Mixtures of Polya trees for flexible spatial frailty survival modelling

BY LUPING ZHAO

Eli Lilly & Company, Indianapolis, Indiana 46285, U.S.A.
zhao0117@umn.edu

TIMOTHY E. HANSON AND BRADLEY P. CARLIN

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
hans0058@umn.edu carli002@umn.edu

SUMMARY

Mixtures of Polya trees offer a very flexible nonparametric approach for modelling time-to-event data. Many such settings also feature spatial association that requires further sophistication, either at the point level or at the lattice level. In this paper, we combine these two aspects within three competing survival models, obtaining a data analytic approach that remains computationally feasible in a fully hierarchical Bayesian framework using Markov chain Monte Carlo methods. We illustrate our proposed methods with an analysis of spatially oriented breast cancer survival data from the Surveillance, Epidemiology and End Results program of the National Cancer Institute. Our results indicate appreciable advantages for our approach over competing methods that impose unrealistic parametric assumptions, ignore spatial association or both.

Some key words: Areal data; Bayesian modelling; Breast cancer; Conditionally autoregressive model; Log pseudo marginal likelihood; Nonparametric modelling.

1. INTRODUCTION

In survival studies, the hazard function for individuals within certain groups may depend on risk factors, some of which may be unknown. Vaupel et al. (1979) introduced the notion of unknown group-specific risk factors, or frailties, incorporated into the survival model as random effects to be estimated from the data. The use of both parametric and semiparametric hierarchical frailty survival models has become rather common, since they offer a computationally and conceptually appealing approach for capturing the association among individual survival times within groups. A variety of parametric and nonparametric choices for the baseline hazard function have been explored in the literature. The frailties are typically assumed to be independent and identically distributed with mean zero, but when the groups correspond to geographic regions, a spatially associated distribution may be more natural. For example, Banerjee et al. (2003) developed parametric frailty specifications based on both lattice and point-referenced spatial models, and compared them with standard approaches under a Weibull baseline hazard function in the context of county-level infant mortality data. Banerjee & Carlin (2002, 2003) developed semiparametric Cox frailty models via beta mixture and counting process approaches, and compared the models using the deviance information criterion (Spiegelhalter et al., 2002).
In the lattice case, the discretely indexed regions instead partition the geographic region being studied. The spatial information in this type of model is usually based on the adjacency of regions, rather than on any continuous distance metric. The most commonly used lattice model is the conditionally autoregressive model (Besag, 1974). Li & Ryan (2002) developed a class of semiparametric proportional hazards spatial frailty models by allowing a set of spatial random effects to enter the baseline hazard function multiplicatively. Banerjee et al. (2003) fitted Cox proportional hazards frailty models and Banerjee & Dey (2005) fitted proportional odds models, both in spatially correlated survival data settings. Banerjee & Carlin (2003) developed semiparametric spatio-temporal frailty models using hierarchical Bayesian methods, which were further extended by Jin & Carlin (2005), who proposed a multivariate conditionally autoregressive model for areally referenced multiple disease data.

In this paper, we consider three models commonly used with survival data: the accelerated failure time model, the proportional hazards model, and the proportional odds model. All three provide useful summary information in the absence of an estimate of the baseline survival distribution, and hence are often fitted using semiparametric methods. The parametric part provides acceleration factors, relative risk factors or relative odds, respectively, which associate the patient risk to a typically small number of regressors. The nonparametric part is for the baseline hazard or survival function, which we may wish to leave as arbitrary as possible.

We compare several aspects of modelling: choice of accelerated failure time, proportional hazards or proportional odds; two types of frailty model or absence thereof; and parametric versus nonparametric assumptions on baseline survival $S_0$. The proportional hazards, accelerated failure time, and proportional odds models all make rather stringent, overarching assumptions about the data generating mechanism for the sake of obtaining succinct data summaries. A novel aspect of the present paper is that we compare competing survival models assuming the same, flexible nonparametric prior for baseline survival. The mixtures of Polya trees baseline hazard can be taken to be the same across the three models, so differences in predictive performance may be attributed to the survival and frailty models only, rather than to additional possible differences in quite different nonparametric priors. In recent, related work, Zhang & Davidian (2008) fit accelerated failure time, proportional hazards, and proportional odds models without frailties, assuming a flexible class of normalized polynomials for baseline survival $S_0$. Li & Lin (2006) and Hennerfeind et al. (2006) developed flexible spatial survival models assuming proportional hazards; alternative specifications were not considered.

Several interesting overarching models have been proposed, including transformation models that include proportional hazards and proportional odds as special cases (e.g. Scharfstein et al., 1998; Mallick & Walker, 2003), transformation and extended regression models that include proportional hazards and additive hazards as special cases (e.g. Yin & Ibrahim, 2005; Martinussen & Scheike, 2006, Ch. 7) and hazard regression models that include both proportional hazards and accelerated failure time as special cases (e.g. Chen & Jewell, 2001). While highly flexible, these models have the deficiency that their regression parameters lack simple interpretability. Furthermore, there may be insufficient information to estimate the additional transformation and regression parameters included in the models. Many of these approaches seem best suited to comparing the appropriateness of two competing models, both embedded within a larger model, with the aim of model reduction and enhanced interpretability. Model interpretation can also proceed via population averaged inference, as recommended by Gustafson (2007) for the transformation model proposed by Yin & Ibrahim (2005). We instead emphasize model interpretability and selection over what essentially amounts to model averaging.

