A DDP Model for Survival Regression

Maria De Iorio
joint with Wesley O. Johnson, Peter Müller
and Gary L. Rosner

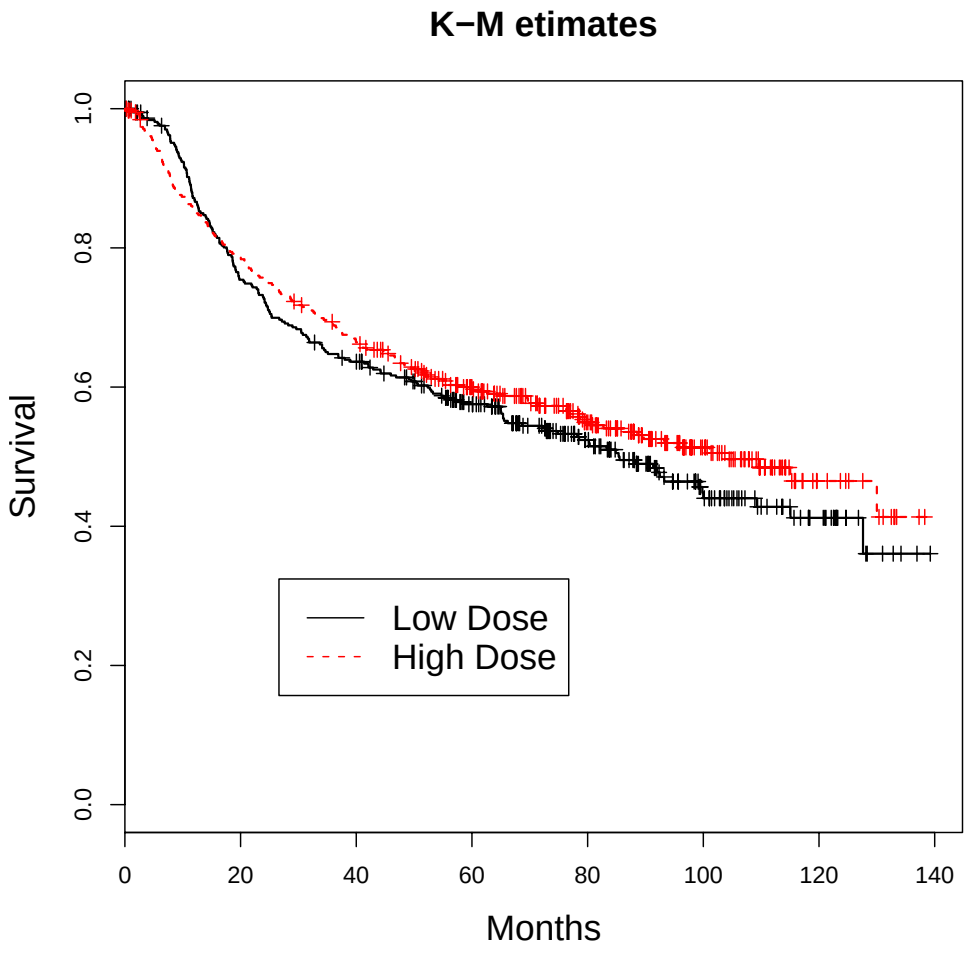
Department of Epidemiology and Public Health
Imperial College London

Construction and Properties of Bayesian Nonparametric Regression Models

Isaac Newton Institute, 8th August 2007

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



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

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 Model

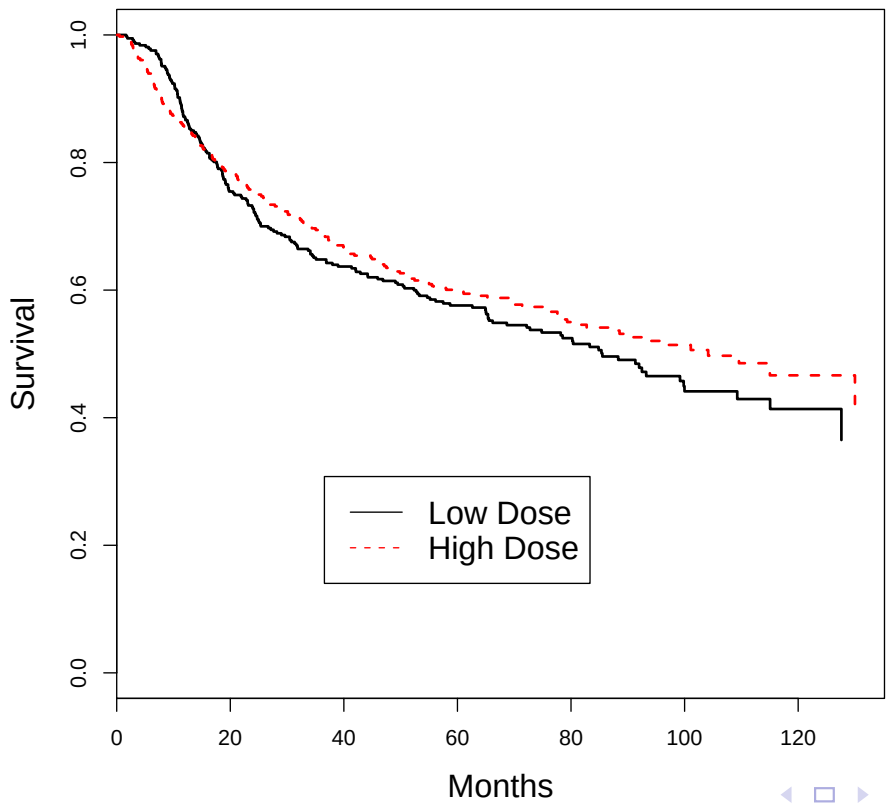
- Let $T > 0$ denote a random survival (event) time.
- Let $P(T > t) = S(t)$: Survival Function
- $h(t)dt = P(T \in [t, t + dt] \mid T \geq t)$: Hazard Function

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.



