Mixtures of Polya Trees for Flexible Spatial Frailty Survival Modeling

Luping Zhao, Timothy E. Hanson, and Bradley P. Carlin[1]

MMC 303, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455–0392, U.S.A.

Correspondence author: Timothy E. Hanson
telephone: (612) 626-7075
fax: (612) 626-0660
email: hanson@biostat.umn.edu

May 3, 2007

[1]Luping Zhao is Graduate Assistant, Timothy E. Hanson is Associate Professor, and Bradley P. Carlin is Mayo Professor in Public Health in the Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, 55455. The work of the three authors was supported in part by NIH grant 1–R01–CA95955–01. The authors thank Dr. Sudipto Banerjee for advice and help constructing the data set.

Mixtures of Polya Trees for Flexible Spatial Frailty Survival Modeling

Abstract

Mixtures of Polya trees offer a very flexible, nonparametric approach for modeling time-to-event data. Many such settings also feature spatial association that requires further sophistication, either at a point (geostatistical) or areal (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 thanks to modern Markov chain Monte Carlo (MCMC) methods. We illustrate the usefulness of our proposed methods with an analysis of spatially oriented breast cancer survival data from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute. Our results indicate appreciable advantages for our approach over previous, competing methods that impose unrealistic parametric assumptions, ignore spatial association, or both.

KEY WORDS: Areal data; Bayesian modeling; Breast cancer; Conditionally autoregressive (CAR) model; Log pseudo marginal likelihood (LPML); Nonparametric modeling.

1 Introduction

In survival studies, the hazard function for individuals within certain groups may depend on a set of risk factors, but some of these risk factors may be unknown. Vaupel, Manton, and Stallard (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 (gamma, lognormal, splines, and so on) have been explored in the literature. The frailties are typically assumed independent and identically distributed (i.i.d.) with mean 0, but in the case where the groups corre-

spond to geographic regions (say, counties or zip codes), a spatially associated distribution may be more natural. For example, Banerjee, Wall, and Carlin (2003) developed parametric frailty specifications based on both areal (lattice) and point-referenced (geostatistical) spatial models, and compared them with traditional i.i.d. frailty and no-frailty approaches under a Weibull baseline hazard function in the context of county-level infant mortality data. Banerjee and Carlin (2002, 2003) developed semiparametric Cox frailty models via beta mixture and counting process approaches, and compared the models using the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002).

As just mentioned, spatial frailty models can be either geostatistical or lattice, depending on whether the frailties are indexed to specific geographical coordinates, or only to a discretely-indexed areal map with associated neighbor (spatial adjacency) information. In the geostatistical case, the frailties are typically modeled as mean-zero Gaussian random variables having a nondiagonal covariance matrix $\mathbf{H}$. The entries of $\mathbf{H}$ often depend on the distances between the corresponding locations, and have various forms corresponding to the various variogram models familiar in the spatial statistics literature (spherical, exponential, powered exponential, Matern, and so on; see e.g. Banerjee et al., 2004, pp.50-51).

In the lattice case, the discretely indexed regions instead form a partition of the geographic region being studied. The spatial information in this type of model is usually based on the adjacency of regions, rather than any continuous distance metric. The most commonly used lattice model is the conditionally autoregressive (CAR) model, introduced by Besag (1974). Li and 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, while Banerjee and Dey (2005) fitted proportional odds models, both in spatially correlated survival data settings. Banerjee and Carlin (2003) developed semiparametric spatio-temporal frailty models using hierarchical Bayesian methods. These methods were further extended by Jin and Carlin (2005), who proposed a multivariate conditionally autoregressive (MCAR) model for areally-referenced multiple disease data.

Thanks to recent advances in computing technology, Bayesian approaches to survival

models are now computationally feasible and increasingly popular. In this paper, we consider three models commonly used with survival data: the accelerated failure time (AFT) model, the proportional hazards (PH) model, and the proportional odds (PO) model. All three models provide useful summary information in the absence of an estimate of the baseline survival distribution, and hence are often fit 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.

The PH model is currently the most widely used for survival data with covariates. Kalbfleisch (1978) first considered a gamma process prior for the baseline hazard in the PH model, and noted that the Cox (1972) partial likelihood can emerge as a limiting case of the marginal likelihood obtained from this model. Ibrahim, Chen, and Sinha (2001, pp.47-94) give a detailed description of semiparametric Bayesian PH models with various nonparametric priors, including the gamma process, beta process, correlated prior processes, and the Dirichlet process (DP; Ferguson, 1973). In each case, they provide a development of the prior process, likelihood function, posterior distributions, and the Markov chain Monte Carlo (MCMC) sampling techniques needed for inference. Sinha and Dey (1997) also provide a review of semiparametric approaches to the PH model. Other PH priors include the extended gamma process and a monotone transformation of a mixture of beta densities.

Christensen and Johnson (1988) provided an analysis for the AFT model, obtaining approximate marginal inference under a DP baseline survival function. However, Johnson and Christensen (1989) showed that it was infeasible to perform a full Bayesian analysis for an AFT model with a DP survival baseline due to an combinatorial explosion in the number of possible configurations of ties in baseline data and the associated bookkeeping required. Kuo and Mallick (1997) eliminated this difficulty by considering a Dirichlet process mixture (DPM) of continuous densities as the baseline survival function. The DPM smoothes the DP via a continuous known kernel with unknown mixing weights. Walker and Mallick (1999) and Hanson and Johnson (2002) developed AFT models with Polya tree (PT) and mixtures

of Polya trees (MPT) survival baselines, respectively.

The PO model has recently gained attention as an alternative to the PH and AFT models. Bayesian approaches, which lend themselves naturally to the type of predictive comparisons we desire, include Banerjee and Dey (2005) and Hanson and Yang (2007), who consider this model with CAR and i.i.d. frailties, respectively. Our work expands Banerjee and Dey (2005) by considering several competing survival models, including parametric models, and a richer model for baseline survival. Hanson and Yang (2007) consider MPT and parametric priors for baseline survival in the PO model; although i.i.d. frailties are developed, they are ultimately not used in data analyses. In contrast we compare several aspects of modeling: (1) choice of AFT, PH, or PO; (2) two types of frailty model or absence thereof, and (3) parametric versus nonparametric assumptions on baseline survival S_0.

Although the DP model is widely used, it is intractable in the PH and PO settings, as is the DPM prior. Ibrahim, Chen, and Sinha (2001, p.94) noted that, "Dirichlet processes are quite difficult to work with in the presence of covariates, since they have no direct representation through either the hazard or cumulative hazard function." Similar problems occur with gamma and beta priors, since they are tailored for use in modeling baseline hazard and integrated hazard functions and are therefore attractive for use in variants of the PH model, but difficult to implement elsewhere (e.g., the AFT model).

The PH, AFT, and PO 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 MPT baseline hazard can be taken to be the same across the three models, placing them on common ground. Differences in predictive performance can therefore be attributed to the *survival* and *frailty* models only, rather than to additional possible differences in quite different (e.g., absolutely continuous versus discrete) nonparametric priors. For the SEER data, we found the PO model to be predictively superior to the PH model. Recently, Li and Lin (2006) and Hennerfeind et al. (2006) proposed highly flexible spatial frailty models for survival analysis. However, both developed models assuming PH and alternative specifications were not considered.

Superior aspects of alternatives to PH (Cox, 1972), such as AFT and PO, have been well argued by many authors (e.g. Wei, 1992; Hutton and Monaghan, 2002; Portnoy, 2003). However, often only empirical measures of model fit such as plots of fitted survival curves or fitted quantile functions are compared across models. Predictive measures, such as the LPML we use, are either impractical or impossible to obtain, especially in models employing rank-based procedures. The distinction between model fit and prediction is important, as it is often the case that a highly parameterized model may fit a given data set very well but is terrible at predicting future data. This is problematic when the main locus of inference is *precisely* the prediction of future survival given a collection of risk factors.

Several interesting "super models" have been proposed, including transformation models that include PH and PO as special cases (e.g. Scharfstein, Tsiatis, and Gilbert, 1998; Mallick and Walker, 2003), transformation and extended regression models that include PH and additive hazards as special cases (e.g. Yin and Ibrahim, 2005; Martinussen and Scheike, 2006, Chapter 7), and hazard regression models that include both PH and AFT as special cases (e.g. Chen and Jewell, 2001). While highly flexible, these models all suffer in that, once fit, the resulting regression parameters lose any simple interpretability. Furthermore, there may not be sufficient information to estimate the additional transformation and regression parameters included in the models. It would seem that many of these approaches are better suited toward testing 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 and Ibrahim (2005). We instead emphasize model interpretability and selection over what essentially amounts to model averaging.

