Estimation of the Latent Distribution in Cure Survival Models using a Flexible Cox Model

Vincent Bremhorst Philippe Lambert

Université catholique de Louvain (UCL)
Institut de Statistique, Biostatistique et sciences Actuarielles (ISBA)

1 February 2013

Outline of the talk

Outline of the talk

Introduction to cure models

- Common hypothesis in classical survival models :

  $\hookrightarrow$ Any observed subject will experience the monitored event if the follow up is sufficiently long.

- Realistic assumption ?

  $\hookrightarrow$ How to deal with subjects who will never experience the event of interest ?

$\Rightarrow$ **Cure models** incorporate the unknown proportion of immune subjects in survival models.

Introduction to cure models

- They are two well known families of cure models :
  - The mixture models (Berkson and Gage 1952 ; Farewell 1982 and 1986 ; Sy and Taylor 2000, for example) ;
  - The promotion time models (Tsodikov 1998 ; Chen Ibrahim and Sinha 1999 ; Zeng, Yin and Ibrahim (2006) ; Liu and Shen (2009), for example).
- Some authors have also defined unified approaches (Yin and Ibrahim 2005 ; Cooner and al. 2007, for example).

Outline of the talk

The promotion time model

The promotion time model has a biological interpretation :

- Each subject is exposed to $N \sim Pois(\theta)$ carcogenic cells.
- Let Y be the incubation time of one cell (Y is often called **latent event time**).
- Assumptions :
  - The cancer mass of each cell is detected independently from each other.
  - Only one cell needs to be detected for subject to fail.

$$\Downarrow$$

  - $Y_1, ..., Y_N$ are independent with a common distribution $F(t)$.
  - The observed failure time is defined as $\mathbf{T = \min_i(Y_i)}$.
  - If $\mathbf{N = 0}$, the subject is not exposed and is considered as **cured**.

The promotion time model

- We enter the covariates through :
  - The mean parameter of the number N of carcogenic cells :

$$\theta(\mathbf{x}) = \exp(\beta_0 + \mathbf{x}'\boldsymbol{\beta})$$

  - The distribution of the latent event times $F(t|\mathbf{z})$ using a Cox PH model :

$$h(t|\mathbf{z}) = h_0(t)\exp(\mathbf{z}'\boldsymbol{\gamma}) \quad ; \quad S(t|\mathbf{z}) = S_0(t)^{\exp(\mathbf{z}'\boldsymbol{\gamma})}$$

- Unconditional (on N) population survival function :

$$\begin{aligned} S^*(t|\mathbf{x},\mathbf{z}) &= \exp\left[-\theta(\mathbf{x})F(t|\mathbf{z})\right] \\ &= \exp\left[-\exp\left(\beta_0 + \mathbf{x}'\boldsymbol{\beta}\right)\left(1 - S_0(t)^{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right)\right] \end{aligned} \tag{1}$$

The promotion time model : Identification issue

- Usual assumptions :
  - i) The vector **z** of covariates does not include an intercept to ensure the identifiability of the Cox PH model.
  - ii) The baseline distribution function $F_0(t) = 1 - S_0(t)$ is a proper distribution function : $lim_{t\to\infty} F_0(t) = 1$
- Under i) and ii) we can show that :
  - A) If the follow up of the study is sufficently long then model (1) is identifiable.
  - B) If the follow up of the study is not sufficently long then model (1) is not identifiable unless vectors **x** and **z** do not share the same components.

The promotion time model : Remarks

- $S^*(t|\mathbf{x}, \mathbf{z})$ is an improper survival function :

$$lim_{t\to\infty} S^*(t|\mathbf{x}, \mathbf{z}) = \exp\left[-\exp\left(\beta_0 + \mathbf{x}'\boldsymbol{\beta}\right)\right] > 0$$

- Remember that if the subject is not exposed to any carcogenic cell (**if N = 0**) then (s)he is considered as cured.