We illustrate our proposed spatial mixtures of Polya trees methodology using a subset of the Surveillance, Epidemiology, and End Results cancer database, as maintained by the
United States National Cancer Institute; see seer.cancer.gov. These data were previously analyzed by Banerjee et al. (2003) and Jin & Carlin (2005) in the context of a proportional hazards spatial frailty model, and comprise the survival times in months for women diagnosed with breast cancer in the state of Iowa during 1995–1998. Important predictors of survival include age at diagnosis, race, number of primaries, i.e. the number of physiologically independent cancers diagnosed, and the stage of the disease, local, regional or distant. Figure 1 shows a county-level choropleth map of the logarithm standardized mortality ratio, defined as the ratio of the observed and expected number of deaths in each county. The expected number of deaths is obtained through internal standardization as the the county population times the overall mortality rate for the state (Banerjee et al., 2004, pp. 158–59). Although there is substantial statewide variability, there does appear to be some local similarity of the rates in neighbouring counties, with clusters of elevated standardized mortality ratios in the east and southwest.

2. Statistical models

2.1. Survival modelling

The proportional hazards, accelerated failure time and proportional odds models can all be formulated in terms of the baseline survival function $S_0$. Let $x_{ij}$ be a $p$-dimensional vector of covariates associated with the $j$th individual in group $i$ ($j = 1, \ldots, n_i$, $i = 1, \ldots, n$) and let $S_{x_{ij}}(\cdot)$ be the associated survival function. We consider patients grouped at the county level, so that $n$ is the number of counties. Let the frailty associated with group $i$ be $\gamma_i$.

The proportional hazards model assumes

$$S_{x_{ij}}(t) = S_0(t)^{\exp(x'_{ij} \beta + \gamma_i)}$$

(1)

while the accelerated failure time model assumes

$$S_{x_{ij}}(t) = S_0\{\exp(x'_{ij} \beta + \gamma_i)t\},$$

(2)

and the proportional odds model assumes

$$\frac{S_{x_{ij}}(t)}{1 - S_{x_{ij}}(t)} = \exp(x'_{ij} \beta + \gamma_i) \frac{S_0(t)}{1 - S_0(t)}.$$

(3)

In each model, $\exp{(x_1 - x_2)' \beta}$ has a useful interpretation comparing relative risk between individuals within a county with covariates $x_1$ and $x_2$ at any time $t$ for proportional hazards; the relative mean, median or any survival quantile for accelerated failure time; or the relative odds of
surviving past any time \( t \) for proportional odds. The factor \( \exp(\gamma_1 - \gamma_2) \) compares county-level risks for any given set of covariates between counties \( i_1 \) and \( i_2 \).

The proportional odds assumption implies conditionally converging hazards for distinct covariate groups and counties, which has been observed for long-term follow-up of breast cancer (e.g. Gore et al., 1984; Jeong et al., 2003). Both proportional odds and accelerated failure time imply conditionally nonproportional hazards, also observed by Yakovlev et al. (1999) for long-term follow-up of breast cancer.

2.2. Spatial frailty modelling

Following Banerjee et al. (2003), we consider a version of the commonly used conditionally autoregressive prior of Besag et al. (1991). Here, frailty terms are conditionally specified as

\[
\gamma_i \mid \{ \gamma_j \}_{j \neq i} \sim N(\bar{\gamma}_i, (\lambda n_i)^{-1}),
\]

where \( \gamma_i \) denotes the frailty in county \( i \), \( n_i \) denotes the number of counties adjacent to county \( i \) and \( \bar{\gamma}_i \) is the sample mean of the \( n_i \) county effects in \( \{ \gamma_j \}_{j \neq i} \) adjacent to county \( i \). For the Iowa data, \( 2 < n_i < 7 \). Banerjee et al. (2003) show that the conditionally autoregressive prior performs similarly to geostatistical alternatives, but in a fraction of the computer time since it avoids inverting large matrices within each Markov chain Monte Carlo iteration.

The proportional hazards model with exchangeable frailties \( \exp(\gamma) \sim \exp(1) \) implies marginal proportional odds in the remaining covariates, i.e. averaged over the frailty distribution. Relaxing this to the more commonly used \( \exp(\gamma) \sim \Gamma(\lambda, \lambda) \), where \( \Gamma(a, b) \) denotes the gamma distribution with mean \( a/b \), yields marginally converging hazards (Clayton & Cuzick, 1985, § 5.1) when \( \lambda > 0 \); proportional hazards arise on the boundary \( \lambda = 0 \). Barker & Henderson (2004) considered a generalization of the proportional hazards model with exchangeable gamma frailties that has both proportional and converging hazards for two partitioning sets of predictors. We compare the exchangeable log-normal \( \exp(\gamma) \) to the spatial Gaussian conditionally autoregressive prior.

2.3. Mixtures of Polya trees priors for the baseline survival function

We consider models (1), (2) and (3) with a mixture of Polya trees prior on \( S_0 \), which smoothes over partitioning effects associated with a simple Polya tree, and includes the underlying centring parametric families as special cases. Consider a mixture of Polya trees prior on \( S_0 \),

\[
S_0 \mid c, \theta \sim PT(c, \rho, G_\theta), \quad \theta \sim p(\theta),
\]

where (4) is shorthand for a particular parameterization (Hanson & Johnson, 2002; Hanson, 2006). We describe the prior below but refer to these references and Lavine (1992) for technical details.