We illustrate our proposed spatial MPT methodology using a subset of the Surveillance, Epidemiology, and End Results (SEER) cancer database, as maintained by the National Cancer Institute (NCI); see `seer.cancer.gov`. These data were previously analyzed by several authors (Banerjee, Wall, and Carlin, 2003; Jin and Carlin, 2005) in the context of a proportional hazards spatial frailty model, and are comprised of the survival times in months

for women diagnosed with breast cancer in the state of Iowa during 1995–1998. Important predictors of survival are well-established in the literature, and 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 log-standardized mortality ratio (SMR), 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.158-159). Note that while there is substantial statewide variability, there does appear to be some local similarity of the rates in neighboring counties, with clusters of elevated SMRs in the east and southwest.

The remainder of our paper is organized as follows. Section 2 gives a detailed description of our statistical models, including computational details related to MCMC implementation of our Polya tree mixtures. Section 3 then offers a detailed analysis of our SEER dataset, including model comparison, parameter estimation, and mapping of smoothed, county-specific fitted rates. Finally, Section 4 discusses our findings, and offers directions for future work in this area.

2 Statistical models

In the frailty literature, competing survival models are rarely considered. This is true in the burgeoning joint longitudinal and survival literature as well. Often the survival portion of a model is more or less "picked" *a priori*, and is often chosen to be a variant of the Cox proportional hazards model, based largely on the considerable momentum this model has gained in the literature. Important predictors are assessed relative to the chosen survival model, and issues concerning the overall fit and predictive ability of the survival model relative to other models is often either ignored or only briefly addressed, perhaps under concluding remarks.

We wish to explore overall survival model choice in tandem with the nonparametric prior and spatial frailty assumptions required by our geographically referenced data. For the SEER data we consider here, we find that the survival model is of greatest importance, and

that both the frailty model and aspects of the nonparametric prior can also markedly affect predictive ability.

2.1 Survival modeling

All three models (PH, AFT, and PO) can be formulated in terms of the baseline survival function S_0. Let $\mathbf{x}_{ij}$ be a p-dimensional vector of explanatory covariates associated with the j^{th} individual in group i, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, and let $S_{\mathbf{x}_{ij}}(\cdot)$ 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 γ_i.

The proportional hazards model assumes

$$S_{\mathbf{x}_{ij}}(t) = S_0(t)^{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta}+\gamma_i)}, \tag{1}$$

while the accelerated failure time model assumes

$$S_{\mathbf{x}_{ij}}(t) = S_0\{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i)t\}, \tag{2}$$

and the proportional odds model assumes

$$\frac{S_{\mathbf{x}_{ij}}(t)}{1 - S_{\mathbf{x}_{ij}}(t)} = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i)\frac{S_0(t)}{1 - S_0(t)}. \tag{3}$$

In each model, $e^{(\mathbf{x}_1-\mathbf{x}_2)'\boldsymbol{\beta}}$ has a useful interpretation, comparing relative risk at any time t (PH); the relative mean, median, or any survival quantile (AFT); or the relative odds of surviving past any time t within a county (PO), between individuals with covariates $\mathbf{x}_1$ and $\mathbf{x}_2$. The factor $e^{\gamma_{i_1}-\gamma_{i_2}}$ compares county-level risks for any given set of covariates between counties i_1 and i_2.

2.2 Spatial frailty modeling

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

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

where γ_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, n_i ranges from 2 to 7. We adopt a vague but proper gamma hyperprior distribution for λ, as in Banerjee, Wall and Carlin (2003). These authors show this model to perform similarly to geostatistical alternatives, but in a fraction of the computer time since it avoids inverting large matrices within each MCMC iteration.

2.3 MPT priors for the baseline survival function

We consider models (1), (2), and (3) with an MPT prior on S_0. An MPT smoothes over partitioning effects associated with a simple Polya tree, and the mixture model includes the underlying centering parametric families as special cases.

Consider a mixture of Polya trees prior on S_0,

$$S_0|\boldsymbol{\theta} \sim PT(c, \rho, G_{\boldsymbol{\theta}}), \quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{4}$$

where (4) is shorthand for a particular parameterization (Hanson and Johnson, 2002; Hanson, 2006). We describe the prior below but leave some technical details to these references and Lavine (1992). Broadly, the baseline S_0 is centered at a parametric family $G_{\boldsymbol{\theta}}$ (e.g. log-logistic) by partitioning $\mathbb{R}^+$ into 2^J intervals of equal probability under $G_{\boldsymbol{\theta}}$, denoted $\{B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_J) : \epsilon_1 \cdots \epsilon_J \in \{0,1\}^J\}$ and placing a particular prior probability on these sets $p_{\mathcal{Y}}(j) = G\{B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_J)\}$ where j is one plus the base-10 representation of binary $\epsilon_1 \cdots \epsilon_J$.

Let J be a fixed, positive integer and let $G_{\boldsymbol{\theta}}$ denote a family of cumulative distribution functions indexed by $\boldsymbol{\theta}$. The distribution $G_{\boldsymbol{\theta}}$ serves to center the random distribution S_0. A Polya tree prior is constructed from a set of nested partitioning sets $\Pi_{\boldsymbol{\theta}} = \{B_{\boldsymbol{\theta}}(\epsilon) : \epsilon \in \bigcup_{l=1}^{J}\{0,1\}^l\}$ and corresponding conditional probabilities $\mathcal{Y} = \{Y_\epsilon : \epsilon \in \bigcup_{l=1}^{J}\{0,1\}^l\}$. A set $B_{\boldsymbol{\theta}}(\epsilon) = B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_k)$ at partition level k is split into two sets: $B_{\boldsymbol{\theta}}(\epsilon 0)$ and $B_{\boldsymbol{\theta}}(\epsilon 1)$. Given that an observation from S_0 is in the parent $B_{\boldsymbol{\theta}}(\epsilon)$, $Y_{\epsilon 0}$ is the conditional probability the observation is in $B_{\boldsymbol{\theta}}(\epsilon 0)$ and $Y_{\epsilon 1}$ is the conditional probability the observation is in $B_{\boldsymbol{\theta}}(\epsilon 1)$. Necessarily, $Y_{\epsilon 0} + Y_{\epsilon 1} = 1$. Under the Polya tree prior, $Y_{\epsilon 0} \overset{ind.}{\sim} \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$. That is, pairs of conditional probabilities $(Y_{\epsilon 0}, Y_{\epsilon 1})$ in $\mathcal{Y}$ are distributed independent Dirichlet with parameters in $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in \bigcup_{l=1}^{J}\{0,1\}^l\}$. At the coarsest level, $(Y_0, Y_1) = (0.5, 0.5)$ for identifiability.

The partition points are quantiles of the centering family: if j is the base-10 representation

of the binary number $\epsilon = \epsilon_1 \cdots \epsilon_k$ at level k, then $B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_k)$ is defined to be the interval $(G_{\boldsymbol{\theta}}^{-1}(j/2^k), G_{\boldsymbol{\theta}}^{-1}((j+1)/2^k)]$; an exception is that the "rightmost" set is $B_{\boldsymbol{\theta}}(11\cdots 1) = (G_{\boldsymbol{\theta}}^{-1}((2^k-1)/2^k), \infty)$. For example, with $k=3$, and $\epsilon = 000$, then $j=0$ and $B_{\boldsymbol{\theta}}(000) = (0, G_{\boldsymbol{\theta}}^{-1}(1/8)]$, but with $\epsilon = 010$, then $j=2$ and $B_{\boldsymbol{\theta}}(010) = (G_{\boldsymbol{\theta}}^{-1}(2/8), G_{\boldsymbol{\theta}}^{-1}(3/8)]$, etc. At each level k, the class $\{B_{\boldsymbol{\theta}}(\epsilon) : \epsilon \in \{0,1\}^k\}$ forms a partition of the positive reals and furthermore $B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_k) = B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_k 0) \bigcup B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_k 1)$ for any binary $\epsilon_1 \cdots \epsilon_k$.