Alternative models

- Accelerated Failure Time Model:

$$S(t \mid x) = S_0(\exp\{x'\beta\}t) \leftrightarrow T = \exp\{x'\beta\}V, \quad V \sim S_0$$

- Proportional Odds Model:

$$\frac{S(t \mid x)}{1 - S(t \mid x)} = \exp\{x'\beta\}\frac{S_0(t)}{1 - S_0(t)}$$

- Others, like additive hazards

Alternative models

- Accelerated Failure Time Model:

$$S(t \mid x) = S_0(\exp\{x'\beta\}t) \leftrightarrow T = \exp\{x'\beta\}V, \quad V \sim S_0$$

- Proportional Odds Model:

$$\frac{S(t \mid x)}{1 - S(t \mid x)} = \exp\{x'\beta\}\frac{S_0(t)}{1 - S_0(t)}$$

- Others, like additive hazards

Alternative models

- Accelerated Failure Time Model:

$$S(t \mid x) = S_0(\exp\{x'\beta\}t) \leftrightarrow T = \exp\{x'\beta\}V, \quad V \sim S_0$$

- Proportional Odds Model:

$$\frac{S(t \mid x)}{1 - S(t \mid x)} = \exp\{x'\beta\}\frac{S_0(t)}{1 - S_0(t)}$$

- Others, like additive hazards

Aim

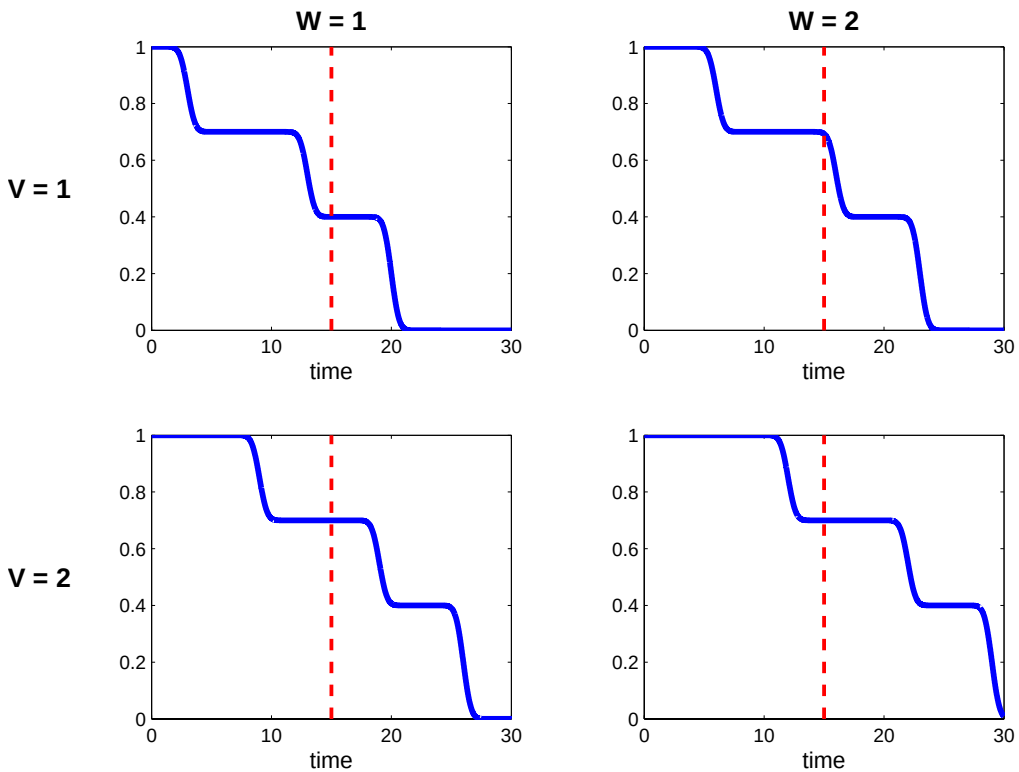
- We develop a DDP model for survival analysis data
- Model extends ANOVA DDP (De Iorio *et al.*, 2004) to handle continuous covariates and censored data
- A major feature is no assumption of proportional hazards

ANOVA for Random Survivor/Hazard Functions

Assume that $S = (S_x, x \in X)$ is an array of survivor funtions, 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



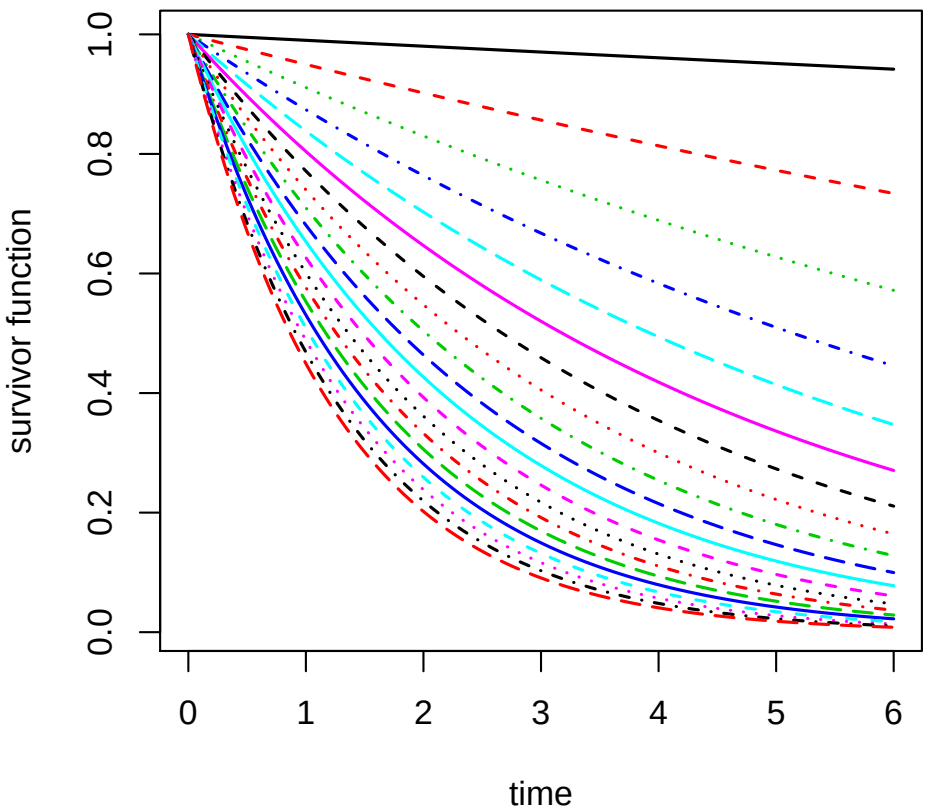
ANOVA for Random Survivor/Hazard Functions