Let \( J \) be a fixed, positive integer and let \( G_\theta \) denote a family of cumulative distribution functions indexed by \( \theta \). The distribution \( G_\theta \) serves to centre the random distribution \( S_0 \). A Polya tree prior is constructed from a set of nested partitioning sets \( \Pi_\theta = \{ B_\theta(e) : e \in \bigcup_{i=1}^J \{0, 1\}^J \} \) and corresponding conditional probabilities \( \mathcal{Y} = \{ Y_e : e \in \bigcup_{i=1}^J \{0, 1\}^J \} \). A set \( B_\theta(e) = B_\theta(e_1 \cdots e_k) \) at partition level \( k \) is split into two sets: \( B_\theta(e_0) \) and \( B_\theta(e_1) \). Given that an observation from \( S_0 \) is in the parent \( B_\theta(e) \), \( Y_{e_0} \) is the conditional probability that the observation is in \( B_\theta(e_0) \) and \( Y_{e_1} \) is the conditional probability that the observation is in \( B_\theta(e_1) \). Necessarily, \( Y_{e_0} + Y_{e_1} = 1 \). Under the Polya tree prior, \( Y_{e_0} \sim \text{Beta}(\alpha_{e_0}, \alpha_{e_1}) \) independently. That is, pairs of conditional probabilities \( (Y_{e_0}, Y_{e_1}) \) in \( \mathcal{Y} \) are distributed independent Dirichlet with parameters in \( \mathcal{A} = \{ \alpha_e : e \in \bigcup_{i=1}^J \{0, 1\}^J \} \). At the coarsest level, \( (Y_0, Y_1) = (0.5, 0.5) \) for identifiability.

The partition points are quantiles of the centring family: if \( j \) is the base-10 representation of the binary number \( \epsilon = \epsilon_1 \cdots \epsilon_k \) at level \( k \), then \( B_\theta(\epsilon_1 \cdots \epsilon_k) \) is defined to
be the interval \((G^{-1}_\theta(j/2^k), G^{-1}_\theta((j+1)/2^k))[\); an exception is that the rightmost set is \(B_\theta(11 \cdots 1) = (G^{-1}_\theta(2^k - 1)/2^k), \infty)\). The family \(\mathcal{A}\) is defined by \(\alpha_{\epsilon_1 \cdots \epsilon_k} = ck^2\) for some \(c > 0\) (Walker & Mallick, 1997, 1999; Hanson & Johnson, 2002). The parameter \(c\) acts like the precision in a Dirichlet process and is directly related to how quickly the data overwhelm the prior. Very large values of \(c\) force conditional probabilities \(Y_i\) to be close to 0-5 regardless of the data, which further forces \(S_0(A) \approx G_\theta(A)\) for sets \(A\); as \(c \to \infty\) we obtain a fully parametric analysis. As \(c \to 0^+\) the posterior baseline is almost entirely data-driven, but this implies essentially zero prior weight on the centring family. Ferguson (1974) notes that \(\alpha_{\epsilon_1 \cdots \epsilon_k} = c/2^k\) yields the special case of a discrete Dirichlet process, in an infinite Polya tree where \(J \to \infty\). The Dirichlet process is the only tailfree process where the partition does not affect inference, but does not have a density with respect to Lebesgue measure.

Within sets at the level \(J\) in \(\Pi_\theta\), we assume that \(S_0\) given \(\mathcal{Y}\) and \(\theta\) follows the baseline distribution \(G_\theta\) (Hanson, 2006). Define the vector of probabilities \(p_\mathcal{Y} = (p_\mathcal{Y}(1), p_\mathcal{Y}(2), \ldots, p_\mathcal{Y}(2^J))\) through

\[
p_\mathcal{Y}(j + 1) = S_0[B_\theta(\epsilon_1 \cdots \epsilon_J) \mid \mathcal{Y}, \theta] = \prod_{i=1}^{J} Y_{\epsilon_1 \cdots \epsilon_i},
\]

where \(\epsilon_1 \cdots \epsilon_J\) is the base-2 representation of \(j\) \((j = 0, \ldots, 2^J - 1)\). After simplification, the baseline survival function is

\[
S_0(t \mid \mathcal{Y}, \theta) = p_\mathcal{Y}[k_\theta(t)]\{k_\theta(t) - 2^J G_\theta(t)\} + \sum_{j=k_\theta(t)+1}^{2^J} p_\mathcal{Y}(j),
\]

where \(k_\theta(t)\) denotes the integer part of \(2^J G_\theta(t) + 1\). The density associated with \(S_0(t \mid \mathcal{Y}, \theta)\) is

\[
f_\theta(t \mid \mathcal{Y}, \theta) = \sum_{j=1}^{2^J} 2^J p_\mathcal{Y}(j) g_\theta(t) I_{B_\theta(\epsilon_j(j-1))}(t) = 2^J p_\mathcal{Y}[k_\theta(t)] g_\theta(t),
\]

where \(g_\theta(\cdot)\) is the density corresponding to \(G_\theta\) and \(\epsilon_J(i)\) is the binary representation \(\epsilon_1 \cdots \epsilon_J\) of the integer \(i\).

Based on previous experience, we consider two priors on \(c\). The prior \(c \sim \Gamma(5, 1)\) places mass on smaller values of \(c\), allowing more flexibility in baseline modelling at the possible expense of predicting future data. The prior \(c \sim \Gamma(20, 2)\) places mass on larger values of \(c\) in an effort to smooth inferences towards the underlying parametric family. Finally, the underlying parametric family is obtained as \(c \to \infty\).

Since the mixture of Polya trees prior builds upon and directly generalizes an underlying parametric family, various methods for obtaining posterior inference can use existing inferential methods based on the parametric model as a starting point. For Markov chain Monte Carlo sampling, intractable full-conditional distributions are often updated using slice sampling or efficient block updates with a Metropolis–Hastings sampler (Carlin & Louis, 2008, pp.131–33). To be effective, the latter approach requires proposal distributions in the general shape of the full conditional distribution, which we simply take to be the underlying centring family of the mixture of Polya trees prior. The individual Polya tree parameters add detail to the overall parametric shape of \(S_0\), and are also updated using random walk Metropolis–Hastings steps. This updating scheme, described in the next section, is quite general and allows the fitting of different models using the same basic strategy. See Hanson (2006) for sampling schemes based on Gaussian centring.
families and Hanson & Yang (2007) for samplers based on Gaussian approximations for the log-logistic centring family.

