Bayesian Nonparametric Nonproportional Hazards Survival Modelling

Maria De Iorio

Department of Statistical Science
UCL

Fundamental Problems in Survival Analysis

4th June 2014
Collaborators

Wesley O. Johnson
Peter Müller
Gary L. Rosner
Alejandro Jara
Fernando Quintana
Emmanuel Lesaffre
Outline

1. Motivation
2. DP and DDP
3. Survival Model
4. Cancer Clinical Trial
Survival probabilities for early times are lower for high dose treatment than for low dose.

The reverse is true for later times, possibly due to toxic effect of high dose.

K–M estimates

Survival

Low Dose  High Dose

Months

0 20 40 60 80 100 120 140

0.0 0.2 0.4 0.6 0.8 1.0

K–M estimates

Survival

Months

0 20 40 60 80 100 120 140

0.0 0.2 0.4 0.6 0.8 1.0

K–M estimates

Low Dose  High Dose
PH Model

- Let $T > 0$ denote a random survival (event) time.
- Let $P(T > t) = S(t)$: Survival Function
- $h(t)dt = P(T \in [t, t + dt] \mid T \geq t)$: Hazard Function
Let $T > 0$ denote a random survival (event) time.

Let $P(T > t) = S(t)$: Survival Function

$h(t)dt = P(T \in [t, t + dt] \mid T \geq t)$: Hazard Function

Denote risk factors (covariates) as $x = (x_1, \ldots, x_p)$.

The PH model relates covariates to the hazard:

$h(t \mid x) = \exp\{x'\beta\}h(t)$, where $h(t)$ is the baseline hazard

Readily available (non-Bayesian) softwares.

Ubiquitous in scientific literature.

Often fails to fit.
PH with stratification

Each stratum is permitted to have a different baseline hazard function.

**Limitation:** inability to formally examine the effects of treatment dose, no borrowing of information between groups.
Alternative models

- Accelerated Failure Time Model:
  \[ S(t \mid x) = G(\exp\{x'\beta\}t) \leftrightarrow T = \exp\{x'\beta\}V, \quad V \sim G \]

- Proportional Odds Model:
  \[
  \frac{S(t \mid x)}{1 - S(t \mid x)} = \exp\{x'\beta\} \frac{G(t)}{1 - G(t)}
  \]

- Others, like additive hazards

All these assumptions may be too strong in many practical applications.
We develop a model for survival analysis data based on a flexible nonparametric prior, the DP and related DDP.

Model extends ANOVA DDP (De Iorio et al., 2004) to handle continuous covariates and censored data.

A major feature is no assumption of proportional hazards.
Assume that \( S = \{ S_x, x \in X \} \) is an array of survivor functions, indexed by categorical covariate \( x \). Let \( x = (v, w) \), with \( v \in \{1, \ldots, V\} \) and \( w \in \{1, \ldots, W\} \).

Want:
"ANOVA" layout with a different survivor function for each combination of covariates.
Want: "ANOVA" layout with a different survivor function for each combination of covariates

\[ x = (v, w) \]
\[ S_{x_i} = S_{x_j}, \text{ if } x_i = x_j \]
\[ S_{x_i} \text{ close to } S_{x_j}, \text{ if } x_i \text{ and } x_j \text{ only differ in one covariate level} \]

\[ : \]
Let \( z \in Z \) be a continuous covariate, we get a collection of random distribution. The level of dependency is controlled by \( z \).
Dirichlet Process (DP)

Probability model on distributions $F \sim DP(M, F^o)$, with measure $F^o = E(F)$ and precision parameter $M$.

$F$ is a.s. discrete
Sethuraman’s stick breaking representation

\[ F = \sum_{h=1}^{\infty} p_h \delta_{m_h} \]

\[ w_h \sim \text{Beta}(1, M) \]

\[ p_h = w_h \prod_{i=1}^{h-1} (1 - w_i), \quad \text{scaled Beta distribution} \]

\[
\begin{array}{c|c|c|c}
& w_1 & w_2(1 - w_1) \\
\hline
m_1 & & & \\
\hline
m_2 & & & \\
\hline
m_h & & & \\
\end{array}
\]

\[ m_h \overset{iid}{\sim} F^0, \quad h = 1, 2, \ldots \]

where \( \delta(x) \) denotes a point mass at \( x \), \( p_h \) are weights of point masses at locations \( m_h \).
In many data analysis applications the discreteness is inappropriate. To remove discreteness: convolution with a continuous kernel

\[
f(y) = \int p(y \mid \mu) dF(\mu)
\]

\[F \sim DP(M, F^o)\]
or with latent variables $\mu_i$

$$F \sim DP(M, F^0)$$
$$\mu_i \sim F$$
$$f(y) = p(y \mid \mu_i)$$

Nice feature: Mixture is discrete with probability one, and with small $M$, there can be high probabilities of a finite mixture.