Given $\boldsymbol{\theta}$ and $\mathcal{A}$, the Polya tree prior is defined up to level J by the random pairs in $\mathcal{Y}$ through the product of conditional probabilities

$$S_0\{B_{\boldsymbol{\theta}}(\epsilon_1 \cdots \epsilon_k)|\mathcal{Y}, \boldsymbol{\theta}\} = \prod_{j=1}^{k} Y_{\epsilon_1 \cdots \epsilon_j}, \tag{5}$$

for $k = 1, 2, \ldots, J$, where we define $S_0(A)$ to be the baseline measure of any set A. This falls out directly from the treelike structure of conditional probabilities on partition sets; see Figure 1 in Ferguson (1974) or the schematic in Hanson and Yang (2007, p.89). For example, given $\mathcal{Y}$ and $\boldsymbol{\theta}$, the S_0-probability of $B_{\boldsymbol{\theta}}(110)$ is $Y_{110}Y_{11}Y_1$ which follows from $B_{\boldsymbol{\theta}}(110) \subset B_{\boldsymbol{\theta}}(11) \subset B_{\boldsymbol{\theta}}(1)$.

The family $\mathcal{A}$ is defined by $\alpha_{\epsilon_1 \cdots \epsilon_k} = ck^2$ for some $c > 0$ (Walker and Mallick, 1997; Walker and Mallick, 1999; Hanson and Johnson, 2002). The prior variability of conditional probabilities, $\text{var}(Y_{\epsilon_1 \cdots \epsilon_k}) = 0.25/(1+2ck^2)$, decreases with the level k at a rate fast enough to ensure the existence of a density in an infinite $J \to \infty$ tree (Ferguson, 1974). The parameter c acts much like the precision in a Dirichlet process and is directly related to how quickly data "take over" the prior. Very large values of c force conditional probabilities Y_ϵ to be close to 0.5 regardless of the data, which further forces $S_0(A) \approx G_{\boldsymbol{\theta}}(A)$ for sets A; as c tends to infinity we obtain a fully parametric analysis. As c tends to zero the posterior baseline is almost entirely data-driven, but this implies essentially zero prior weight on centering family and is problematic from both philosophical and practical viewpoints (Hanson, 2006). In fact, in an infinite $J \to \infty$ Polya tree, as $c \to 0^+$ a predictive density no longer exists and hence the LMPL is undefined.

The Polya tree conditional probabilities $(Y_{\epsilon 0}, Y_{\epsilon 1})$ "adjust" the shape of the survival density f_0 relative to a parametric centering family of distributions. If the data are truly

distributed according to $G_{\boldsymbol{\theta}}$, then observations should be on average evenly distributed among partition sets at any level j. Under the Polya tree posterior, if more observations fall into interval $B_{\boldsymbol{\theta}}(\epsilon 0) \subset \mathbb{R}^+$ than its companion set $B_{\boldsymbol{\theta}}(\epsilon 1)$, the conditional probability $Y_{\epsilon 0}$ of $B_{\boldsymbol{\theta}}(\epsilon 0)$ is stochastically "increased" relative to $Y_{\epsilon 1}$. The Polya tree essentially adds $2^J - 2$ free parameters $\mathcal{Y}$ that allow for focused deviations relative to the parametric model, indexed by $\boldsymbol{\theta}$. However $E(Y_\epsilon) = 0.5$ anchors the random S_0 about the family $\{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Within sets at the level J in $\Pi_{\boldsymbol{\theta}}$, we assume $S_0|\mathcal{Y}, \boldsymbol{\theta}$ follows the baseline $G_{\boldsymbol{\theta}}$ (Hanson, 2006).

Define the vector of probabilities $\mathbf{p}_{\mathcal{Y}} = (p_{\mathcal{Y}}(1), p_{\mathcal{Y}}(2), \ldots, p_{\mathcal{Y}}(2^J))'$ through

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

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

$$S_0(t|\mathcal{Y}, \boldsymbol{\theta}) = p_{\mathcal{Y}}(k_{\boldsymbol{\theta}}(t)) \left[k_{\boldsymbol{\theta}}(t) - 2^J G_{\boldsymbol{\theta}}(t)\right] + \sum_{j=k_{\boldsymbol{\theta}}(t)+1}^{2^J} p_{\mathcal{Y}}(j), \tag{6}$$

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

$$f_0(t|\mathcal{Y}, \boldsymbol{\theta}) = \sum_{j=1}^{2^J} 2^J p_{\mathcal{Y}}(j) g_\theta(t) I_{B_\theta(\epsilon_J(j-1))}(t) = 2^J p_{\mathcal{Y}}(k_{\boldsymbol{\theta}}(t)) g_{\boldsymbol{\theta}}(t), \tag{7}$$

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

The mixture of Polya trees (MPT) prior provides an intermediate choice between a strictly parametric analysis and allowing S_0 to be completely arbitrary. In areas where data are sparse, such as the tails, the MPT prior places relatively more posterior mass on the underlying parametric family $\{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. In areas where data are plentiful the posterior is more data driven, and features not allowed in the strictly parametric model, such as left-skew and multimodality, become apparent. The weight c controls how closely the posterior follows $\{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with larger values of c yielding inference closer to that obtained from the underlying parametric model. Based on previous experience with Polya tree models, we consider two priors on c, the first of which is $c \sim \Gamma(5, 1)$ where $\Gamma(a, b)$ denotes the gamma

distribution with mean a/b and variance a/b^2. This prior places mass on smaller values of c, allowing more flexibility in baseline modeling at the possible expense of predicting future data. The second prior we consider is $c \sim \Gamma(20, 2)$, which places mass on larger values of c in an effort to "smooth" our inferences toward the underlying parametric family. Finally, we also consider the underlying parametric family, obtained as $c \to \infty$. Essentially, we are considering three model specifications: the parametric log-logistic family ($c \to \infty$), a specification that allows some variability about this family ($c \sim \Gamma(20, 2)$), and a specification that allows for substantially more variability ($c \sim \Gamma(5, 1)$).

Although c has the flavor of a smoothing parameter, large values of c smooth in the direction of a fixed parametric family for S_0. Hanson (2006) discusses simulations where there is evidently posterior information for c when the true data distribution is unlike any member of the centering family $\{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. However, when data are evenly distributed among partition sets at level J, i.e. when data are "perfectly" distributed according to $G_{\boldsymbol{\theta}}$, the posterior distribution of c is flat and improper under an improper flat prior. That is, posterior learning cannot take place when the data generating mechanism is very close to the centering family. Berger and Guglielmi (2001) consider testing a univariate parametric family against a Polya tree alternative by finding the value of (essentially) c^{-1} that maximizes the Bayes factor in favor of the Polya tree alternative. We could try this approach were it computationally feasible in the complex spatial/survival models fit herein. Instead we consider the three priors described above to determine the impact of c, and the degree of "parametric-ness" the prior implies, on posterior inference, and in particular prediction. Considering $c \sim \Gamma(5, 1)$ and $c \sim \Gamma(20, 2)$ allows for a bit more data-driven flexibility than fixing $c = 5$ and $c = 10$.

2.4 Summary and computational notes

In this paper we consider an MPT prior on S_0 centered at the log-logistic family of densities. The log-logistic family has tails that die off slower that the Weibull or log-normal families, and we have found it to be numerically stable across the three survival models considered. For the PH model, the log-logistic model also provides slightly better prediction

than the Weibull family. Hanson and Johnson (2002) found the choice of underlying family to make little difference in density estimation with an MPT prior. The number of tree levels was capped at $J = 4$, achieving good MCMC mixing, and allowing the comparison of dozens of models in a reasonable amount of computer time. Adding a level to the tree essentially doubles the number of parameters defining S_0 and hence doubles computation time. Adding levels to the tree allows the MPT to accommodate greater detail, but can also slow MCMC mixing (Hanson, 2006) greatly increasing computational burden, resulting in a "law of diminishing returns." Each model we fit took on the order of an hour to run on a 3.2Ghz Pentium 4 with 2GB of RAM in compiled FORTRAN.