2.4. Summary and computational notes

The mixture of Polya trees prior on $S_0$ is centred at the log-logistic or Weibull family of densities. Hanson & Johnson (2002) found the choice of underlying family to make little difference in density estimation with a mixtures of Polya trees prior; this also holds for the data analysis in § 3-3. The number of tree levels was capped at $J = 4$, achieving good Markov chain Monte Carlo mixing, and allowing the comparison of dozens of models in reasonable computer time. Adding a level to the tree doubles the number of Polya tree parameters, allowing $S_0$ to accommodate greater detail, but also greatly increases the computational burden. Hanson (2006) notes a law of diminishing returns in terms of a models’ predictive ability versus the level. Each model we fitted took around an hour to run on a 3-2 GHz Pentium 4, with 2 GB of RAM, in compiled Fortran.

For each of (1), (2) and (3), given $\theta = (\alpha, \eta)$ and $c$, the baseline model is $S_0 \sim PT(c, \rho, G_{\alpha,\eta})$, where the log-logistic family is parameterized $G_{\alpha,\eta}(t) = 1 - (1 + e^{t(\alpha - 1)})^{-1}$ and the Weibull $G_{\alpha,\eta}(t) = 1 - \exp \left( \left( t / e^{\eta} \right)^{\alpha} \right)$. The weight parameters follow $c \sim \Gamma(5, 1)$ or $c \sim \Gamma(20, 2)$. There are $n = 99$ counties in the state of Iowa; for models with frailty terms $\gamma = (\gamma_1, \ldots, \gamma_n)$, we jointly specify them as conditionally autoregressive with precision parameter $\lambda$ or as multivariate normal with mean zero and covariance $\lambda^{-1}I_n$, where $I_n$ is the $n \times n$ identity matrix, and $\lambda \sim \Gamma(0-1, 0-1)$. This seemingly focused prior on $\lambda$ actually induces a vague conditional prior on the county effect $e^{\gamma}$ relative to the overall mean neighbouring county effect $e^{\bar{\gamma}}$ under the conditionally autoregressive model. For Iowa state, three counties have only $n_i = 2$ Iowan neighbours and one has $n_i = 7$ neighbours. The induced conditional prior for $\gamma_i - \bar{\gamma}_i$ yields a 95% prior credible interval of $(-10^{12}, 10^{12})$ when $n_i = 2$ and $(-10^{11}, 10^{11})$ when $n_i = 7$; the county effect relative to that obtained from the neighbours’ average, $e^{\gamma_i - \bar{\gamma}_i}$ has then a 95% prior credible interval of $(e^{-10^{11}}, e^{10^{11}})$. Similarly, under the independent and identically distributed frailty model, $\lambda \sim \Gamma(0-1, 0-1)$ yields a 95% prior credible interval of $(-10^{12}, 10^{12})$ for $\gamma_i - E(\gamma_i)$.

Given $(\gamma_i, \beta, \gamma, \theta)$ and covariates $x_{ij}$, the density for a survival time is denoted by $p(x_{ij} \mid \gamma_i, \beta, \gamma, \theta)$. Frailty models add $\gamma$ and $\lambda$, and the mixture of Polya tree parameters $p(\gamma \mid c)$ follow the product of $2^J - 2$ beta densities. Define the likelihood as

$$ L(\gamma, \beta, \gamma, \theta) = \prod_{i=1}^{n_i} \prod_{j=1}^{n_i} p_{x_{ij}}(t_{ij} \mid \gamma_i, \beta, \gamma, \theta)^{\delta_{ij}} S_{x_{ij}}(t_{ij} \mid \gamma_i, \beta, \gamma, \theta)^{1-\delta_{ij}}. $$

The posterior density, given data $D$, $p(\gamma, \beta, \gamma, \theta, \lambda \mid D)$, is thus proportional to

$$ L(\gamma, \beta, \gamma, \theta) p(\theta, \beta) p(\gamma \mid \lambda) p(\lambda) p(\gamma) p(\gamma \mid c) p(c). $$

We assume a vague but proper hyperprior $(\theta, \beta) \sim N(p+2(m, S)$, with the restriction $\alpha > 0$, where the notation denotes a multivariate normal distribution of dimension $(p + 2)$ with mean $m$ and covariance $S$. Here $m = 0$ and $S = 1000I_{p+2}$. Prior information could be elicited and implemented based on the parametric model (Bedrick et al., 2000).

The Markov chain Monte Carlo scheme successively samples each parameter given current values of the remaining parameters and the data $D$. The basic parametric survival model without frailties includes parameters $(\beta, \theta)$. Frailty models add $\gamma$ and $\lambda$, and the mixture of Polya trees prior adds conditional probabilities $\gamma$.

The frailties $\gamma_1, \ldots, \gamma_n$ given the remaining parameters are sampled using Metropolis–Hastings steps. For the conditionally autoregressive prior, the candidate $\gamma_i^* \sim N(\bar{\gamma}_i, (\lambda m_i)^{-1})$ is sampled; for independent and identically distributed $\gamma$, the candidate is $\gamma_i^* \sim N(\bar{\gamma}, (\lambda n)^{-1})$. In either case,
\[ \gamma^*_i \text{ is accepted with probability} \]
\[ \min \left\{ 1, \frac{\prod_{j=1}^n p_{x_{ij}}(t_{ij} | y^*_i, \gamma, \beta, \theta, \delta_j) s_{x_{ij}}(t_{ij} | y^*_i, \gamma, \beta, \theta, \delta_j)^{1-\delta_j}}{\prod_{j=1}^n p_{x_{ij}}(t_{ij} | y_i, \gamma, \beta, \theta, \delta_j) s_{x_{ij}}(t_{ij} | y_i, \gamma, \beta, \theta, \delta_j)^{1-\delta_j}} \right\}. \]