  $\hookrightarrow$ The probability of being cured is :

$$\begin{aligned} P(N=0) &= \exp\left[-\theta\left(\mathbf{x}\right)\right] \\ &= \exp\left[-\exp\left(\beta_0 + \mathbf{x}'\boldsymbol{\beta}\right)\right] \\ &= \lim_{t\to\infty} S^*(t|\mathbf{x}, \mathbf{z}) \end{aligned}$$

Outline of the talk

Estimation

- We estimate the latent distribution $F(t|\mathbf{z})$ of (1) using a linear combinaison of cubic B-splines on $\log(h_0(t))$:

$$\hat{h}_0(t) = \exp\left(\sum_{k=1}^{K} b_k(t)\phi_k\right) = \exp\left(B_t\boldsymbol{\phi}\right)$$

where $(b_k(.), k = 1, ...K)$ denote the cubic B-splines basis associated to a predefined number of equidistant knots defined on $[0, t_{Rcens}]$, where t_{Rcens} is the upper bound of the follow up.

Estimation

- Knowing the relation between the survival and the hazard functions, we obtain the estimation of the baseline survival function :

$$\hat{S}_0(t) = \exp\left(-\int_0^t \exp\left(B_u\phi\right) du\right)$$

$\hookrightarrow$ The integral in this expression has no analytic form and needs to be evaluated numerically. We use **the rectangle method**.

Estimation : Rectangle method

- Partition of $[0, t_{Rcens}]$ into J small intervals of equal width : $J_j = [\tau_{j-1}, \tau_j]$ where $0 = \tau_0 < \tau_1 < ... < \tau_J = t_{Rcens}$.
- Define
  - The middle of the j^{th} interval : $u_j = \frac{\tau_j - \tau_{j-1}}{2}$;
  - The length of the j^{th} interval : $\delta_j = \tau_j - \tau_{j-1}$

$\hookrightarrow$ The baseline survival function can be approximated by :

$$\hat{S}_0(t) \approx \exp\left(-\sum_{j=1}^{j(t)} \left[\exp\left(B_{u_j}\phi\right)\right]\delta_j\right)$$

where $j(t)$ corresponds to the interval containing t.