For each of (1), (2), and (3), the baseline model is

$$S_0|\alpha, \eta, c \sim PT(c, \rho, G_{\alpha,\eta}), \quad G_{\alpha,\eta}(t) = 1 - (1 + e^{\eta} t^{\alpha})^{-1},$$

where $c \sim \Gamma(5, 1)$ or $c \sim \Gamma(20, 2)$ and $p(\alpha, \eta) \propto 1$. There are $n = 99$ counties in the state of Iowa; for models with frailty terms we jointly specify them either spatially as

$$\boldsymbol{\gamma}|\lambda \sim \text{CAR}(\lambda),$$

or nonspatially as

$$\boldsymbol{\gamma}|\lambda \sim N(\mathbf{0}_n, \lambda^{-1}\mathbf{I}_n),$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ and $\lambda \sim \Gamma(0.1, 0.1)$. This seemingly focused prior on λ actually induces a vague conditional prior on the county effect e^{γ_i} relative to the overall mean neighboring county effect $e^{\bar{\gamma}_i}$ under the CAR model. For the Iowa SEER data three counties have only $n_i = 2$ (Iowan) neighbors, while one county has $n_i = 7$ neighbors. 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 neighbors' average, $e^{\gamma_i - \bar{\gamma}_i}$ has then a 95% prior credible interval of $(e^{-10^{11}}, e^{10^{11}})$. Similarly, under the i.i.d. 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 $(\gamma_i, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})$ and covariates $\mathbf{x}_{ij}$ the pdf of a survival time is denoted $p_{\mathbf{x}_{ij}}(\cdot|\gamma_i, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})$. The frailty model is further denoted $p(\boldsymbol{\gamma}|\lambda)$ and the Polya tree parameters $p(\mathcal{Y}|c)$ follow

the product of $2^J - 2$ beta densities. We place a flat prior on the remaining parameters $p(\boldsymbol{\beta}, \boldsymbol{\theta}) \propto 1$. Define the likelihood

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} p_{\mathbf{x}_{ij}}(t_{ij}|\gamma_i, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})^{\delta_{ij}} S_{\mathbf{x}_{ij}}(t_{ij}|\gamma_i, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})^{1-\delta_{ij}} \ ,$$

where $\delta_{ij} = 0$ if observation ij is censored, and 1 if not. The posterior density given data $\mathcal{D}$, $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}, \lambda|\mathcal{D})$, is thus proportional to $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}) p(\boldsymbol{\gamma}|\lambda) p(\lambda) p(\mathcal{Y}|c) p(c)$. We adopted the following strategy for sampling the parameters:

- Sampling $\boldsymbol{\gamma}|\boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}, \lambda, \mathcal{D}$:

  We use a Metropolis-Hastings (M-H) step for each γ_i. For $\boldsymbol{\gamma} \sim \text{CAR}(\lambda)$, sample $\gamma_i^* \sim N(\bar{\gamma}_i, (\lambda m_i)^{-1})$; for i.i.d. $\boldsymbol{\gamma}$, sample $\gamma_i^* \sim N(\bar{\gamma}, (\lambda n)^{-1})$. In either case, accept γ_i^* with probability

  $$\min \left\{ 1, \frac{\prod_{j=1}^{n_i} p_{\mathbf{x}_{ij}}(t_{ij}|\gamma_i^*, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})^{\delta_{ij}} S_{\mathbf{x}_{ij}}(t_{ij}|\gamma_i^*, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})^{1-\delta_{ij}}}{\prod_{j=1}^{n_i} p_{\mathbf{x}_{ij}}(t_{ij}|\gamma_i, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})^{\delta_{ij}} S_{\mathbf{x}_{ij}}(t_{ij}|\gamma_i, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})^{1-\delta_{ij}}} \right\},$$

  for $i = 1, \ldots, K$.

- Sampling $\lambda|\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}, \mathcal{D}$:

  Under $\boldsymbol{\gamma}|\lambda \sim \text{CAR}(\lambda)$, sample

  $$\lambda|\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}, \mathcal{D} \sim \Gamma \left(0.1 + 0.5(n-1) \ , \ 0.1 + \sum_{i=1}^{n} n_i (\gamma_i - \bar{\gamma}_i)^2 \right).$$

  Under $\boldsymbol{\gamma}|\lambda \sim N_n(\mathbf{0}, \mathbf{I}_n \lambda^{-1})$, sample

  $$\lambda|\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta}, \mathcal{D} \sim \Gamma \left(0.1 + 0.5n \ , \ 0.1 + \sum_{i=1}^{n} \gamma_i^2 \right).$$

- Sampling $(\boldsymbol{\beta}, \boldsymbol{\theta})|\boldsymbol{\gamma}, \mathcal{Y}, \lambda, \mathcal{D}$:

  We updated $(\boldsymbol{\beta}, \boldsymbol{\theta})$ jointly starting with estimates and an estimated covariance matrix from fitting the parametric non-frailty model. We then "refined" the covariance matrix by running a crude M-H random walk sampler for the full model for 5000 steps. The resulting empirical covariance matrix $\mathbf{V}$ was then used as a scaled (by $k > 0$) M-H proposal covariance matrix for $(\boldsymbol{\beta}, \boldsymbol{\theta})$. We then sample $(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*) \sim N_{p+2}((\boldsymbol{\beta}, \boldsymbol{\theta}), k\mathbf{V})$ and accept these candidates with probability

  $$\min \left\{ 1, \frac{\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}^*, \mathcal{Y}, \boldsymbol{\theta}^*)}{\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})} \right\}.$$

- Sampling $\mathcal{Y}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \lambda, \mathcal{D}$:

  Sample a candidate $(Y^*_{\epsilon 0}, Y^*_{\epsilon 1}) \sim Dirichlet(m(Y_{\epsilon 0}, Y_{\epsilon 1}))$, $\epsilon \in \bigcup_{j=1}^{J-1}\{0,1\}^j$, where we choose $m = 20$ here. Accept the candidate as the new $(Y_{\epsilon 0}, Y_{\epsilon 1})$ with probability

$$\min\left\{1, \frac{\Gamma(mY_{\epsilon 0})\Gamma(mY_{\epsilon 1})(Y_{\epsilon 0})^{mY^*_{\epsilon 0}-cj^2}(Y_{\epsilon 1})^{mY^*_{\epsilon 1}-cj^2}\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}^*, \boldsymbol{\theta})}{\Gamma(mY^*_{\epsilon 0})\Gamma(mY^*_{\epsilon 1})(Y^*_{\epsilon 0})^{mY_{\epsilon 0}-cj^2}(Y^*_{\epsilon 1})^{mY_{\epsilon 1}-cj^2}\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{Y}, \boldsymbol{\theta})}\right\},$$

  where $\mathcal{Y}^*$ is the set $\mathcal{Y}$ with $(Y^*_{\epsilon 0}, Y^*_{\epsilon 1})$ replacing $(Y_{\epsilon 0}, Y_{\epsilon 1})$.

- Sampling $c|\mathcal{Y}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \lambda, \mathcal{D}$:

  Sample $c^* \sim N(c, \tau^2)$, where τ^2 is chosen from 2 to 6 such that the accept rate varied from about 30% to 60%. Accept c^* with probability

$$\min\left\{1, \frac{(Y_{\epsilon 0})^{c^*j^2}(Y_{\epsilon 1})^{c^*j^2}\Gamma(2c^*j^2)(\Gamma(cj^2))^2(c^*)^{a-1}e^{-bc^*}}{(Y_{\epsilon 0})^{cj^2}(Y_{\epsilon 1})^{cj^2}\Gamma(2cj^2)(\Gamma(c^*j^2))^2c^{a-1}e^{-bc}}\right\}.$$

3 Analysis of the Iowa SEER data

As mentioned earlier, the SEER database (`seer.cancer.gov`) provides survival data on a cohort of breast cancer patients observed progressively through time for a collection of US states. Our Iowa breast cancer data were extracted from this cohort, and include information on a cohort of 1073 women in Iowa, who were diagnosed with malignant breast cancer starting in 1995, with enrollment and follow-up continued through the end of 1998. Only deaths which were identified as being due to metastasis of cancerous nodes in the breast were considered to be events, while the rest (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.

For each individual, the dataset records the survival time in months (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 (physiologically independent cancers diagnosed), and the stage of the disease: local (confined to the breast), regional (spread beyond the breast tissue), or distant (metastatis). 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 covariates, we do not include these in our analysis. Thus we include only the two stage dummies and the centered age covariates.