The hyperparameter \( \lambda \) is sampled according to its full conditional distribution. Under the conditionally autoregressive prior, \( \lambda \sim \Gamma(0.1 + 0.5(n - 1), 0.1 + 0.5 \sum_{i=1}^n n_i \gamma_i (y_i - y^*_i)). \) Under the exchangeable Gaussian frailty model, we instead have \( \lambda \sim \Gamma(0.1 + 0.5n, 0.1 + 0.5 \sum_{i=1}^n y_i^2). \)

The survival and centring distribution parameters (\( \beta, \theta \)) are updated as a block. These correspond to parameters in the simple survival model that can accordingly provide good starting values for the Markov chain. Additionally, the scaled inverse Fisher information matrix from fitting the parametric nonfrailty model provides an initial multivariate normal Metropolis–Hastings random walk proposal that is refined by running a Metropolis–Hastings random walk sampler for the full model for 5000 steps. The resulting empirical covariance matrix, \( V \), say, is then used as a scaled Metropolis–Hastings proposal covariance matrix for (\( \beta, \theta \)). The candidates (\( \beta^*, \theta^* \)) \( \sim \mathcal{N}_{p+2}(\beta, \theta, kV) \) are accepted with probability
\[ \min \left\{ 1, \frac{\mathcal{L}(y, \beta^*, \gamma, \theta^*) p(\theta^*, \beta^*)}{\mathcal{L}(y, \beta, \gamma, \theta) p(\theta, \beta)} \right\}. \]

The nonparametric part of the model is encapsulated in elements of \( \mathcal{Y} \), which are updated using simple random walk Metropolis–Hastings steps. Each candidate pair is sampled as \( (Y^*_{e0}, Y^*_{e1}) \sim \text{Dirichlet}(m(Y_{e0}, Y_{e1})) \) and accepted with probability
\[ \min \left\{ 1, \frac{\Gamma(m Y_{e0}) \Gamma(m Y_{e1}) (Y_{e0})^{m Y^*_{e0} - c j^2} (Y_{e1})^{m Y^*_{e1} - c j^2} \mathcal{L}(y, \beta, \gamma^*, \theta)}{\Gamma(m Y^*_{e0}) \Gamma(m Y^*_{e1}) (Y_{e0})^{m Y^*_{e0} - c j^2} (Y_{e1})^{m Y^*_{e1} - c j^2} \mathcal{L}(y, \beta, \gamma, \theta)} \right\}, \]

where \( \mathcal{Y}^* \) is the set \( \mathcal{Y} \) with \( (Y^*_{e0}, Y^*_{e1}) \) replacing \( (Y_{e0}, Y_{e1}); m = 20 \) worked well. Finally, the weight \( c \) is sampled via a random walk Metropolis–Hastings step: \( c^* \sim N(c, \tau^2) \), where \( \tau^2 \) is chosen from 2 to 6 such that the acceptance rate is between 30% and 60%. The acceptance probability is
\[ \min \left\{ 1, \frac{\prod_{(Y_{e0}, Y_{e1}) \in \mathcal{Y}} (Y_{e0})^{c^* j^2} (Y_{e1})^{c^* j^2} \Gamma(2c^* j^2)(\Gamma(c^* j^2))^2 (c^* \gamma)^{a-1} e^{-bc^*}}{(Y_{e0})^{c j^2} (Y_{e1})^{c j^2} \Gamma(2c j^2)(\Gamma(c j^2))^2 c \gamma^{a-1} e^{-bc}} \right\}, \]

where, in the last two expressions, \( j \) is the number of digits of \( \epsilon \).

3. Data analysis

3.1. Data description

The Surveillance, Epidemiology, and End Results database provides survival data on a cohort of breast cancer patients observed progressively through time for certain U.S. states. We analyzed a cohort of 1073 Iowan women, who were diagnosed with malignant breast cancer starting in 1995, and enrollment and follow-up continued through the end of 1998. Only deaths that were identified as being due to metastasis of cancerous nodes in the breast were considered to be events, while the remainder, including death from metastasis of other types of cancer or from other causes, were considered to be censored observations. By the end of 1998, 488 of the patients had died of breast cancer, while the remaining 585 women were censored, either because they survived until the end of the study period, died of other causes or were lost to follow-up.
Table 1. Summary statistics for breast cancer data

| Continuous variables | Mean | Std |
|----------------------|------|-----|
| Follow-up time in months | 20·81 | 3·4 |
| Age in years | 68·8 | 15·8 |

| Categorical variables | Level | Count | Percentage |
|-----------------------|-------|-------|------------|
| Status | Event | 488 | 45·5 |
| | Censored | 585 | 54·5 |
| Race | White | 1069 | 99·6 |
| | Black | 4 | 0·4 |
| Number of primaries | 1 | 953 | 88·8 |
| | 2 | 111 | 10·3 |
| | 3 | 9 | 0·8 |
| Stage | Local | 510 | 47·5 |
| | Regional | 355 | 33·1 |
| | Distant | 208 | 19·4 |

Std, standard deviation.

For each individual, the dataset records the survival time in months, from 1 to 48, and her county of residence at diagnosis. Several individual-level covariates are also available, including race, white or black, age in years at diagnosis, number of primaries, i.e. physiologically independent cancers diagnosed, and the stage of the disease: local, regional or distant. Local is confined to the breast, regional means spread beyond the breast tissue and distant implies metastasis. We treat local as the baseline and create two dummy variables for regional and distant, respectively. Table 1 shows several summary statistics for our dataset. Since there are insufficient sample sizes for some levels of the race and number of primaries, we do not include these in our analysis. Thus we include only the two-stage dummies and the centred age covariates.