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

$$
\begin{array}{rcll}
x & = & (v, w) & \\
S_{x_i} & = & S_{x_j}, & \text{if } x_i = x_j \\
S_{x_i} & \text{close to} & S_{x_j}, & \text{if } x_i \text{ and } x_j \text{ only differ in one covariate level} \\
 & \vdots & &
\end{array}
$$

Continuous covariate

Let $z \in Z$ be a continuous covariate, we get a collection of random distribution. The level of dependency is controlled by z.



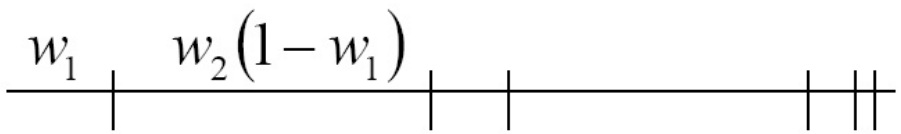
Dirichlet Process (DP)

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

$$
\begin{aligned}
F &= \sum_{h=1} p_h \delta_{m_h} \\
w_h &\sim \text{Beta}(1, M) \\
p_h &= w_h \prod_{i=1}^{h-1} (1 - w_i), \quad \text{scaled Beta distribution}
\end{aligned}
$$



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

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

Dirichlet Process Mixtures (DPM)

In many data analysis applications the discreteness is inappropriate.
To remove discreteness: convolution with a continuous kernel

$$\begin{aligned} f(y) &= \int p(y \mid \mu) \mathrm{d}F(\mu) \\ F &\sim DP(M, F^o) \end{aligned}$$

Dirichlet Process Mixtures (DPM)

or with latent variables μ_i

$$\begin{aligned} F &\sim DP(M, F^o) \\ \mu_i &\sim F \\ f(y) \mid \mu_i &= p(y \mid \mu_i) \end{aligned}$$

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

Often $p(y \mid \mu) = N(\mu, \sigma^2) \longrightarrow f(y) = \sum_{h=1}^{\infty} p_h N(\mu_h, \sigma^2)$

Dependent Dirichlet Process (DDP)

- (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

$$
\begin{aligned}
x = 1: &\quad F_1 = p_1 \delta_{m_{11}} + p_2 \delta_{m_{12}} + \dots \\
x = 2: &\quad F_2 = p_1 \delta_{m_{21}} + p_2 \delta_{m_{22}} + \dots \\
x = 3: &\quad F_3 = p_1 \delta_{m_{31}} + p_2 \delta_{m_{32}} + \dots \\
&\dots
\end{aligned}
$$

- $m_h = \{m_{xh}, x \in X\} \overset{iid}{\sim} p(m)$, which defines a stochastic process indexed by x, for each fixed h

DDP

- F_x and $F_{x^\star}$ are dependent by virtue of the modelled relationship between the random pairs $\{(m_{xh}, m_{x^\star h}) : h = 1, 2, \ldots\}$
- Marginally: $F_x \sim DP(M, F_x^o)$, for all $x \in X$, $m_{xh} \overset{iid}{\sim} F_x^o$
- 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. $M_h \sim \mathrm{N}(\mu_h, \tau^2)$, 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

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 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 $\{F_x : x \in X\} \sim$ Linear DDP(M, p^o)
- The model is easily generalised to more than one continuous covariate

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 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 $\{F_x : x \in X\} \sim$ Linear DDP(M, p^o)
- The model is easily generalised to more than one continuous covariate

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 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 $\{F_x : x \in X\} \sim$ Linear DDP(M, p^o)
- The model is easily generalised to more than one continuous covariate

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 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 $\{F_x : x \in X\} \sim$ Linear DDP(M, p^o)
- The model is easily generalised to more than one continuous covariate

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 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 $\{F_x : x \in X\} \sim$ Linear DDP(M, p^o)
- The model is easily generalised to more than one continuous covariate

Survival Regression

- 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 $\beta \longrightarrow$ See Wes and Alessandra's talks
- These are partially parametric models

DDP Regression

$$
\begin{aligned}
T \mid x, F_x &\sim \int f(t \mid \theta) \mathrm{d}F_x(\theta) \\
\{F_x, x \in X\} &\sim \text{Linear DDP}(M, F^o)
\end{aligned}
$$

Recall $F_x = \sum_h p_h \delta_{m_{xh}}$, with $m_{xh} \overset{iid}{\sim} F_{0x}$

Choices for $f(t|\theta_x)$

- Standard choice would be a log normal pdf

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

- 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
- 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

$$\begin{aligned} p(t \mid x, F) &= \int f(t \mid \alpha d_x, \sigma^2) \mathrm{d}F(\alpha, \sigma^2) \\ F &\sim DP(M, F^o) \end{aligned}$$

where $F^o = (p_M, p_A, p_\beta, p_{\sigma^2})$

Large M

- When M is large, F concentrates on F^o, and the model becomes a traditional parametric Bayesian LM or log LM
- that is,

$$p(t \mid x, \theta) = \int f(t \mid \alpha d_x, \sigma^2) \mathrm{d}F^o(\alpha, \sigma^2)$$

- With the additional prior on the "hyperparameters" of F^o, this is a hierarchical model