Often $p(y \mid \mu) = N(\mu, \sigma^2) \rightarrow f(y) = \sum_{h=1}^{\infty} p_h N(\mu_h, \sigma^2)$
(MacEachern, 1999) introduces a probability model for a collection of random distribution \( \{F_x, x \in X\} \)

Introduce dependence across \( x \) by assuming \( m_h = (m_{xh}, x \in X) \) dependent

\[
\begin{align*}
  x = 1 : & \quad F_1 = p_1 \delta m_{11} + p_2 \delta m_{12} + \ldots \\
  x = 2 : & \quad F_2 = p_1 \delta m_{21} + p_2 \delta m_{22} + \ldots \\
  x = 3 : & \quad F_3 = p_1 \delta m_{31} + p_2 \delta m_{32} + \ldots \\
  \vdots \\
\end{align*}
\]

\( m_h = \{m_{xh}, x \in X\} \sim iid \ p(m) \), which defines a stochastic process indexed by \( x \), for each fixed \( h \)
$F_x$ and $F_{x^*}$ are dependent by virtue of the modelled relationship between the random pairs
\{$(m_{xh}, m_{x^*h}) : h = 1, 2, \ldots$\}

Marginally: $F_x \sim DP(M, F_x^0)$, for all $x \in X$, $m_{xh} \overset{iid}{\sim} F_x^0$

Computationally easy

Special case: ANOVA DDP (De Iorio et al., 2004)
ANOVA DDP

- Categorical factors \( x = (v, w) \)
- Recall \( F = \sum p_h \delta_{m_h} \)
- Induce dependence across \( F_x \) by inducing dependence on point masses
- Introduce dependence across \( x = (v, w) \) by assuming an ANOVA model on the locations \( \{m_{xh}, x = (v, w), v = 1, \ldots, V, w = 1, \ldots, W\} \)

\[
m_{xh} = M_h + A_{vh} + B_{wh}
\]

with \( M_h \sim p_M(M_h), A_{vh} \sim p_{A_v}(A_{vh}), B_{wh} \sim p_{B_w}(B_{wh}) \)
e.g. \( M_h \sim N(\mu_h, \tau^2) \), etc. and \( A_{0h} \equiv B_{0h} \equiv 0 \)

- Independence across \( h \), dependent - as desired - across \( x \)
Interpretation

- Model for the \( \{ m_{xh} \} \): ordinary ANOVA
- Interpretation \( M_h \): "overall mean"
  \( A_h, B_h \): "main" effects for \( v \) and \( w \)
- Model is easily generalised to a \( p \)-dimensional covariate vector \( x = (x_1, \ldots, x_p) \)
- Include "interactions", additional factors, inference on contrasts etc. as in ANOVA
- Model allows us to incorporate differential prior information for the various covariate levels
- Easy to include constraints on the estimated effects
Extension to continuous covariates
Linear DDP

- Extension to continuous covariates
- Consider simple case with bivariate covariates
  \[ x = (v, z) \] where \( v \) is categorical and \( z \) is continuous
Extension to continuous covariates

Consider simple case with bivariate covariates $x = (v, z)$ where $v$ is categorical and $z$ is continuous

Dependence across random distribution by imposing a linear model on the locations (random effects LM)

$$m_{xh} = M_h + A_{vh} + \beta_h z$$

with $M_h \sim p_M(M_h), A_{vh} \sim p_{Av}(A_{vh})$ and $\beta_h \sim p_\beta(\beta_h)$

and independence across $h$
Extension to continuous covariates

Consider simple case with bivariate covariates $x = (v, z)$ where $v$ is categorical and $z$ is continuous.