3·2. Model selection

We compare the predictive ability of competing models using the log pseudo marginal likelihood originally suggested by Geisser & Eddy (1979). The conditional predictive ordinate statistic for the $ij$th observation is

$$CPO_{ij} = px_{ij}(t_{ij} \mid D_{(-ij)})^{\delta_{ij}} S_{x_{ij}}(t_{ij} \mid D_{(-ij)})^{1-\delta_{ij}},$$

where $t_{ij}$ denotes the response for the $ij$th observation, and $D_{(-ij)}$ denotes the data with the $ij$th observation held out. Thus $CPO_{ij}$ is the marginal posterior predictive density or survival function of the observed $t_{ij}$ given the remaining data $D_{(-ij)}$. If $CPO_{ij}$ is larger under one model than another, then datum $ij$ is better supported under that model. The product of these conditional density values has been termed a pseudo-marginal likelihood, and gives an aggregate summary of a model’s predictive utility. The log pseudo marginal likelihood is simply the log of this measure:

$$LPML = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log(CPO_{ij}).$$

Carlin and Louis (2008) suggest that differences in the deviance information criterion (Spiegelhalter et al., 2002) that are less than five are hardly worth mentioning, with differences greater than ten perhaps decisively indicating a preferred model in terms of prediction. Analogously, we look for differences in $LPML$ that exceed, say, 5 or 10 in magnitude.
Table 2. Log pseudo marginal likelihood for the three survival models, with the mixtures of Polya trees centred at the log-logistic or Weibull baseline. The parametric model is obtained when $c \to \infty$.

| Model           | $c$ prior | PH Log-logistic | Weibull AFT Log-logistic | Weibull Log-logistic | PO Log-logistic |
|-----------------|-----------|-----------------|--------------------------|----------------------|-----------------|
| CAR frailty     | $\Gamma(5, 1)$ | $-25.8$         | $-24.5$                  | $-31.1$              | $-25.2$         | $-9.2$          | $-8.7$         |
|                 | $\Gamma(20, 2)$ | $-26.1$         | $-28.2$                  | $-33.8$              | $-26.3$         | $-12.7$         | $-12.0$        |
|                 | $c \to \infty$ | $-33.0$         | $-40.6$                  | $-33.1$              | $-29.6$         | $-20.9$         | $-29.5$        |
| iid frailty     | $\Gamma(5, 1)$ | $-28.2$         | $-25.8$                  | $-31.7$              | $-26.2$         | $-12.5$         | $-11.9$        |
|                 | $\Gamma(20, 2)$ | $-27.7$         | $-29.1$                  | $-37.6$              | $-27.9$         | $-15.9$         | $-15.2$        |
|                 | $c \to \infty$ | $-34.8$         | $-42.3$                  | $-34.9$              | $-32.5$         | $-23.2$         | $-32.4$        |
| Nonfrailty      | $\Gamma(5, 1)$ | $-44.2$         | $-40.1$                  | $-40.7$              | $-34.7$         | $-23.6$         | $-22.7$        |
|                 | $\Gamma(20, 2)$ | $-44.3$         | $-41.5$                  | $-43.0$              | $-35.9$         | $-24.9$         | $-24.5$        |
|                 | $c \to \infty$ | $-47.7$         | $-54.8$                  | $-47.9$              | $-39.5$         | $-30.8$         | $-39.2$        |

Tabled values have 2200 added. PH, proportional hazards; AFT, accelerated failure time; PO, proportional odds; CAR, conditionally autoregressive; iid, independent and identically distributed.

3.3. Results for the breast cancer data

For each of proportional hazards, accelerated failure time and proportional odds (1)–(3), we fitted a model that has a mixture of Polya trees prior on baseline survival centred at the log-logistic or Weibull family, with conditionally autoregressive spatial frailty terms. Various simpler, competing models were also fitted including models with exchangeable frailties, no frailties and versions with a parametric log-logistic or Weibull baseline survival distribution. All models were fitted using variants of the algorithm in § 2.4 implemented in Fortran 90. Despite the high dimension of our models, the Markov chain mixed reasonably well. For each model, we retained 100 000 iterations for posterior estimation following a burn-in of 50 000 iterations.

Table 2 reports the log pseudo marginal likelihood statistics for the competing survival models under both parametric centring families. The more nonparametric mixture of Polya trees model is always preferred over the underlying parametric model; the difference is greatest for proportional odds models, slightly less for proportional hazards and much less for accelerated failure time models. For mixture of Polya trees models, the log-logistic and Weibull centring families perform about equally well for proportional odds and proportional hazards; the log-logistic family is slightly preferred for all accelerated failure time models. For strictly parametric models, the Weibull family is preferred by proportional odds and proportional hazards.

Across all mixture of Polya trees models, proportional odds provides significantly better prediction as measured by LPML, with differences in the range of 10–20. For the parametric log-logistic family, the LPML statistics are almost identical under proportional odds and accelerated failure time models, because the intersection of these two models is the log-logistic regression model. Similarly, the intersection of proportional hazards and accelerated failure time is the Weibull regression model; almost identical LPML statistics are observed for the parametric Weibull family under these two different assumptions. In fact, for data that follow either Weibull or log-logistic regression models, the full mixture of the Polya trees modelling approach will not be able to distinguish between, for example, accelerated failure time and proportional hazards in the Weibull case, because neither generalization is needed beyond the simple parametric regression model.

Within the context of a given survival model, the frailty models significantly improve prediction. Although the conditionally autoregressive is slightly preferred over independent and identically distributed frailties, differences in LPML are only in the range of 1–3.
common disease stage, a twenty-year increase in age at diagnosis is associated with a factor of
exp(0.017) ≈ 20 times longer than a woman with regional stage, and exp(1.019) • 20 times longer than a woman with distant stage. In the same county, a woman with local stage of malignant breast cancer typically survives nearly so, in each model. Higher age at diagnosis increases the hazard; e.g. a twenty-year increase in age at diagnosis increases the hazard rate of women of the same age who live in the same county will be exp(0.017) • 21 times larger if detected at the regional stage, and exp(1.019) • 20 times longer than a patient who has the same stage of disease and lives in the same county. Among patients of the same age and living in the same county who have local stage of disease as the reference, the hazard rate of women of the same age who live in the same county will be exp(0.017) • 21 times larger if their cancer is detected at the regional stage, and exp(1.019) • 20 times longer than a woman with distant stage.

