

# Resolution-invariant Binary Partition Priors, and a few other interesting problems in (Bayesian) statistics

---

Vanja M. Dukic  
University of Chicago

Based on joint work with: XL Meng (Harvard),  
P Bouman (Northwestern), J Dignam (Chicago)

## Outline

---

1. Bayesian statistics and inference
2. Curve as a parameter (survival and hazard curves)
3. Resolution-invariant priors for curves
  1. Good things about them
  2. Bad things about them
  3. Ugly (or at least not very elegant) ways around the bad
4. Computation: MCMC algorithms
5. Breast cancer/tamoxifen survival study
6. Other potentially interesting questions
  1. HMM with unknown number of states

## Bayesian statistics

---

- likelihood function

$$\mathcal{L}(\theta | Y) = \mathcal{P}(Y | \theta)$$

- prior probability distributions/ prior “beliefs”

$$p(\theta)$$

- posterior distribution (Bayes theorem)

$$\pi(\theta | Y) = p(\theta) \mathcal{P}(Y | \theta) / Z$$

- all inference based on the posterior of parameters,  $\pi(\theta | Y)$

## Curve as a parameter

---

- infinite-dimensional  $\theta$ 
  - composed of an unknown curve  $h(t)$
  - and other parameters  $\theta^*$
- examples:
  - rate of an event of interest, but instead of parametrically specifying its form we let it change “non-parametrically” over time
  - some unknown functional relationship in the model between outcome and predictor:
$$g(Y) = h(t) + \beta X$$
  - probability distribution of a trait of interest in a population

## Tamoxifen Trial Example

NCI-sponsored multicenter cancer trial:

- 168 centers in US and Canada
- 2818 women randomized after surgical removal of a breast tumor to tamoxifen or placebo
- observed for 10 years
- clinical outcome: disease-free survival time **T**
- Clinical questions of interest:
  - Differences in the **distributions of survival time T** between women on and off treatment?
  - Differences in the **risk (hazard) of cancer recurrence over time** in the two groups?

## Hazard and Survival functions

- survival function  $S(t)$  is the probability of surviving up to and beyond time  $t$ :

$$S(t) = P(T \geq t)$$

where  $T$  denotes the time of event ("failure")

- think of it as proportion of women who are still disease-free at time  $t$

- $S(0)=1$ ,  $S(t)$  non-increasing

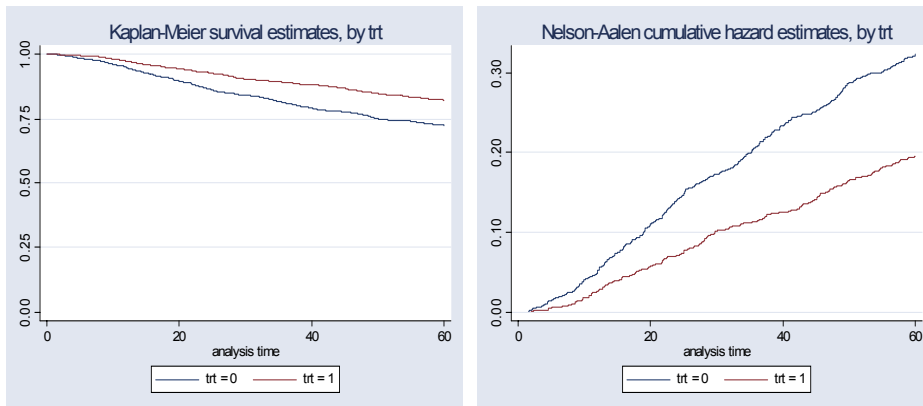
- hazard function:  $h(t)$  is instantaneous chance of failure, given no failure so far:

$$h(t) = \frac{d(-\log(S(t)))}{dt}$$

- Cumulative hazard (H):

$$H(t) = \int_0^t h(u) du$$

## Survival and cumulative hazard



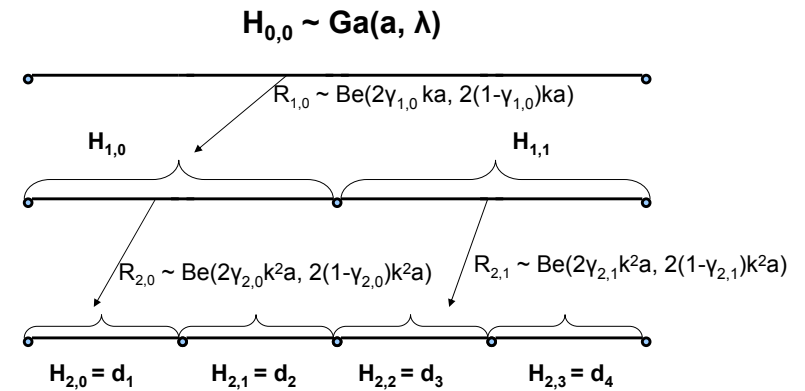
## Flexible Hazard Estimation

- hazard curve  $h(t)$  can be complicated, multimodal - many parametric models are unimodal
- Example: clustering of events around critical time points; those "high activity" times usually of interest per se
- Flexible and smooth hazard estimator needed
- Relatively easy to estimate (semi-parametric)
- with a tractable measure of uncertainty