Dependence across random distribution by imposing a linear model on the locations (random effects LM)

$$m_{xh} = M_h + A_{vh} + \beta_h z$$

with $M_h \sim p_M(M_h)$, $A_{vh} \sim p_{A_v}(A_{vh})$ and $\beta_h \sim p_\beta(\beta_h)$ and independence across $h$

We say $\{F_x : x \in X\} \sim \text{Linear DDP}(M, p^o)$
Linear DDP

- Extension to continuous covariates
- Consider simple case with bivariate covariates $x = (v, z)$ where $v$ is categorical and $z$ is continuous
- Dependence across across random distribution by imposing a linear model on the locations (random effects LM)
  \[ m_{xh} = M_h + A_{vh} + \beta_h z \]
  with $M_h \sim p_M(M_h)$, $A_{vh} \sim p_{A_v}(A_{vh})$ and $\beta_h \sim p_\beta(\beta_h)$ and independence across $h$
- We say $\{F_x : x \in X\} \sim \text{Linear DDP}(M, p^o)$
- The model is easily generalised to more than one continuous covariate
Our goal is to model survival time $T$ as a function of covariate information, $x$.

- In the Cox Model, $S(t|x) = (S_0(t))^{e^{x'\beta}}$
- In the AFT model, $S(t|x) = S_0(\exp(x'\beta)t)$
- In the PO model,
  
  \[ S(t|x) = e^{x'\beta}S_0(t)/[1 - S_0(t) + e^{x'\beta}S_0(t)] \]

- NP priors are placed on $S_0$, and independent parametric priors are placed on $\beta$
- These are partially parametric models
DDP Regression

\[ T \mid x, F_x \sim \int f(t \mid \theta) dF_x(\theta) \]

\{F_x, x \in X\} \sim \text{Linear DDP}(M, F^o)

Recall \( F_x = \sum_h \rho_h \delta_{m_{xh}}, \) with \( m_{xh} \overset{iid}{\sim} F_{0x} \)
Choices for $f(t|\theta_x)$

- Standard choice would be a log normal pdf

$$f(t|\theta_x = (\mu_x, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-0.5 \frac{(\log(t) - \mu_x)^2}{2\sigma^2}}$$

- The model would then be a discrete mixture of log normals

- Alternatively, could specify a simple normal; not unreasonable due to flexibility of the mixture. We could still model the log-data.

- Weibull, log logistic, gamma etc.
Formulation of Linear DDP as DPM

- Consider case with bivariate covariate $x = (v, z)$
- Let $\alpha_h = [M_h, A_{2h}, \ldots, A_{Vh}, \beta_h]$ denote the row vector corresponding to the $h$-th point mass
- Let $d_x$ denote a design vector such that $\mu_{xh} = \alpha_h d_x$
- Then the linear DDP model can be written as

$$p(t \mid x, F) = \int f(t \mid \alpha d_x, \sigma^2) dF(\alpha, \sigma^2)$$

$$F \sim DP(M, F^o)$$

where $F^o = (p_M, p_A, p_\beta, p_{\sigma^2})$
When $M$ is large, $F$ concentrates on $F^o$, and the model becomes a traditional parametric Bayesian LM or log LM

that is,

$$p(t \mid x, \theta) = \int f(t \mid \alpha d_x, \sigma^2) dF^o(\alpha, \sigma^2)$$

With the additional prior on the "hyperparameters" of $F^o$, this is a hierarchical model
For the normal linear model formulation,

\[ E(T|x, \alpha, F) = m + A_v + \beta z \]

\[ (\alpha, \sigma^2) \sim F, \quad F \sim DP(M, F^0) \]

We are just mixing the linear model using the random mixture \( F \), which for small \( M \) will tend to be a finite mixture.

In the case of log normal kernel,

\[ \text{med}(T|x, \alpha, F) = e^{(m + A_v + \beta z)} \]
The Data

- Standard censored survival data: \{ (t_i, \nu_i) : i = 1, \ldots, n \}
- Under the Linear DDP, all observations will be independent
- Introducing latent variables \((\alpha_i, \sigma_i^2)\), we can rewrite model hierarchically

\[
\begin{align*}
  t_i \mid x_i, \alpha_i, \sigma_i^2 &\sim \mathcal{N}(t_i \mid \alpha_i d_{x_i}, \sigma_i^2) \\
  (\alpha_i, \sigma_i^2) \mid F &\sim \text{iid} F, \quad F \sim \text{DP}(M, F^o)
\end{align*}
\]