Table 3 provides the posterior medians and equal-tailed 95% credible intervals for main effects under the full proportional hazards, accelerated failure time and proportional odds models. These and all subsequent inferences given in Table 3 and Figs. 2, 3 and 4 are under the log-logistic centring family; differences between Weibull and log-logistic model parameters and estimated survival and density curves are negligible for the c • 9(5, 1) model. All effects are significant, or nearly so, in each model. Higher age at diagnosis increases the hazard; e.g. a twenty-year increase in age is associated with an exp(0.018 • 20) • 43-fold increase in hazard. Using women with local stage of disease as the reference, the hazard rate of women of the same age who live in the same county will be exp(0.22) • 1.25 times larger if their cancer is detected at the regional stage, and exp(1.65) • 5.21 times larger if detected at the distant stage.

Turning to the accelerated failure time assumption, a patient who is twenty years younger typically has a mean lifetime exp(0.017 • 20) • 1.40 times longer than a patient who has the same stage of disease and lives in the same county. Among patients of the same age and living in the same county, a woman with local stage of malignant breast cancer typically survives exp(0.18) • 1.20 times longer than a woman with regional stage, and exp(1.49) • 4.44 times longer than a woman with distant stage.

Finally, for the proportional odds model, for women living in the same county and having common disease stage, a twenty-year increase in age at diagnosis is associated with a factor of exp(−(−0.030) • 20) • 1.82 greater odds of dying from breast cancer before any time t. After
adjusting for the age at diagnosis and the county of residence, the odds of dying from breast cancer before any time \( t \) are \( \exp\{-(-0.47)\} \approx 1.60 \) greater for regional stage versus local stage, and are \( \exp\{-(-2.68)\} \approx 14.59 \) greater for distant stage versus local stage.

These findings are confirmed by Fig. 2, which shows the fitted survival densities for women aged 68.8 years at study entry, the median in our dataset, for three disease stages under the three competing mixtures of Polya trees conditionally autoregressive models, and assuming a spatial frailty of zero. These fitted densities are overlaid on histograms of the observed survival times for study participants with entry ages 58.8–78.8. Since 585 of our 1073 observations are censored, to incorporate both the censored and uncensored observations, we take the Kaplan–Meier survival function estimates and convert them back to an approximate histogram (Huzurbazar, 2005). In all three plots, the predicted density curves from the proportional odds model best mirror the data. These trends are also consistent with the LPML values in Table 2.

Table 3 further compares posterior medians and equal-tailed 95% credible intervals for main effects under the full proportional hazards, accelerated failure time and proportional odds models with those obtained under corresponding parametric models with and without independent and identically distributed frailties. The standard results were obtained using the survival package in R 2.3. As is often the case with main effects, which are typically well identified, the estimates change little across models.

Figure 3 offers a geographic summary of the overall fitted spatial frailty pattern for our best-fitting model, the proportional odds mixtures of Polya trees with conditionally autoregressive frailties. We see clusters of high frailty, and so poorer survival, in the southwest, northeast, central and east-central parts of the state. This map is essentially spatially smoothed version of the raw data map in Fig. 1.

We compare the hazard ratios for two age groups and two stages across the proportional hazards, proportional odds and accelerated failure time models in two counties with disparate observed experience, Mahaska and Mills. Mahaska county had two events out of 26 diagnoses, while Mills county had six events out of just nine diagnoses. Figure 4 shows the predictive hazard ratio for ages 88.8 and 68.8 years, for distant versus local stage in Mills and Mahaska counties. These ratios are constant across time in the proportional hazards model by construction, but decreasing for proportional odds and irregular for accelerated failure time. The rate of decrease in the proportional odds case is concave up for Mills, but concave down for Mahaska. The mixtures of Polya trees conditionally autoregressive proportional odds model thus offers appealing estimates of how the relative status of two groups changes over time and county.
Fig. 4. Hazard ratio for ages 88-8 versus 68-8 years (left), and for distant versus local stage (right) in Mills (upper) and Mahaska (lower) counties from the three competing conditionally autoregressive mixtures of Polya trees models: proportional odds (solid), proportional hazards (dashed) and accelerated failure time (dotted).

4. DISCUSSION

In this paper we have developed flexible survival models for time-to-event data that incorporate spatially varying or independent and identically distributed frailties. The models assume the same nonparametric baseline $S_0$ centred at a parametric family of distributions. Often in the literature for these type of models, the focus is on either the frailty structure or the choice of nonparametric baseline. A variant of the proportional hazards model is often chosen for the survival part of the model. Our findings for the breast cancer data indicate that three model aspects are important for predicting patient survival: the conditional survival model itself, a nonparametric or parametric baseline $S_0$ and whether frailties are included. Of lesser importance is the choice of conditionally autoregressive versus exchangeable frailties for these data.

The conditionally autoregressive model imposes more spatial structure on the frailty terms, smoothing a particular county’s frailty towards neighbouring values, while the independent and identically distributed frailties are simply shrunk towards the statewide global mean. As in many geographically oriented data settings, this seems to have led to slight predictive overfitting of the data by our unstructured frailty models. For the proportional odds mixtures of Polya trees models, the posterior correlations between the 222 unique pairs of neighbouring county frailties from the conditional autoregressive models have median 0.44 and 2.5th and 97.5th percentiles (0.31, 0.54), higher than those from the independent and identically distributed frailty models, with median 0.02 and percentiles (−0.06, 0.17).
To check the robustness of the prior assumptions, we also employed an improper flat prior for \((\theta, \beta)\). All posterior inferences, including estimated regression effects, frailties and LPML statistics, were virtually identical.