3.1 Model selection

Model comparison is a crucial part of any statistical analysis. Our primary tool here is based on the *log pseudo marginal likelihood* (LPML), originally suggested by Geisser and Eddy (1979). To develop this measure, we begin by defining the *conditional predictive ordinate* (CPO) statistic for the ij^{th} observation as

$$\text{CPO}_{ij} = p_{\mathbf{x}_{ij}}(t_{ij}|\mathcal{D}_{(-ij)})^{\delta_{ij}} S_{\mathbf{x}_{ij}}(t_{ij}|\mathcal{D}_{(-ij)})^{1-\delta_{ij}},$$

where t_{ij} denotes the response for the ij^{th} observation, and $\mathcal{D}_{(-ij)}$ 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 $\mathcal{D}_{(-ij)}$. If CPO_{ij} is larger under one model relative to another, then datum ij is "better supported" or "better predicted," indicating greater predictive ability for the model. Thinking of the product of these conditional density values as a "psuedo marginal likelihood," this gives an aggregate summary measure of fit. The LPML is simply the log of this measure,

$$\text{LPML} = \log\left\{\prod_{i=1}^{n}\prod_{j=1}^{n_i} \text{CPO}_{ij}\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \log(\text{CPO}_{ij}) \ ,$$

the log being added primarily for computational convenience. In the context of survival data, the LPML has been discussed by Gelfand and Mallick (1995) and Sinha and Dey (1997). Unlike Bayes facors, the LPML remains well defined under improper priors (provided the posterior does), and is quite stable computationally. Ibrahim, Chen, and Sinha (2001) offer a detailed discussion of the use of CPO and LPML with survival data.

In a similar vein, the posterior mean deviance, $\bar{D} = E_{\boldsymbol{\theta}|y} D(\boldsymbol{\theta})$, is often used to summarize the information content in a model; see for example Dempster (1997) and Zeger and Karim (1991). Spiegelhalter et al. (2002) recommend use of $\bar{D}$ as a measure of Bayesian model adequacy. For omnibus model selection, these authors instead recommend the DIC measure,

which is the sum of $\bar{D}$ and a deviance-based model complexity measure, p_D. Draper and Krnjajic (2007, Sec. 4.1) have shown that DIC approximates the LPML for approximately Gaussian posteriors. However, our models also lie fairly far outside the exponential family, where recent work (van der Linde, 2004; 2005) suggests the use of DIC may or may not be sensible. In what follows we investigate and compare the performance of LPML and DIC in model selection.

3.2 Results for SEER data

For the PH, AFT, and PO survival forms (1)–(3), we first fit what we call the "full model": an MPT prior on baseline survival centered at the log-logistic family, with a CAR prior for the spatial frailty terms. We denote these full models as MPT CAR frailty models. We also fit MPT i.i.d. frailty models, in which the frailties are modeled as independent (i.e., having no spatial pattern), and MPT non-frailty models, which do not include frailty terms at all. Similarly, we fit corresponding non-MPT CAR frailty models, non-MPT i.i.d. frailty models, and non-MPT non-frailty models, which do not include the MPT aspect (i.e., the baseline survival simply follows a parametric log-logistic density) but include spatial frailties, i.i.d. frailties, or no frailties, respectively. We fit all of these models using the algorithm described implemented in `Fortran 90`. Despite the high dimension of our models, the MCMC chains mixed reasonably well. For each model, we retained 100,000 iterations for posterior estimation following a burn-in of 50,000 iterations. For the MPT CAR frality, MPT i.i.d. frailty and non-frailty models we used our two priors for the c parameter in the MPT, the $\Gamma(5, 1)$ and the $\Gamma(20, 2)$. Tables 2, 3, and 4 show DIC and LPML scores for the competing PH, AFT, and PO models, respectively. Up to the degree of Monte Carlo accuracy in our results, the $\Gamma(5, 1)$ prior (which we recall favors smaller c values and hence a more flexible specification) performs better for all MPT models (PH, PO, and AFT) using either the DIC or LPML criterion. Among the non-MPT models, DIC and LPML performance is sometimes acceptable in the CAR frailty case, but degrades noticeably in the i.i.d. frailty and non-frailty cases.

In the PH and PO models, both DIC and LPML indicate similar same trends for goodness

of fit, with the MPT CAR frailty models outperforming the MPT i.i.d. frailty models, which outperform the MPT non frailty models, which outperform the non-MPT models. Among the MPT models, both the frailty terms and their spatial arrangement seem important to model prediction and fit. Moreover, the MPT part also offers a contribution to improve model fit over the fully parametric alternative. Among the PH, AFT, and PO models, the latter score best while the AFT models fare worst; in fact, the worst PO model outperforms the best AFT model. This finding is somewhat confirmed by the integrated hazard plots based on Cox-Snell residuals, shown in Figure 2. The plot for the PO MPT CAR frailty model is closest to a line with slope 1 (which indicates perfect model fitting) among those for the three full models. Finally, there is no obvious trend among the AFT models according to DIC or LPML, though the full model with the $\Gamma(5, 1)$ prior is among the best. This may be because the AFT models are not appropriate for this dataset, hence neither the MPT nor the spatial frailty extensions can improve the fit substantially.

Table 5 provides the posterior medians and equal-tailed 95% credible intervals for main effects (components of $\boldsymbol{\beta}$) under the full PH, AFT, and PO models. The PH model indicates that all of the predictors except regional stage are significant at the 0.05 level. Higher age at diagnosis increases the hazard; e.g., a twenty-year increase in age is associated with a $e^{0.018\times 20} \approx 1.43$-fold increase in hazard rate. 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 $e^{0.22} \approx 1.25$ times larger if their cancer is detected at the regional stage, and $e^{1.64} \approx 5.16$ times larger if detected at the distant stage.

Turning to the AFT assumption, a patient who is twenty years younger typically has a mean lifetime $e^{0.017\times 20} \approx 1.40$ 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 $e^{0.19} \approx 1.21$ times longer than a woman with regional stage, and $e^{1.50} \approx 4.48$ times longer than a woman with distant stage.

Finally, for the PO 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

$e^{-(-0.030)\times 20} \approx 1.82$ 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 $e^{-(-0.49)} \approx 1.63$ greater for regional stage versus local stage, and are $e^{-(-2.70)} \approx 14.88$ greater for distant stage versus local stage. Note that the regression effect for the regional stage is significant under the PO model, while marginally insignificant under the PH and AFT models. For these data, the model with the best LPML and DIC scores, the PO model, helps tease out a bit more signal relative to noise over the AFT and PH models. The main regression effects across all PO models are in fact quite similar and significant (not shown).

These findings are further confirmed by Figure 3, which shows the fitted survival densities for women aged 68.8 years at study entry (the mean in our dataset) for three disease stages under the three competing MPT CAR models, and assuming a spatial frailty of 0. These fitted densities are overlaid onto histograms of the observed survival times for study participants with entry ages 58.8 to 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; see `www.stat.unm.edu/~aparna/cdh.html`). In all three plots, the predicted density curves from the PO model (solid line) best mirror the data, while the AFT results (dotted line) appear oversmoothed. We remark that these trends are also consistent with the LPML and DIC values in Tables 2–4.

Proceeding without the poorly-fitting AFT model, Table 6 compares posterior medians and equal-tailed 95% credible intervals for main effects (components of $\boldsymbol{\beta}$) under the full PH and PO models with those obtained under standard semiparametric partial likelihood-based PH and parametric log-logistic PO models with and without i.i.d. frailties. The standard results were obtained using the `survival` package in `R 2.3.1`. As is often the case with main effects (which are typically well-identified), the estimates change little across models with the possible exception of the "distant" stage group.

Figure 4 offers a geographic summary of the overall fitted spatial frailty pattern for our best-fitting model, the PO MPT CAR. We see clusters of high frailty (poorer survival) in

the southwest, northeast, central, and east-central parts of the state. Note that this map is essentially spatially smoothed version of the raw data map in Figure 1, which enables spatial epidemiologists to better understand patterns of breast cancer mortality across the state. While the overall fitted survival patterns (not shown) are similar for the PH and AFT models, this is not the case across all important covariate groups. Figure 5 maps the differences in estimated one-year survival rates between two covariate groups (distant versus local stage) for PH and PO random effects models with and without the Polya tree structure, and without and without the spatial aspect (a total of eight models) for a woman of mean age at entry. In the figure, the first column gives results for MPT models while the second column reports non-MPT results; also, the upper four maps are from CAR frailty models, while the lower four use only i.i.d. frailties. Dramatic differences in the spatial patterns are now evident, with the PO models suggesting much larger differences, and the spatial models greatly clarifying the spatial pattern. These maps clarify that differences in identified spatial patterns of cancer incidence and survival can vary across groups depending on the statistical model assumed.