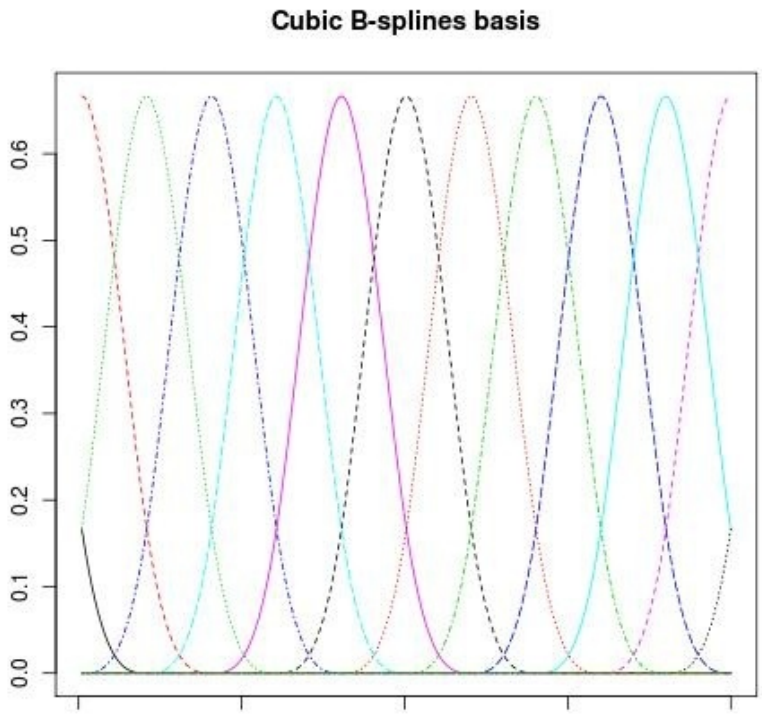
Regularisation

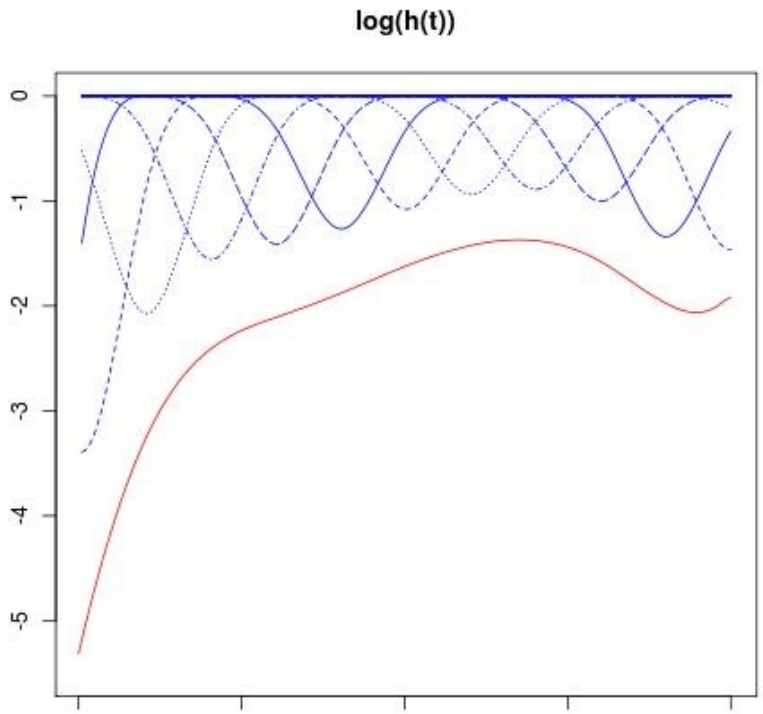
- **Roughness penalty** : We choose a large number (say K) of B-splines and countrebalance the flexibility by introducing a penalty on finite differences of adjacent B-spline parameters :
  $\tau \sum_k (\Delta^r \phi_k)^2 = \tau \phi' \mathbf{D}' \mathbf{D} \phi$, where τ is the penalty parameter. (Eilers and Marx 1996)

- For example, when we use a third order penalty, the matrix $\mathbf{D}$ is given by :

$$\mathbf{D} = \begin{bmatrix} 1 & -3 & 3 & -1 & 0 & \dots & 0 \\ 0 & 1 & -3 & 3 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -3 & 3 & -1 \end{bmatrix}$$

Graphical illustrations





- Number of knots : 12
- $b_k(t) > 0$ for only 3 values of k

- *blue curves* $\equiv \mathbf{B_t}\hat{\phi}$
- *red curve* $\equiv$ Estimation of $\log(h_0(t))$

Likelihood

- Contribution of the i^{th} subject to the likelihood :

$$L(D_i, \Phi) = \hat{S}^*(t_i|\mathbf{x}_i, \mathbf{z}_i)(\hat{h}^*(t_i|\mathbf{x}_i, \mathbf{z}_i))^{\nu_i}$$

where

  - $D_i = (t_i, \nu_i, \mathbf{x_i}, \mathbf{z_i})$;
  - ν_i is the event indicator ;
  - Φ is the set of all parameters specific to the model ;
  - $\hat{S}^*(t|\mathbf{x}, \mathbf{z})$ is the survival function defined in (1) where the baseline survival function $S_0(t)$ was substituted by $\hat{S}_0(t)$;
  - $\hat{h}^*(t|\mathbf{x}, \mathbf{z})$ is the corresponding hazard function.

Outline of the talk

Bayesian model

- In a Bayesian setting, the roughness penalty is translated into a prior distribution for ϕ :

$$\pi(\phi|\tau) \propto \tau^{\frac{\rho(\mathbf{P})}{2}} \exp\left(-\frac{\tau}{2}\phi'\mathbf{P}\phi\right) \qquad \text{where} \qquad \mathbf{P} = \mathbf{D}'\mathbf{D}$$