Linear DDP as DPM

- For the normal linear model formulation,

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

$$(\alpha, \sigma^2) \sim F, \quad F \sim DP(M, F^o)$$

- 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: $\{(t_i, \nu_i) : i = 1, ..., n\}$
- Under the Linear DDP, all observations will be independent
- Introducing latent variables (α_i, σ_i^2), we can rewrite model hierarchically

$$t_i \mid x_i, \alpha_i, \sigma_i^2 \sim \mathrm{N}(t_i \mid \alpha_i d_{x_i}, \sigma_i^2)$$

$$(\alpha_i, \sigma_i^2) \mid F \overset{iid}{\sim} F, \quad F \sim DP(M, F^o)$$

- In words, the observations t_i are sampled from a mixture of heteroscedastic linear models, with a DP prior on the unknown mixing measure

Gibbs Sampling

- 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

Completion of the Model

- The conditionally conjugate base measure is

$$\sigma^2 \sim \text{Inverse-Gamma}\,(s_0/2, s_0 S/2)$$

$$\alpha \sim N(m, B)$$

- Conjugate hyperpriors are:

$$S \sim \text{Ga}(q_0/2, q_0/2R_0)$$

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

- $E(S) = R_0, \qquad E(B^{-1}) = C_0^{-1}$
- $M \sim \text{Ga}(\gamma_0, \lambda_0)$. See West (1992).

Simulation Study

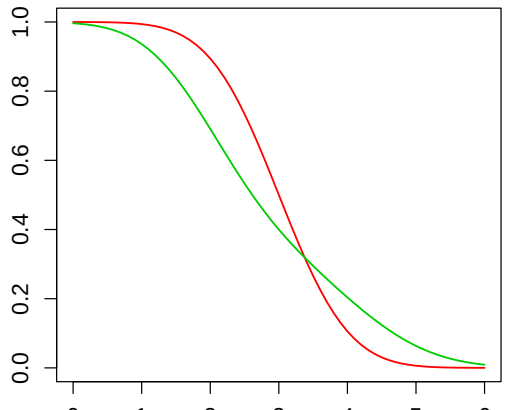
- Simulated data for 2 groups of hypothetical patients, e.g. patients in 2 different treatment groups (low/high dose)
- Low dose group

$$y = \log(t) \sim \mathrm{N}(3, 0.8^2)$$

- High dose group

$$y = \log(t) \sim 0.4\mathrm{N}(4, 1) + 0.6\mathrm{N}(2, 0.8^2)$$

The survivor functions cross around $y = 3.4$



Simulation Study

- We generated data for different sample sizes: $n = 25, 50, 100, 250, 500$ observations per group and $n = 500$ with censoring
- We compare the DDP model to a PH model with time dependent covariates

PH with time dependent covariate

- 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:

$$z = \begin{cases} x & \text{if } y > y^\star \\ 0 & \text{if } y \leq y^\star \end{cases}$$

PH with time dependent covariate

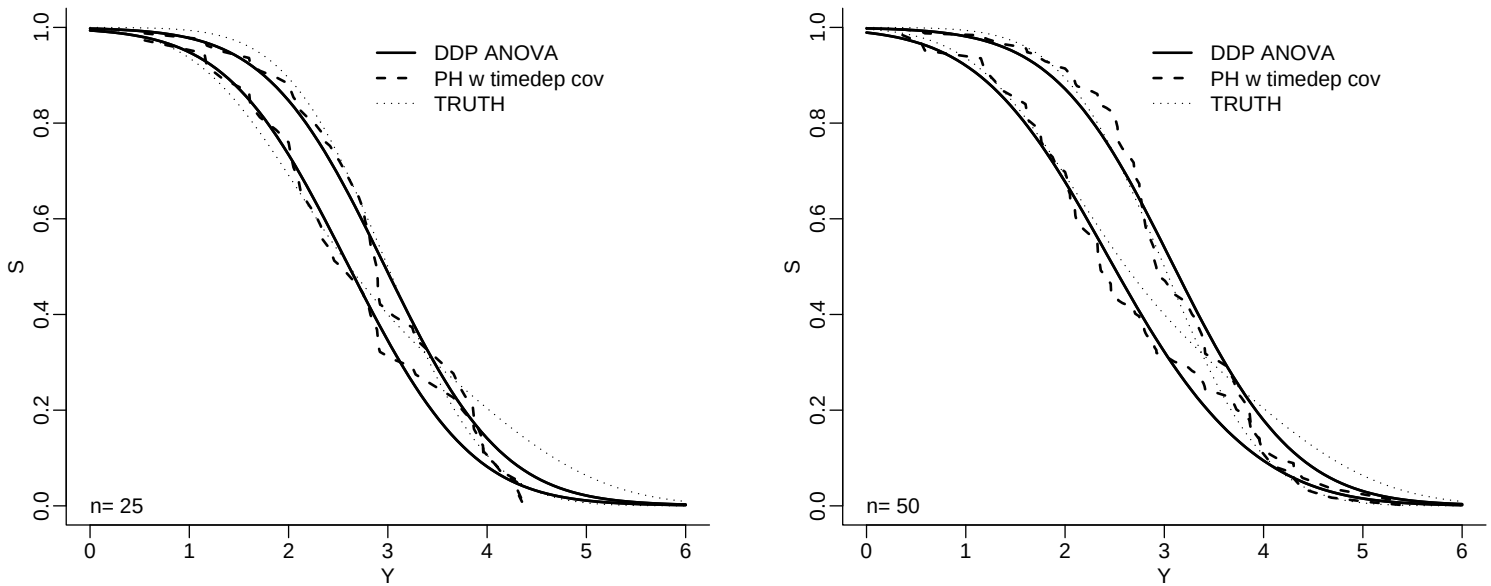
- The hazard rate for the PH model becomes

$$h(y \mid x, z) = \begin{cases} h_o(y)\exp(\beta_1 x) & \text{if } y \leq y^\star \\ h_o(y)\exp((\beta_1+\beta_2)x) & \text{if } y > y^\star \end{cases}$$

  where $h_o(y)$ is the baseline hazard rate.