Finally, we compare the hazard ratios for two age groups and two stage groups across the PH, PO, and AFT models in two specific counties with disparate observed experience, Mahaska and Mills. From the original data, we know that Mahaska county had 2 events out of 26 diagnoses, while Mills county had 6 events out of just 9 diagnoses. Figure 6 provides the predictive hazard ratio for ages 88.8 and 68.8 (the mean age; left column) and for distant versus local stage (right column) in Mills (upper row) and Mahaska (lower row). These ratios are constant across time in the PH model (dashed line) by construction, but decreasing for PO (solid) and irregular for AFT (dotted). Note that the rate of decrease in the PO case is concave up for Mills, but concave down for Mahaska. The MPT CAR PO model thus offers appealing estimates of how the relative status of two groups changes over time and county.

4 Discussion and future work

In this paper, we have developed highly flexible survival models for time-to-event data that incorporate spatially varying or i.i.d. frailties. The models assume the same nonpara-

metric (really highly parametric) baseline S_0 centered at the log-logistic family of distributions. Often in the literature for these type of models, focus is on either the frailty structure or the choice of nonparametric baseline. More often than not, a variant of the proportional hazards model is chosen for the survival part of the model. Our findings for the SEER data indicate that all three model aspects are important for predicting patient survival. Roughly speaking, the LPML measure indicates that the *survival model itself* is most important, followed by whether the baseline is modeled as nonparametric or parametric, followed by the frailty model, or absence thereof.

The spatial frailty models outperformed the i.i.d. frailty models in terms of DIC and LPML. The CAR model imposes more spatial structure on the frailty terms, smoothing a particular county's frailty towards neighboring values, while the i.i.d. frailties are simply shrunk toward the statewide global mean. As in many geographically-oriented data settings, this seems to have led to predictive overfitting of the data by our unstructured frailty models. Coupled with the appealing spatial smoothness in the fitted maps, we found ample reason to prefer the MPT CAR models for our data.

To study the robustness of our MPT prior specification to the choice of baseline "centering" survival model, we replaced our log-logistic function with a Weibull. Table 7 shows the resulting DIC and LPML scores for this Weibull alternative. The results suggest similar trends across models in the PH case, and in fact the corresponding map of fitted frailties for the PH MPT CAR frailty model also indicates a very similar spatial pattern as the one seen in Figure 4. The PH MPT models thus appear robust with respect to the selection of the baseline function. However, when we tried to fit the AFT and PO models using the Weibull function, we encountered numerical instability stemming from numerical roundoff in the Weibull density tails.

Hanson (2006) observed a "leveling off effect" in LPML 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 $J = 6$. The first case led to some improvement in LPML score, but at the cost of roughly doubling the necessary computing time. The $J = 6$ case resulted in sufficiently poor MCMC convergence that we felt results could not

be reliably reported. Clearly the required level of computational intensity as we add levels to the Polya tree becomes infeasible at some point using today's computers, but we believe this problem will resolve on its own as computers continue to become faster.

Future work in this area looks toward extending the survival and frailty models to longitudinal specifications, allowing temporal change in a county's mortality. Temporally dynamic frailties anticipate county level variability in implementation of programs such as Iowa's *Care for Yourself* program, a part of the larger federal government's National Breast and Cervical Cancer Early Detection Program, established in 1991. Also, extending the survival part of the model to accommodate subject-specific time dependent covariates allows for temporal changes in personal behavior and multiple diagnoses of stage at differing time points.

References

Banerjee, S. and Carlin, B.P. (2002). Spatial semiparametric proportional hazards models for analyzing infant mortality rates in Minnesota counties. In *Case Studies in Bayesian Statistics, Volume VI*, eds. C. Gatsonis et al. New York: Springer-Verlag, pp. 137–151.

Banerjee, S. and Carlin, B.P. (2003). Semiparametric spatio-temporal frailty modeling. *Environmetrics*, **14**, 523–535.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press.

Banerjee, S. and Dey, D.K. (2005). Semi-parametric proportional odds models for spatially correlated survival data. *Lifetime Data Analysis*, **11**, 175-191.

Banerjee, S., Wall, M.M., and Carlin, B.P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, **4**, 123–142.

Berger, J.O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus a nonparametric alternatives. *Journal of the American Statistical Association*, **96**, 174-184.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**, 192–236.

Chen, Y.Q. and Jewell, N.P. (2001). On a general class of semiparametric hazards regression models. *Biometrika*, **88**, 687-702.

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

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

Dempster, A. P. (1997) The direct use of likelihood for significance testing. *Statistics and Computing*, **7**, 247-252.

Draper, D. and Krnjajic, M. (2007). Bayesian model specification. Technical report, Department of Applied Mathematics and Statistics, University of California – Santa Cruz.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615-629.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153-160.

Gelfand A.E. and Mallick, B.K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics*, **51**, 843–852.

Gustafson, P. (2007). On robustness and model flexibility in survival analysis: transformed hazards models and average effects. *Biometrics*, **63**, 69-77.

Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, **101**, 1548-1565.

Hanson, T. and Johnson, W.O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**, 1020-1033.

Hanson, T. and Yang, M. (2007). Bayesian semiparametric proportional odds models. *Biometrics*, **63**, 88-95.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065-1075.

Hutton, J. L. and Monaghan, P. F. (2002). Choice of parametric accelerated life and proportional hazards models for survival data: Asymptotic results. *Lifetime Data Analysis*, **8**, 375-393.

Huzurbazar, A. V. (2005). A censored data histogram. *Communications in Statistics-Simulation and Computation*, **34**, 113-120.

Ibrahim, J.G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag: New York.

Jin, X. and Carlin, B.P. (2005). Multivariate parametric spatio-temporal models for county level breast cancer survival data. *Lifetime Data Analysis*, **11**, 5–27.

Jin, X., Carlin, B.P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, **61**, 950-961.

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

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214-221.

Kuo, L. and Mallick, B.K. (1997). Bayesian semiparametric inference for the accelerated failured-time model. *Canadian Journal of Statistics* **25**, 457-472.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics*, **20**, 1222-1235.

Li, Y. and Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association*, **101**, 591-603.

Li, Y. and Ryan, L. (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics*, **58**, 287–297.

Mallick, B.K. and Walker, S.G. (2003). A Bayesian semiparametric transformation model incorporating frailties. *Journal of Statistical Planning and Inference*, 112, 159-174.

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

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, **98**, 1001-1012.

Scharfstein, D.O., Tsiatis, A.A., and Gilbert, P.B. (1998). Efficient estimation in the generalized odds-Rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis*, **4**, 355-391.

Sinha, D. and Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**, 1195-1212.

Spiegelhalter, D.J., Best, N., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, Series B, **64**, 583-639.

van der Linde, A. (2004). On the association between a random parameter and an observable. *Test*, **13**, 85–111.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, **59**, 45–56.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

Walker, S.G. and Mallick, B.K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B*, 59, 845-860.

Walker, S.G. and Mallick, B.K. (1999). Semiparametric accelerated life time model. *Biometrics*, **55**, 477-483.

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

Yin, G. and Ibrahim, J.G. (2005). A class of Bayesian shared gamma frailty models with multivariate failure time data. *Biometrics*, **61**, 208-216.

Zeger, S.L. and Karim, M.R. (1991). Generalised linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79-86.

continuous variables		mean	std deviation
follow-up time (months)		20.8	13.4
age (years)		68.8	15.8
categorical variables	**level**	**count**	**proportion (%)**
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

Table 1: Summary statistics for follow-up time and available covariates, Iowa SEER breast cancer data.

model	c prior	DIC	LPML
MPT CAR frailty	$\Gamma(5,1)$	4428.1	–2224.7
	$\Gamma(20,2)$	4429.1	–2229.0
MPT i.i.d. frailty	$\Gamma(5,1)$	4436.3	–2233.7
	$\Gamma(20,2)$	4438.9	–2239.5
MPT non-frailty	$\Gamma(5,1)$	4464.3	–2239.7
	$\Gamma(20,2)$	4479.4	–2241.5
non-MPT CAR frailty	—	4477.5	–2243.1
non-MPT i.i.d. frailty	—	4488.1	–2253.7
non-MPT non-frailty	—	4508.9	–2254.7

Table 2: DIC and LPML scores for the competing PH models, with the MPT centered at the log-logistic baseline.

model	c prior	DIC	LPML
MPT CAR frailty	$\Gamma(5,1)$	4460.6	–2236.1
	$\Gamma(20,2)$	4464.1	–2235.2
MPT i.i.d. frailty	$\Gamma(5,1)$	4471.4	–2252.3
	$\Gamma(20,2)$	4485.4	–2252.9
MPT non-frailty	$\Gamma(5,1)$	4466.1	–2234.6
	$\Gamma(20,2)$	4469.3	–2235.9
non-MPT CAR frailty	—	4473.2	–2237.0
non-MPT i.i.d. frailty	—	4504.4	–2256.7
non-MPT non-frailty	—	4478.7	–2239.5