- In words, the observations \(t_i\) are sampled from a mixture of heteroscedastic linear models, with a DP prior on the unknown mixing measure
Likelihood for censored observations

\[ L(\theta) = \prod_{i=1}^{n} \rho(t_i \mid x_i, \theta)^{\nu_i} S(t_i \mid x_i, \theta)^{1-\nu_i} \]

\[ \rho(t \mid x, \theta) = \int f(t \mid \theta) dF_x(\theta) \]
This representation implies that any Markov chain Monte Carlo (MCMC) scheme for DP mixture models can be used for posterior simulation cf. MacEachern and Mueller (1998), Neal (2000), Jain and Neal (2004)

De Iorio et al (2004) give relevant modifications needed for the ANOVA DDP model

The conjugate nature of the base measure $F^o$ and the kernel $f(t|x, \alpha, \sigma^2)$ greatly simplifies posterior simulation

Extension to handle censored observations

R packages available on Peter’s webpage
The conditionally conjugate base measure is

\[ \sigma^2 \sim \text{Inverse-Gamma} \left( \frac{s_0}{2}, \frac{s_0 S}{2} \right) \]

\[ \alpha \sim N(m, B) \]

Conjugate hyperpriors are:

\[ S \sim \text{Ga} \left( \frac{q_0}{2}, \frac{q_0}{2R_0} \right) \]

\[ m \sim N(a_0, A_0), \quad B^{-1} \sim \text{Wish} \left( c_0, (c_0 C_0)^{-1} \right) \]

\[ E(S) = R_0, \quad E(B^{-1}) = C_0^{-1} \]

\[ M \sim \text{Ga} \left( \gamma_0, \lambda_0 \right). \text{ See West (1992).} \]
Extension to Poisson-Dirichlet Process

\[ F = \sum_{h=1}^{\infty} p_h \delta_{m_h} \]

\[ w_h \sim \text{Beta}(1 - a, M + ha), \quad 0 \leq a < 1 \text{ and } M > -a \]

\[ p_h = w_h \prod_{i=1}^{h-1} (1 - w_i), \quad \text{scaled Beta distribution} \]

\[ m_h \overset{iid}{\sim} F^o, \quad h = 1, 2, \ldots \]

where \( \delta(x) \) denotes a point mass at \( x \), \( p_h \) are weights of point masses at locations \( m_h \).

Easy to extend Linear Dependent framework to the atoms of the process.
High-dose chemotherapy with bone marrow or stem cell transplantation is controversial therapy for treating women with breast cancer. It consists in giving ultra-high doses of toxic anti-cancer drugs, high enough to wipe out the woman’s blood-cell producing marrow and, hopefully, any residual cancer cells circulating in the body. Patients receive substantial supportive care to help the patient regenerate blood cells and avoid life-threatening infections.
In the 1990s, the Cancer and Leukemia group carried out a randomized clinical trial of high-dose (HD) chemotherapy with transplantation versus lower-dose chemotherapy.

The primary endpoint was disease-free survival (time until death from any cause, relapse, or diagnosis with a second malignancy).

High doses of treatment are known to be associated with a high risk of treatment related mortality early on — researchers expected the hazard functions to cross when comparing HD therapy to LD.

Those advocating HD hope the initial risk is subsequently offset by a substantial reduction in mortality and disease recurrence, justifying a more aggressive therapy.
The data record the event-free survival time in months for 761 women. 53% of the observations are censored. We consider 2 categorical covariates, 1 continuous plus an interaction term:

- Treatment Dose (low/high)
- Estrogen receptor (ER) status (pos/neg)
- Tumor Size (TS)
- Dose by ER interaction
<table>
<thead>
<tr>
<th>$t$</th>
<th>(months)</th>
<th>$\nu$</th>
<th>(freq.)</th>
<th>Dose</th>
<th>(freq.)</th>
<th>TS</th>
<th>(cm)</th>
<th>ER</th>
<th>(freq.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Med</td>
<td>21.88</td>
<td>Cens</td>
<td>400</td>
<td>High</td>
<td>385</td>
<td>Mean</td>
<td>3.8</td>
<td>Pos</td>
<td>528</td>
</tr>
<tr>
<td>IQR</td>
<td>33.54</td>
<td>Event</td>
<td>361</td>
<td>low</td>
<td>376</td>
<td>STD</td>
<td>2.4</td>
<td>Neg</td>
<td>233</td>
</tr>
</tbody>
</table>
The primary motivation was to compare low versus high dose.