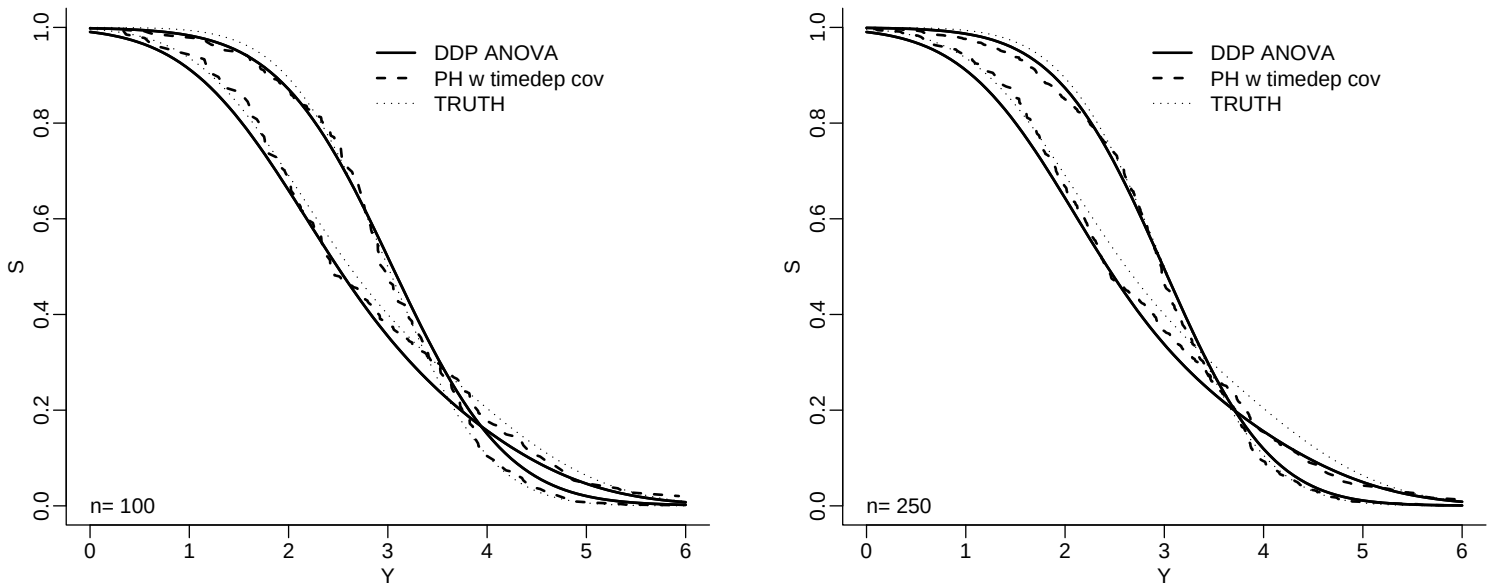
- $\exp(\beta_1)$ is the relative risk prior to time $y^\star$ for the high dose group relative to the low dose group
- $\exp(\beta_1+\beta_2)$ is the relative risk after time $y^\star$.
- $\exp(\beta_2)$ is the increase in relative risk after time $y^\star$, *change point* for the relative risk
- To fix $y^\star$: we fit the model on a grid of values for $y^\star$ and we choose $y^\star$ as the value with the largest log likelihood.

Results



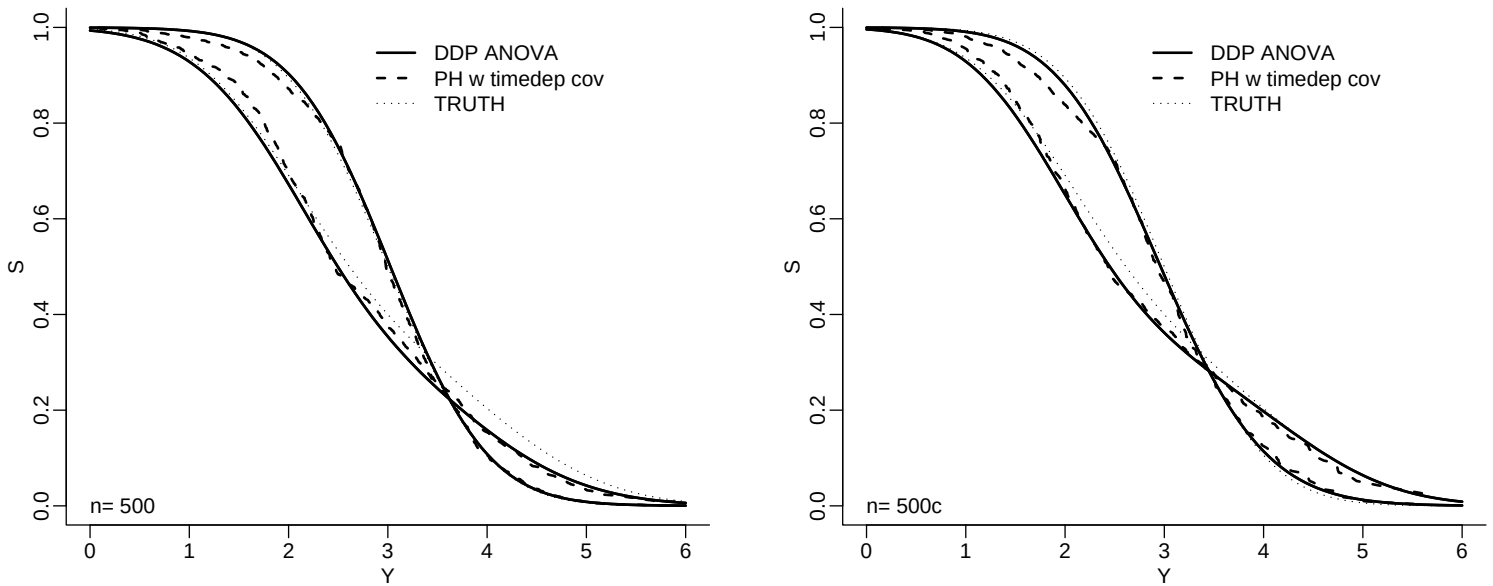
$n = 25$ on the left and $n = 50$ on the right