# How do we specify priors on curves?

- (Beta and Gamma) process priors
  - independent (Kalbfleisch 1978; Dykstra and Laud 1981; Hjort 1990)
  - correlated (Nieto-Barajas and Walker 2000; Arjas and Gasbarra 1994; Aslanidou, Dey and Sinha, 1998)
- Discretized curve priors
  - discretized using  $J = 2^M$  time points between time 0 and final time  $T_J$
  - piece-wise constant (but arbitrary) hazard rates within intervals
  - possibly correlated (smooth) process models
  - Likelihood for event in interval  $(t_{j-1}, t_j)$  constant and same for observations with same characteristics
  - resolution-invariant priors (Nowak and Kolaczyk 2002)

# Binary partition prior



# Properties of the prior

We can set up the expected value of the piece-wise bits as we wish, by choosing  $a$  and  $\lambda$ :

- Start with the bottom level values – pick a desired mean of each  $d_j$

$$E(d_j) = d_j^* \quad (\text{in the prev picture } j = 1, \dots, 4)$$

- Recursively set  $\gamma_{2,0} = d_1^* / (d_1^* + d_2^*) \dots$
- Let  $a\lambda = \sum d_j^*$

# Properties of the prior (cont)

- With this parametrization:
  - the prior on the curve is self-consistent:
    - it does not depend on the total “resolution” level (depth of the tree), ie:
    - the joint prior of H’s at any given level is the same regardless of the final depth of resolution
  - We can directly control prior on smoothness by manipulating  $k$  and  $a$

## Properties of the prior (cont)

- prior  $\text{corr}(d_j, d_i) = \frac{\lambda^2 a^2}{4^M} \left[ \left( \prod_{l=1}^L \frac{2k^l a + 2k}{2k^l a + 1} \right) - 1 \right]$

where  $d_i$  and  $d_j$  are hazard increments that share first  $L-1$  levels of splits

- For  $k < 0.5$  hazard increments ( $d_j$ ) negatively correlated
- For  $k > 0.5$  hazard increments positively correlated
- For  $k = 0.5$  hazard increments independent: at the  $M$ -th level they are independent  $\text{Gamma}(a/2^M, \lambda)$

## Properties of Bayesian MR Model

- **Models with fixed  $a$**

- “Blocky” *a-priori* correlation structure in MR models due to binary partition tree construction of the model and not the usual Euclidean topology
- Nearby hazard increments may have lower correlation than those far apart

- **Mixing over  $a$  important:**

- equalizes *a-priori* between-hazard increment correlations

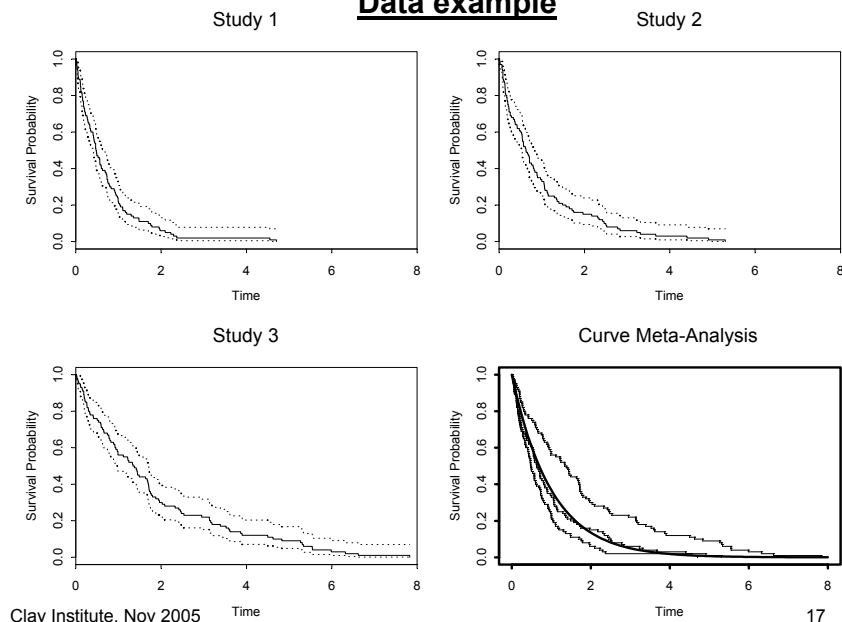
## Estimation

- Parameter posterior does not exist in closed form
- Very high dimensional
- Not easy to simulate from
- Model fitted via Markov Chain Monte Carlo (MCMC) technique: Metropolis-Hastings used to sample from sequential uni-dimensional conditional distributions

## Modeling possibly heterogeneous survival data

- Motivation:
  - meta-analysis of individual patient survival data
  - analysis of data from multiple trial centers
- Need:
  - a smooth estimate of the “common” survival curves
  - a smooth estimate of the “common” baseline hazards
- Capable of:
  - accounting for covariate effects
  - accounting for unobserved heterogeneity of data from different sources (studies or centers)
  - Center-specific survival predictions (borrowing strength from neighbors)