Preliminary analysis: PH analysis using tumour size and ER status as covariates and stratifying by treatment dose.

Limitation: Difficulty to examine treatment effects.

Proposed model-based Bayesian inference provides a full probabilistic description of uncertainties in addition to the point estimates of the survivor function.

The model includes inference about any functional of interest of the survivor function.
Results: only treatment dose

Posterior estimated survivor functions for high versus low dose. For comparison the right panel shows the data (KM).
Comparison with PH Models

- PH with time dependent covariate.
- PH with time-varying coefficients.
When PH assumption does not hold and interest focuses on a binary covariate, one approach consists in introducing an indicator function as a time dependent covariate (Klein and Moeschberger 1997).

Categorical variable $x$ for treatment dose with $x = 1$ high dose and 0 otherwise.

We introduce a second indicator variable $z$ as time dependent covariate:

$$z = \begin{cases} 
  x & \text{if } y > y^* \\
  0 & \text{if } y \leq y^* 
\end{cases}$$
The hazard rate for the PH model becomes

\[
h(y \mid x, z) = \begin{cases} 
  h_0(y) \exp(\beta_1 x) & \text{if } y \leq y^* \\
  h_0(y) \exp((\beta_1 + \beta_2)x) & \text{if } y > y^* 
\end{cases}
\]

where \( h_0(y) \) is the baseline hazard rate.

\(-\exp(\beta_1) \) is the relative risk prior to time \( y^* \) for the high dose group relative to the low dose group

\(-\exp(\beta_1 + \beta_2) \) is the relative risk after time \( y^* \).

\(-\exp(\beta_2) \) is the increase in relative risk after time \( y^* \), \textit{change point} for the relative risk

To fix \( y^* \): we fit the model on a grid of values for \( y^* \) and we choose \( y^* \) as the value with the largest log likelihood.
PH with time-varying coefficients

- It assumes that the effect of a covariate changes over time: $\beta(t)$
- The covariate itself is fixed
- We used spline basis functions to model the effect over time
Estimated survivor functions: the solid line refers to low-dose group, while the dashed line refers to high dose group.
Posterior estimated survivor function and hazard hazard for high versus low dose, for tumour size = 2 cm and positive ER status. The grey shaded bands show point-wise central 50% credible intervals.
Posterior distribution of the difference in survival probabilities at 10, 20, 40, 60, 80 and 100 months between HD versus LD. The boxplots show the posterior distributions of the differences for positive ER status and tumour size = 2 cm. Note how the difference changes sign from 20 to 40 months.
Comparison with AFT model

Posterior survivor functions using the AFT median regression model (Hanson and Johnson, 2002) and using the DDP model. Note the almost vanishing difference between the solid and the dashed line.
Extension to more complex censoring

- **Interval Censoring.** A failure time $T$ cannot be observed but can only be determined to lie in an interval obtained from a sequence of observation/examination times.

- **Doubly Interval-Censored Data.** Survival time is defined as the elapsed time between two related events. Observations on the occurrence of both events could be right- or interval-censored.
  - Common in disease progression/epidemiological studies: initial event represents infection and subsequent event represents onset of disease
  - Follow up studies of patients who are at risk of being infected by the HIV and thus of developing AIDS. Interest on AIDS incubation time (survival time), time between HIV infection and diagnosis of AIDS.
Doubly-Interval Censored Data

\[ T^O = \text{true onset time} \]
\[ T^E = \text{true event time} \]
\[ T^T = T^E - T^O = \text{true time-to-event, survival time} \]

\[ (T^O, T^T) \mid x, F_x \sim \int f(t^O, t^T \mid \theta) dF_x(\theta) \]
\[ \{F_x, x \in X\} \sim \text{Linear DPD/DDP} \]
Programs are available as a function in the R package `ddpanova` at
http://www.ma.utexas.edu/users/pmueller/prog.html

The function `ddpsurvival(.)` implements the proposed DDP survival regression model.

Some function available in the R package `DPpackage` by A. Jara.
Discussion

- We have introduced a flexible nonparametric model that can be used to introduce categorical and continuous covariates in survival model based on DP priors.
- ease of interpretation
- facility to impose structure
- efficient computation
- MCMC scheme relies on the conjugacy of the base measure and mixing kernel
- it could be extended to more complex nonparametric prior
- We have extended the model to more complex censoring structure