Table 3: DIC and LPML scores for the competing AFT models, with the MPT centered at the log-logistic baseline.

model	c prior	DIC	LPML
MPT CAR frailty	$\Gamma(5,1)$	4404.8	–2208.8
	$\Gamma(20,2)$	4409.4	–2211.6
MPT i.i.d. frailty	$\Gamma(5,1)$	4417.2	–2218.4
	$\Gamma(20,2)$	4426.9	–2221.2
MPT non-frailty	$\Gamma(5,1)$	4420.0	–2223.5
	$\Gamma(20,2)$	4441.5	–2224.1
non-MPT CAR frailty	—	4460.0	–2230.3
non-MPT i.i.d. frailty	—	4479.4	–2242.6
non-MPT non-frailty	—	4478.6	–2239.4

Table 4: DIC and LPML scores for the competing PO models, with the MPT centered at the log-logistic baseline.

covariates	PH	AFT	PO
β_1 (centered age)	0.018 (0.012, 0.024)	0.017 (0.011, 0.024)	−0.030 (−0.038, −0.021)
β_2 (regional stage)	0.22 (−0.02, 0.45)	0.19 (−0.03, 0.40)	−0.49 (−0.75, −0.21)
β_3 (distant stage)	1.64 (1.41, 1.88)	1.50 (1.23, 1.77)	−2.70 (−3.01, −2.35)

Table 5: Posterior medians and 95% equal-tail credible intervals, MPT CAR frailty model fixed effects.

model	β_1 (centered age)	β_2 (regional stage)	β_3 (distant stage)
MPT CAR frailty PH	0.018 (0.012, 0.024)	0.22 (−0.02, 0.45)	1.64 (1.41, 1.88)
standard i.i.d. frailty PH	0.019 (0.013, 0.025)	0.26 (0.04, 0.49)	1.68 (1.45, 1.92)
standard non-frailty PH	0.019 (0.013, 0.025)	0.30 (0.08, 0.52)	1.64 (1.42, 1.87)
MPT CAR frailty PO	−0.030 (−0.038, −0.021)	−0.49 (−0.75, −0.21)	−2.70 (−3.01, −2.35)
standard i.i.d. frailty PO	−0.028 (−0.036, −0.020)	−0.37 (−0.66, −0.08)	−2.58 (−2.92, −2.24)
standard non-frailty PO	−0.029 (−0.037, −0.020)	−0.40 (−0.68, −0.12)	−2.53 (−2.86, −2.21)

Table 6: Point estimates and 95% equal-tail credible intervals, standard and MPT CAR PH and PO model fixed effects.

model	c prior	DIC	LPML
MPT CAR frailty	$\Gamma(5,1)$	4428.8	–2226.8
	$\Gamma(20,2)$	4433.4	–2227.0
MPT i.i.d. frailty	$\Gamma(5,1)$	4442.6	–2236.6
MPT non-frailty	$\Gamma(5,1)$	4475.4	–2244.0
non-MPT CAR frailty	—	4460.7	–2235.3

Table 7: DIC and LPML scores for the competing PH models, with the MPT centered at the Weibull baseline.

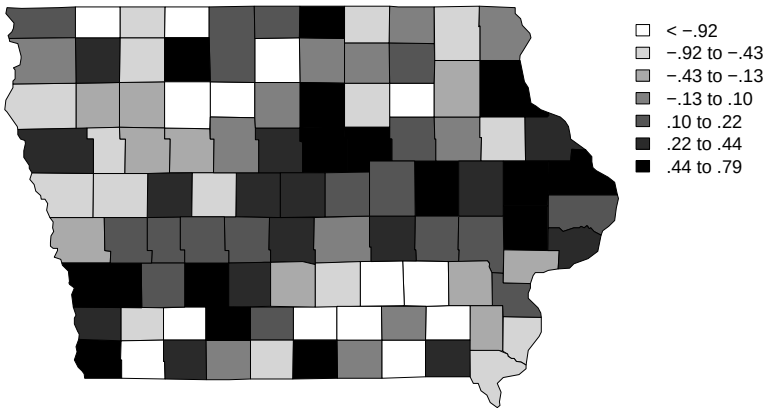

Figure 1: Log standardized mortality ratio by county.

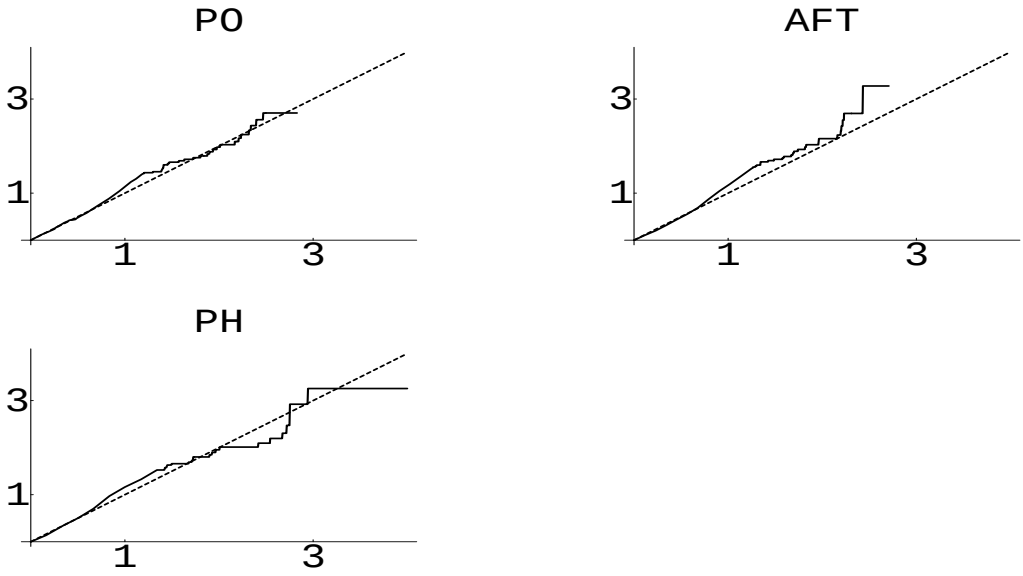

Figure 2: Integrated hazard plots for Cox-Snell residuals from PO, AFT, and PH MPT CAR frailty models under the $\Gamma(5, 1)$ prior for c.

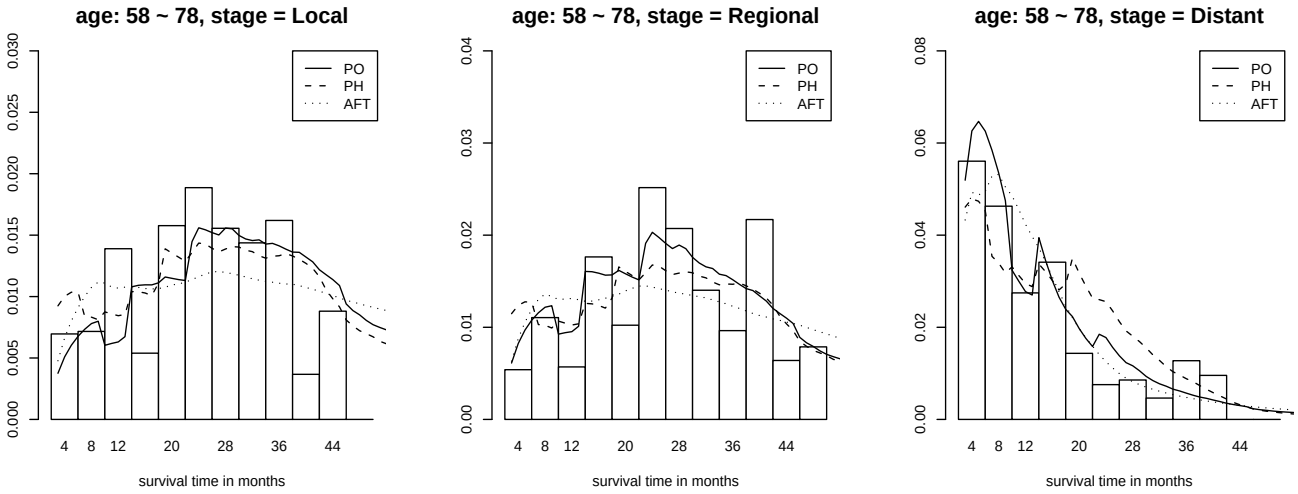

Figure 3: Histogram of raw times of death in months for observations with diagnosis ages between 58.8 and 78.8, and stage "local" (left), "regional" (middle), or "distant" (right), with fitted predictive density curves from the three competing MPT CAR models overlaid.