## Data example



## Proportional Hazards MR model

- Hazard discretized using  $J = 2^M$  time points
- Assume constant hazard within intervals
- Likelihood for “failure” in interval  $(t_{j-1}, t_j)$  is

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

(random effects can also be incorporated)

## Priors for Bayesian MR Model

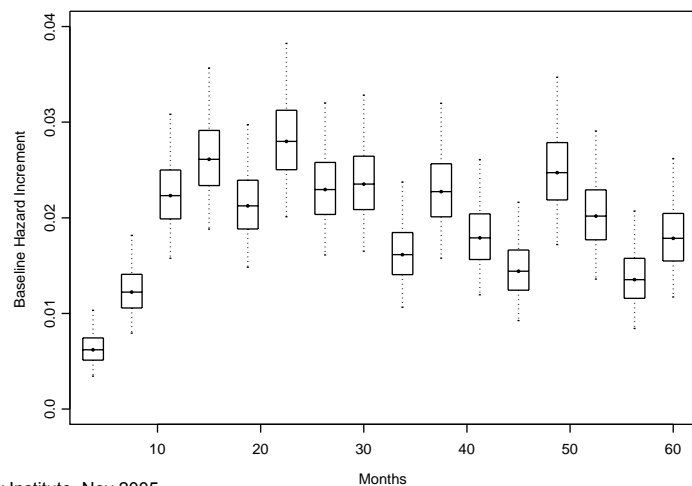
- $k \sim \text{Exp}(\mu_k)$
- $\lambda \sim \text{Exp}(\mu_\lambda)$
- $a \sim \text{Zero Truncated Poisson prior}$
- Random-effects:
  - $\sim \text{common } N(0, \tau^2)$  prior
  - Gamma hyperprior on  $\tau^2$
- $\beta$ 's  $\sim$  vague Normal priors

## Tamoxifen results

Marginal Posterior 95% Credible Intervals for Enrollment and Patient Covariate Effects

Log Hazard Ratio $\beta$	$\beta$ post. (2.5%, 50%, 97.5%)	$\exp(\beta)$ post. (2.5%, 50%, 97.5%)
Overall Treatment Effect	(-0.69, -0.52, -0.36)	(0.50, 0.59, 0.70)
Tumor Size 2.1-4 cm	(0.23, 0.40, 0.57)	(1.26, 1.49, 1.76)
Tumor Size $\geq 4.1$ cm	(0.42, 0.74, 1.04)	(1.52, 2.10, 2.84)
Prog. Receptor Level $\geq 10$ fmol/mg	(-0.56, -0.39, -0.21)	(0.57, 0.68, 0.81)
Stdized Age at Enrollment, Linear	(-0.07, 0.01, 0.10)	(0.93, 1.01, 1.11)
Stdized Age at Enrollment, Quadratic	(0.05, 0.11, 0.18)	(1.05, 1.12, 1.20)

## hazard curve estimate



Clay Institute, Nov 2005

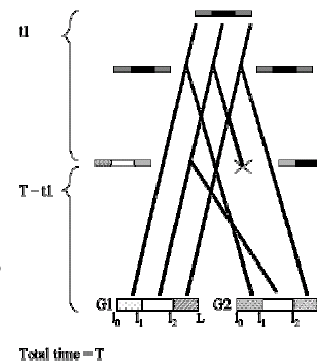
Months

21

## Other (potentially) interesting questions

HMM with unknown number of states:

Identifying multiple gene conversion regions  
Possibly many unknown divergence rates  
Additional conversion more likely within converted regions?



**A gene conversion event:** Horizontal bars represent gene sequences and the difference in pattern signifies the amount of divergence. At time  $t = 0$ , a gene is duplicated and the two copies begin to diverge. At time  $t = t_1$  a gene conversion occurs in which the central section of gene  $G_2$  is overwritten by the central section of gene  $G_1$ . At the present (time  $t = T$ ), the central region of both genes appears to have a more recent date of duplication than the flanking regions.

Clay Institute, Nov 2005

22

thanks 😊

Clay Institute, Nov 2005

23

## Bibliography

- Bouman, Dukic, Meng: A Bayesian multiresolution hazard model. *Stat Sinica* (2005).
- Bouman, Dignam, Meng, Dukic 2004: A multiresolution hazard model for multicenter survival studies: application to tamoxifen treatment in early stage breast cancer. *JASA* (under review).
- Gray 1994: A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*.
- Hjort 1990: Nonparametric Bayes estimators based on Beta processes in models for life history data. *Ann of Stat*.
- Kim and Lee 2004: Bayesian analysis of PH models. *Ann of Stat*.
- Nieto-Barajas and Walker 2002: Markov Beta and Gamma processes for modeling hazard rates. *Scand. J of Stat*.
- Ibrahim et al. 2001 : "*Bayesian Survival Analysis*". Springer.

Clay Institute, Nov 2005

24