Results



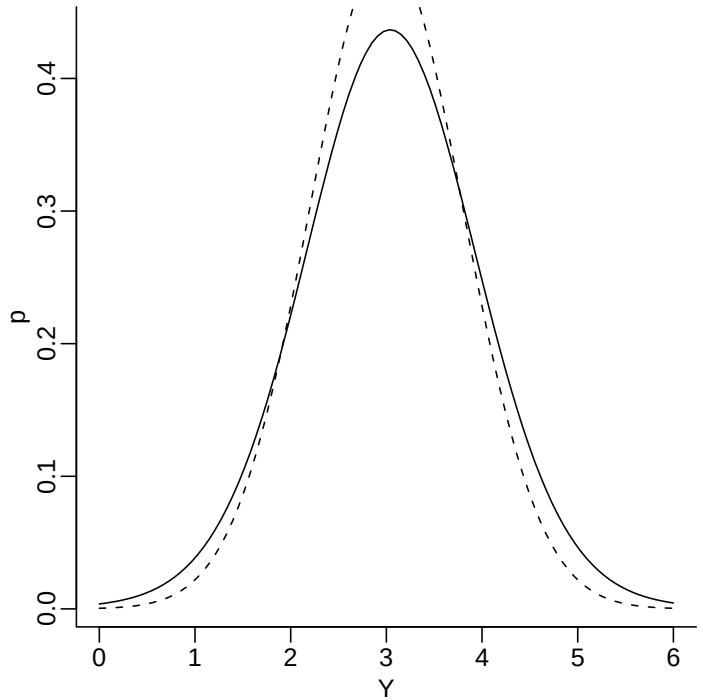
$n = 100$ on the left and $n = 250$ on the right

Results

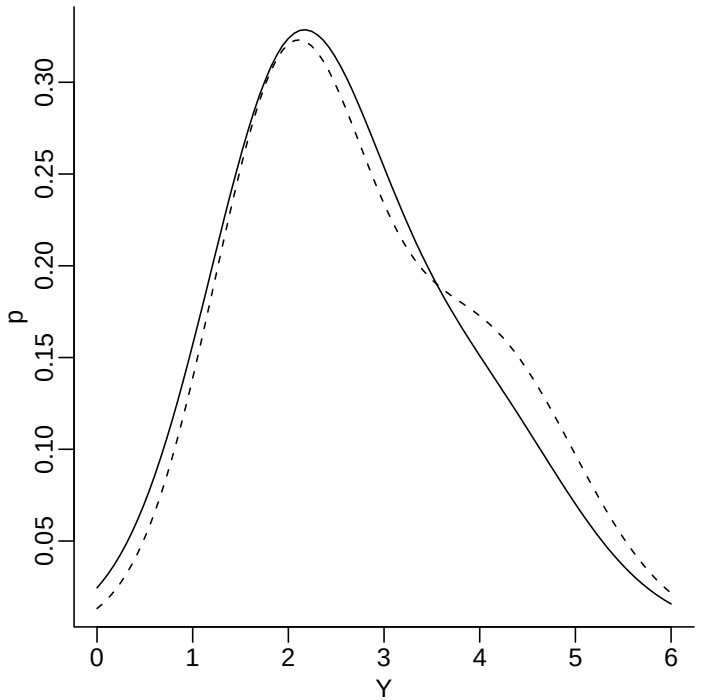


$n = 500$ on the left and $n = 500$ with 23% of censoring on the right

Results



$p_1(y)$



$p_2(y)$

Estimated distributions under the two groups with $n = 100$.

Cancer Clinical Trial

- 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.

Cancer Clinical Trial

- In the 1990s, the Cancer and Leukemia Group B 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 $\longrightarrow$ 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.

Data

- The data record the event-free survival time in months for 761 women
- 53% of the observations are censored
- We consider 3 categorical covariates plus an interaction term:
  - Treatment Dose (low/high)
  - Estrogen receptor (ER) status (pos/neg)
  - Tumor Size (TS) ($<$ 3.8 cm / $>$ 3.8 cm)
  - Dose by ER interaction