- We use a common non-informative prior for τ :

$$\pi(\tau) \propto G(a(=1), b(=10^{-4}))$$

- We use an improper uniform prior for all the regression parameters :

$$\pi(\beta_i) \propto 1 \quad \forall i \in I_{\boldsymbol{\beta}}$$

$$\pi(\gamma_j) \propto 1 \quad \forall i \in I_{\boldsymbol{\gamma}}$$

Posterior distribution

- Using the Bayes' theorem, the joint posterior distribution is given by :

$$\pi(\boldsymbol{\phi}, \tau, \beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{D}, I) \propto L(\mathbf{D}, \Phi) \pi(\boldsymbol{\phi} | \tau) \pi(\tau)$$

- Only the conditional posterior distribution for the penalty parameter τ comes from a well known family :

$$\pi(\tau | \boldsymbol{\phi}, D, I) \propto G\left(a + \frac{\rho(P)}{2}, b + \frac{\boldsymbol{\phi}' P \boldsymbol{\phi}}{2}\right)$$

MCMC algorithm

- Gibbs step for τ ;
- Adaptive univariate Metropolis step for ϕ, β_0, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (Haario et Al. 2001) ;
- Experience shows that the mixing of the posterior chain of the spline parameters can be very slow.

  $\hookrightarrow$ Applying the adaptive Metropolis step on a reparametrized posterior distribution can improve the mixing of the chain ;

  $\hookrightarrow$ We estimate the correlation structure of the spline parameters using the link between survival data and the Poisson GLM ;

  $\hookrightarrow$ We reparametrize the posterior distribution of the spline parameters using the obtained estimation.

Outline of the talk

Set up

- The baseline hazard function $h_0(t)$ is related to a Weibull distribution with mean 10.8 and standard deviation 5.64
- 2 covariates are taken into account :

$$W_1 \sim N(0,1) \qquad \text{and} \qquad W_2 \sim B(1,0.5)$$

- Both covariates enter through the parameter θ and through the Cox PH model : $\mathbf{x} = (W_1, W_2) = \mathbf{z}$
- Sample size : n $= 500$
- Number of replications : S $= 200$
- Splines are defined on 12 knots and a 3^{rd} order penalty is used

Investigated Scenarios

- In each scenario, a sufficient follow up is assumed.

Cured (%)	Cured censored (%)	Suceptible censored (%)	Global censoring (%)
20%	0%	0%	0%
20%	47%	8%	18%
40%	0%	0%	0%
40%	46%	7%	25%

Results of the simulations : Regression parameters

- 20% of cured individuals

Global censoring (%)		Estimation	$L_{95\%}$	$U_{95\%}$	ESE	RMSE
0%	$\beta_0 = 0.65$	0.624	0.396	0.888	0.129	0.017
	$\beta_1 = 1.2$	1.146	0.932	1.394	0.127	0.018
	$\beta_2 = 0.5$	0.521	0.227	0.809	0.144	0.021
	$\gamma_1 = -1$	-0.910	-1.169	-0.670	0.130	0.150
	$\gamma_2 = 2.5$	2.386	1.974	2.826	0.212	0.239
18%	$\beta_0 = 0.65$	0.625	0.369	0.933	0.142	0.020
	$\beta_1 = 1.2$	1.170	0.932	1.437	0.133	0.018
	$\beta_2 = 0.5$	0.528	0.199	0.847	0.163	0.027
	$\gamma_1 = -1$	-0.941	-1.204	-0.679	0.137	0.143
	$\gamma_2 = 2.5$	2.411	1.971	2.884	0.237	0.256

Results of the simulations : Regression parameters

- 40% of cured individuals

Global censoring (%)		Estimation	$L_{95\%}$	$U_{95\%}$	ESE	RMSE
0%	$\beta_0 = 0.4$	0.337	0.112	0.619	0.120	0.017
	$\beta_1 = 1.75$	1.620	1.384	1.889	0.132	0.032
	$\beta_2 = -0.75$	-0.668	-1.015	-0.352	0.182	0.040
	$\gamma_1 = -1$	-0.816	-1.082	-0.557	0.147	0.229
	$\gamma_2 = 2.5$	2.350	1.964	2.769	0.202	0.243
25%	$\beta_0 = 0.4$	0.355	0.081	0.687	0.144	0.022
	$\beta_1 = 1.75$	1.635	1.365	1.929	0.144	0.032
	$\beta_2 = -0.75$	-0.659	-1.079	-0.316	0.208	0.048
	$\gamma_1 = -1$	-0.837	-1.134	-0.552	0.157	0.218
	$\gamma_2 = 2.5$	2.369	1.942	2.856	0.226	0.250

Conclusions of the simulations

- Regression parameters :
  - The point estimates are quite similar ;
  - The width of the confidence interval for each parameter increases when we introduce right censoring ;
  - For each parameter, the RMSE is quite similar whatever the considered scenario.