Hanson (2006) observed a levelling off of log pseudo marginal likelihood across a variety of models as the Polya tree level increased from \(J\) to \(J + 1\). We used \(J = 4\) for all the models reported in this paper, but also experimented with \(J = 5\) and \(J = 6\). Taking \(J = 5\) led to some improvement in LPML score, but roughly doubled the necessary computing time. Taking \(J = 6\) gave sufficiently poor Markov chain Monte Carlo convergence that we felt results could not be reliably reported.

ACKNOWLEDGEMENT

This work was supported by the U.S. National Institutes of Health. The authors thank the referees and associate editor for suggestions that improved the paper.

REFERENCES

BANERJEE, S. & CARLIN, B. P. (2002). Spatial semiparametric proportional hazards models for analyzing infant mortality rates in Minnesota counties. In Case Studies in Bayesian Statistics, Vol. VI, Ed. C. Gatsonis, R. E. Kass, A. Carriquiry, A. Gelman, D. Higdon, D. K. Pauler & I. Verdinelli, pp. 137–51. New York: Springer.
BANERJEE, S. & CARLIN, B. P. (2003). Semiparametric spatio-temporal frailty modelling. Environmetrics 14, 523–35.
BANERJEE, S., CARLIN, B. P. & GELFAND, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data. Boca Raton, FL: Chapman and Hall/CRC Press.
BANERJEE, S. & DEY, D. K. (2005). Semi-parametric proportional odds models for spatially correlated survival data. Lifetime Data Anal. 11, 175–91.
BANERJEE, S., WALL, M. M. & CARLIN, B. P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics 4, 123–42.
BARKER, P. & HENDERSON, R. (2004). Modelling converging hazards in survival analysis. Lifetime Data Anal. 10, 263–81.
BEDRICK, E. J., CHRISTENSEN, R. & JOHNSON, W. O. (2000). Bayesian accelerated failure time analysis with application to veterinary epidemiology. Statist. Med. 19, 221–37.
BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Statist. Soc. B 36, 192–236.
BESAG, J., YORK, J. C. & MOLLIE, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math. 43, 1–59.
CARLIN, B. P. & LOUIS, T. A. (2008). Bayesian Methods for Data Analysis, 3rd ed. Boca Raton, FL: Chapman and Hall/CRC Press.
CHEN, Y. Q. & JEWELL, N. P. (2001). On a general class of semiparametric hazards regression models. Biometrika 88, 687–702.
CLAYTON, D. & CZUZICK, J. (1985). Multivariate generalizations of the proportional hazards model. J. R. Statist. Soc. A 148, 82–117.
FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. Ann. Statist. 2, 615–29.
GEISSER, S. & EDDY, W. F. (1979). A predictive approach to model selection. J. Am. Statist. Assoc. 74, 153–60.
GORE, S. M., POCOCK, S. J. & KERR, G. R. (1984). Regression models and non-proportional hazards in the analysis of breast cancer survival. Appl. Statist. 33, 176–95.
GUSTAFSON, P. (2007). On robustness and model flexibility in survival analysis: transformed hazards models and average effects. Biometrics 63, 69–77.
HANSON, T. E. (2006). Inference for mixtures of finite Polya tree models. J. Am. Statist. Assoc. 101, 1548–65.
HANSON, T. E. & JOHNSON, W. O. (2002). Modeling regression error with a mixture of Polya trees. J. Am. Statist. Assoc. 97, 1020–33.
HANSON, T. E. & YANG, M. (2007). Bayesian semiparametric proportional odds models. Biometrics 63, 88–95.
HENNERFEIND, A., BREZGER, A. & FAHRMEIR, L. (2006). Geoadditive survival models. J. Am. Statist. Assoc. 101, 1065–75.
HUZURBAZAR, A. V. (2005). A censored data histogram. Commun. Statist. Simul. 34, 113–20.
JEONG, J.-H., JUNG, S.-H. & WIEAND, S. (2003). A parametric model for long-term follow-up data from phase III breast cancer clinical trials. Statist. Med. 22, 339–52.
Jin, X. & Carlin, B. P. (2005). Multivariate parametric spatio-temporal models for county level breast cancer survival data. *Lifetime Data Anal.* **11**, 5–27.

Laíne, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* **20**, 1222–35.

Li, Y. & Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. *J. Am. Statist. Assoc.* **101**, 591–603.

Li, Y. & Ryan, L. (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics* **58**, 287–97.

Mallick, B. K. & Walker, S. G. (2003). A Bayesian semiparametric transformation model incorporating frailties. *J. Statist. Plan. Iner.* **112**, 159–74.

Martinussen, T. & Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. New York: Springer.

Scharfstein, D. O., Tsiatis, A. A. & Gilbert, P. B. (1998). Efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Anal.* **4**, 355–91.

Spiegelhalter, D. J., Best, N., Carlin, B. P. & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc. B* **64**, 583–639.

Vaupel, J. W., Manton, K. G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–54.

Walker, S. G. & Mallick, B. K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. R. Statist. Soc. B* **59**, 845–60.

Walker, S. G. & Mallick, B. K. (1999). Semiparametric accelerated life time model. *Biometrics* **55**, 477–83.

Yakovlev, A. Y., Tsodikov, A. D., Boucher, K. & Kerber, R. (1999). The shape of the hazard function in breast carcinoma: curability of the disease revisited. *Cancer* **85**, 1789–98.

Yin, G. & Ibrahim, J. G. (2005). A class of Bayesian shared gamma frailty models with multivariate failure time data. *Biometrics* **61**, 208–16.

Zhang, M. & Davidian, M. (2008). “Smooth” semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics* **64**, 567–76.

[Received August 2007. Revised September 2008]