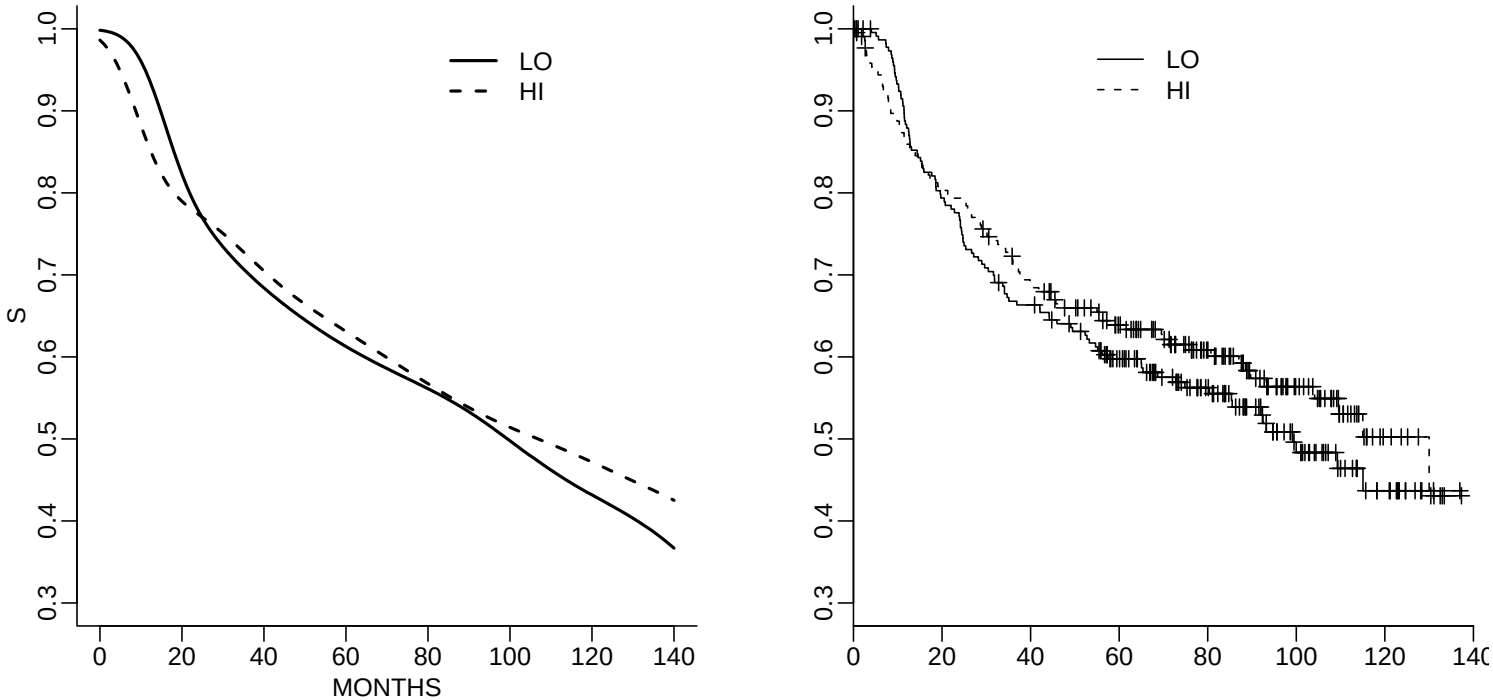
Summary of Cancer Data

t	(months)	ν	(freq.)	Dose	(freq.)	TS	(cm)	ER	(freq.)
Med	21.88	Cens	400	High	385	Mean	3.8	Pos	528
IQR	33.54	Event	361	low	376	STD	2.4	Neg	233

Goal

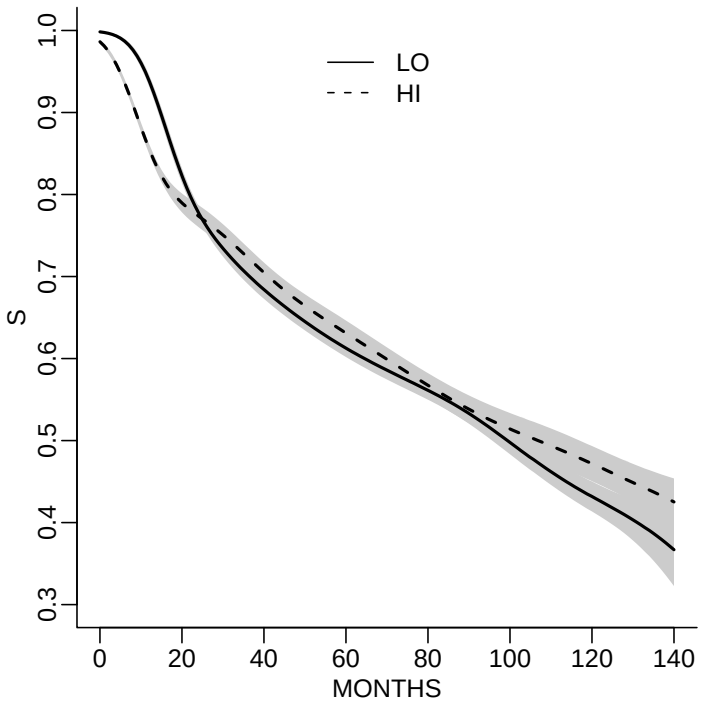
- The primary motivation was to compare low versus high dose
- Preliminary analysis: PH analysis using tumour size and ER status as covariates and stratifying by treatment dose
- Limitation: Difficulty to examine treatment effects
- Proposed model-based Bayesian inference provides a full probabilistic 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

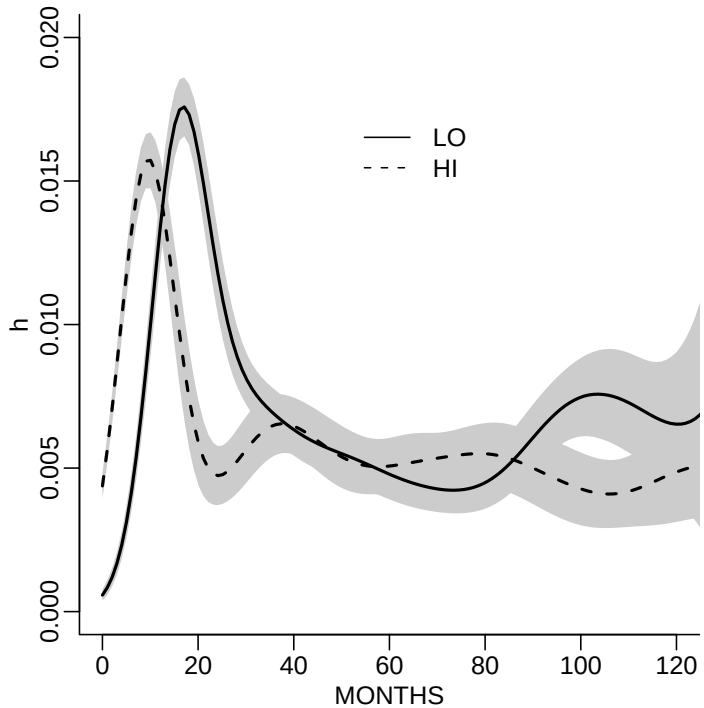
Results



Posterior estimated survivor functions for high versus low dose, for tumor size < 3.8 cm and ER positive. For comparison the right panel shows the data (combining positive and negative ER status).

