theory* **46**, 1811-1825.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer Verlag, New York.

Sahu, S. K., Dey, D. K., Aslanidou, H. and Sinha, D. (1997). A Weibull regression model with Gamma frailties for multivariate survival data. *Lifetime Data Anal.* **3**, 123-137.

Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data. *J. Amer. Statist. Assoc.* **88**, 979-983.

Sinha, D. (1998). Posterior likelihood methods for multivariate survival data. *Biometrics* **54**, 1463-1474.

Sinha, D. and Dey, D. K. (1997). Semiparameteric Bayesian analysis of durvival data. *J. Amer. Statist. Assoc.* **92**, 1195-1212.

Spiegelhalter D. J., Best, N. G., Carlin, B. P. and Angelike van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. Ser. B* **64**, 583-616.

Tu, X. M., Meng, X.-L. and Pagano, M. (1993). The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Amer. Statist. Assoc.* **88**, 26-36.

van Dyk, D., Connors, A., Esch, D., Freeman, P., Kang, H., Karovska, M., Kashyap, V., Siemiginowska, A. and Zezas, A. (2004). Deconvolution in high energy astrophysics: science, instrumentation and methods. *Bayesian Anal.* To appear.

Willett, R. M. and Nowak, R. D. (2003). *Multiscale density estimation.* Technical report. Department of Electrical and Computer Engineering, Rice University.

Kellogg School of Management, 2001 Sheridan Rd, Evanston, IL 60208, U.S.A.

E-mail: p-bouman@kellogg.northwestern.edu

Department of Health Studies, University of Chicago, Chicago, IL 60637, U.S.A.

E-mail: vdukic@health.bsd.uchicago.edu

Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138-2901, U.S.A.

E-mail: meng@stat.harvard.edu

(Received December 2003; accepted April 2004)