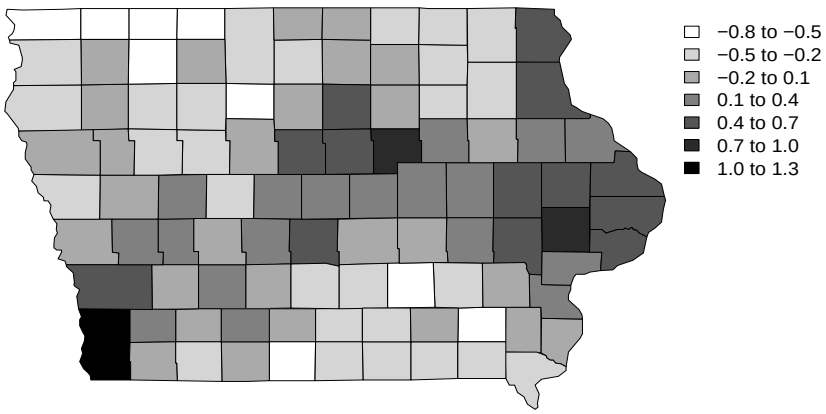

Figure 4: Fitted frailties for 1995–1998 Iowa SEER data, MPT CAR frailty PO model with $c \sim \Gamma(5, 1)$.

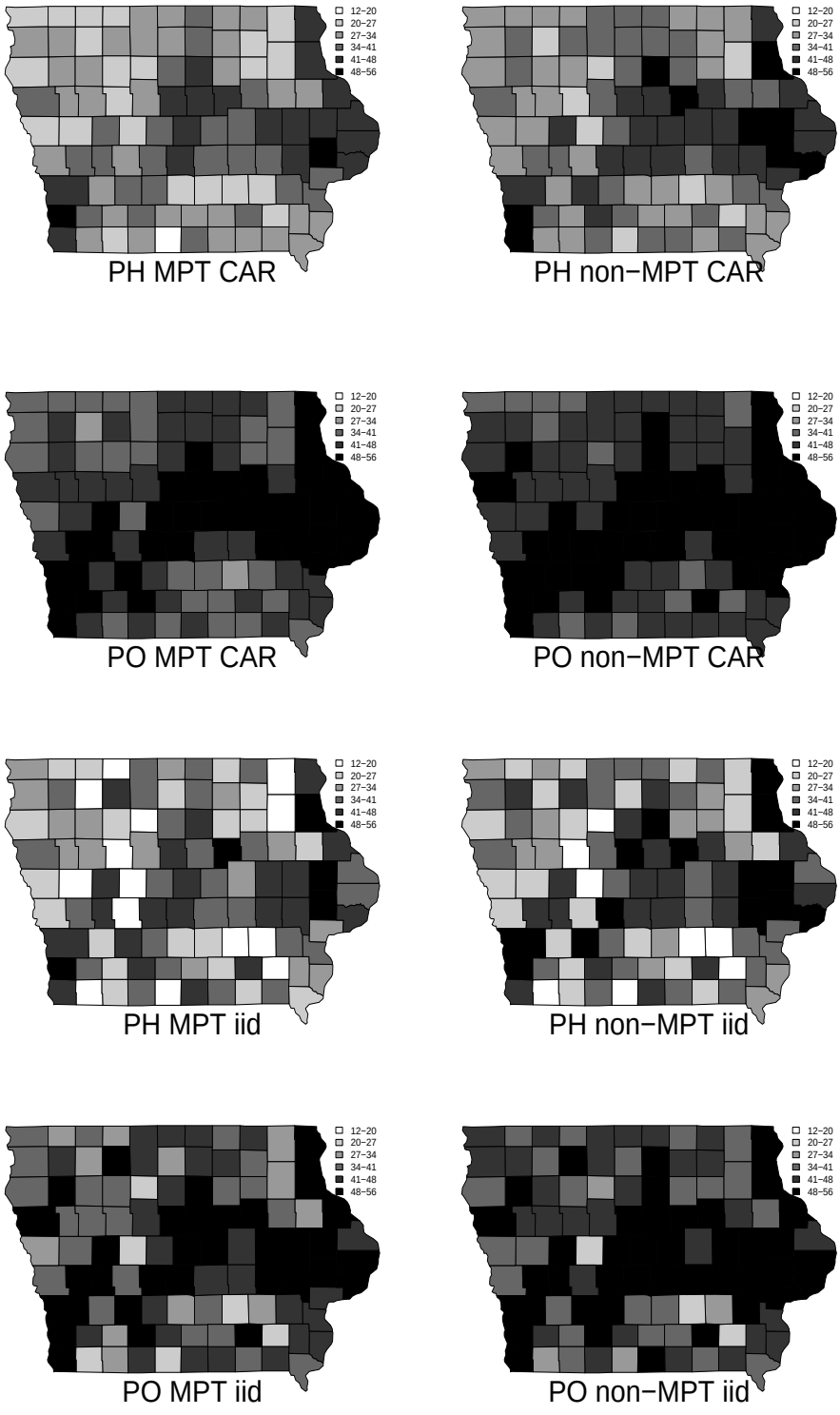

Figure 5: Predicted survival difference (%) between stages "local" and "distant" at the mean age after 1 year of follow up.

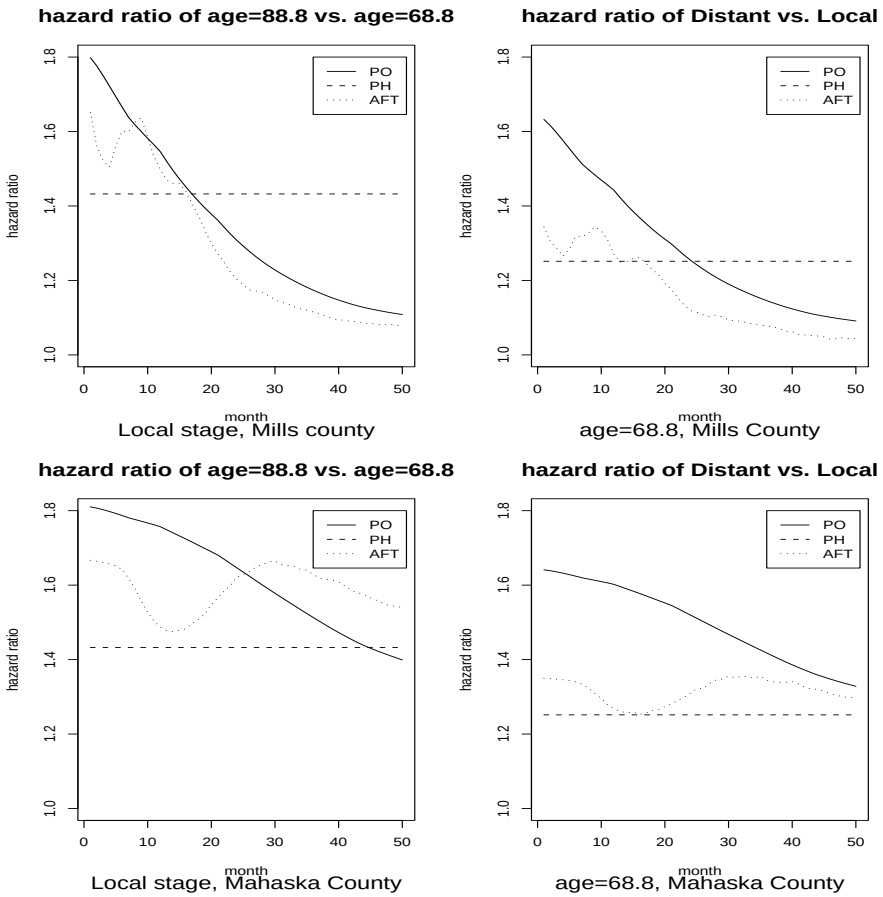

Figure 6: Fitted predictive hazard ratio of ages 88.8 and 68.8 (left column) and that of stages "distant" and "local" (right) in Mills County (upper row) and in Mahaska County (lower row) from the three competing MPT CAR models.