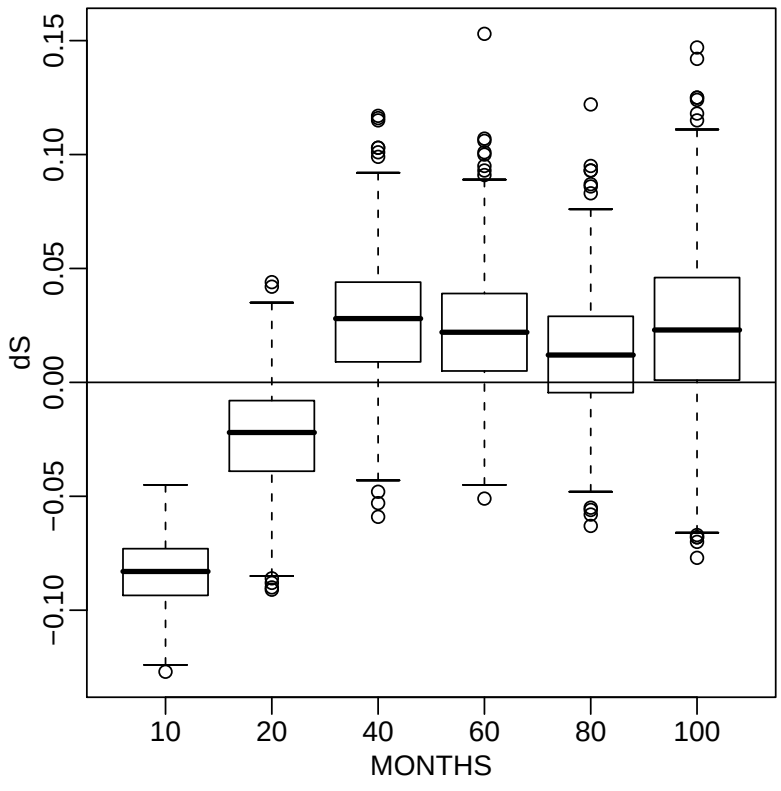
Results





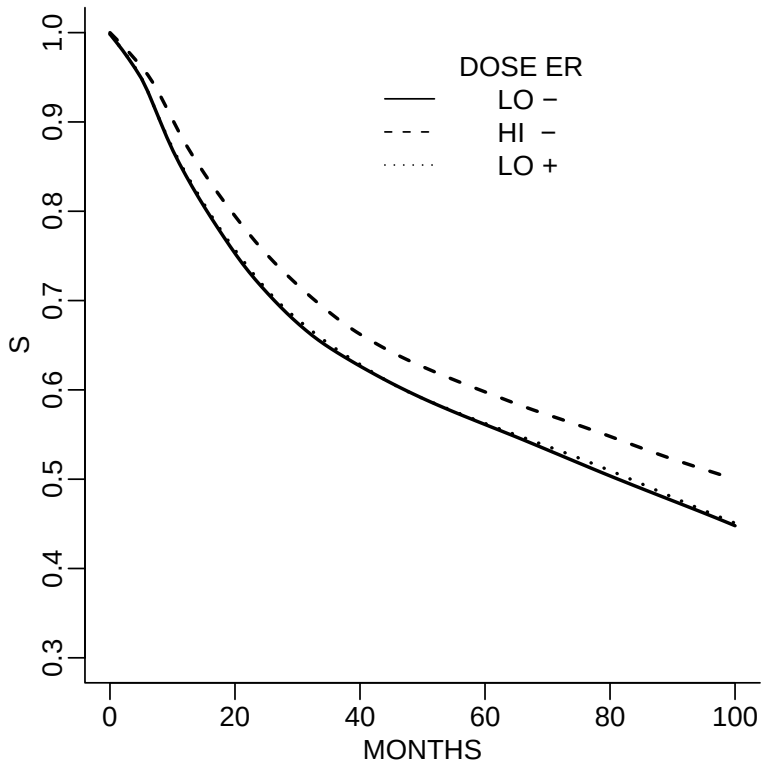
Posterior estimated survivor function and hazard hazard for high versus low dose, for tumour size $<$ 3.8 cm and positive ER status. The grey shaded bands show point-wise central 50% credible intervals.

Results

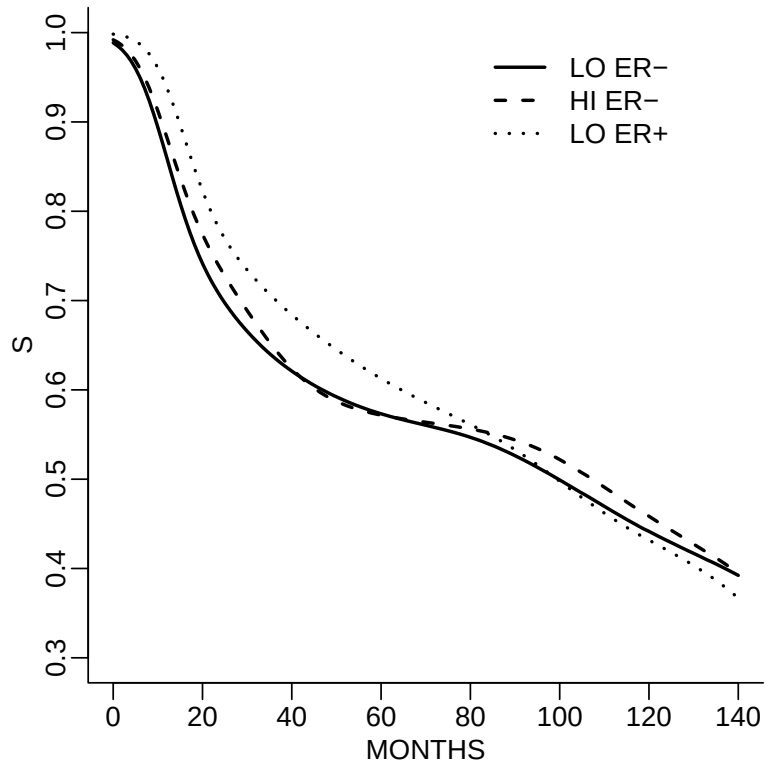


Posterior distribution of the difference in survival probabilities at 10, 20, 40, 60, 80 and 100 months between HD versus LD. The boxplots show the posterior distributions of the differences for positive ER status and tumour size less than 3.8 cm. Note how the difference changes sign from 20 to 40 months.

Comparison with AFT model



(a) AFT



(b) ANOVA DDP

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.

R package

Programs are available as a function in the R package `ddpanova` at `http://odin.mdacc.tmc.edu/~pm/prog.html`

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

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.
- We have extended the ANOVA DDP model to continuous covariates and to handle censored observations
- 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 DDP settings