Conclusion of the simulations

- Baseline survival function :

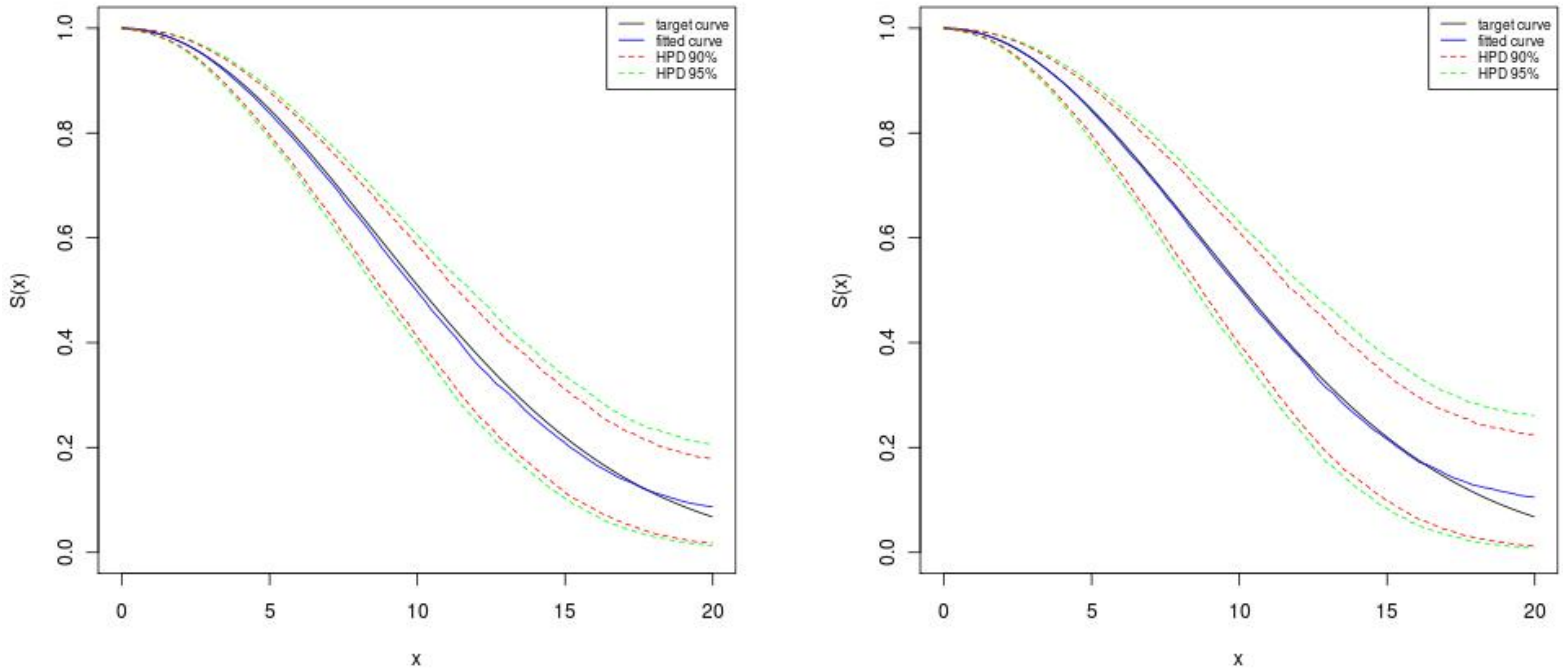


FIGURE: 20% of cured subjects. Left : Without right censoring ; Right : With right censoring

Conclusion of the simulations

- Baseline survival function :

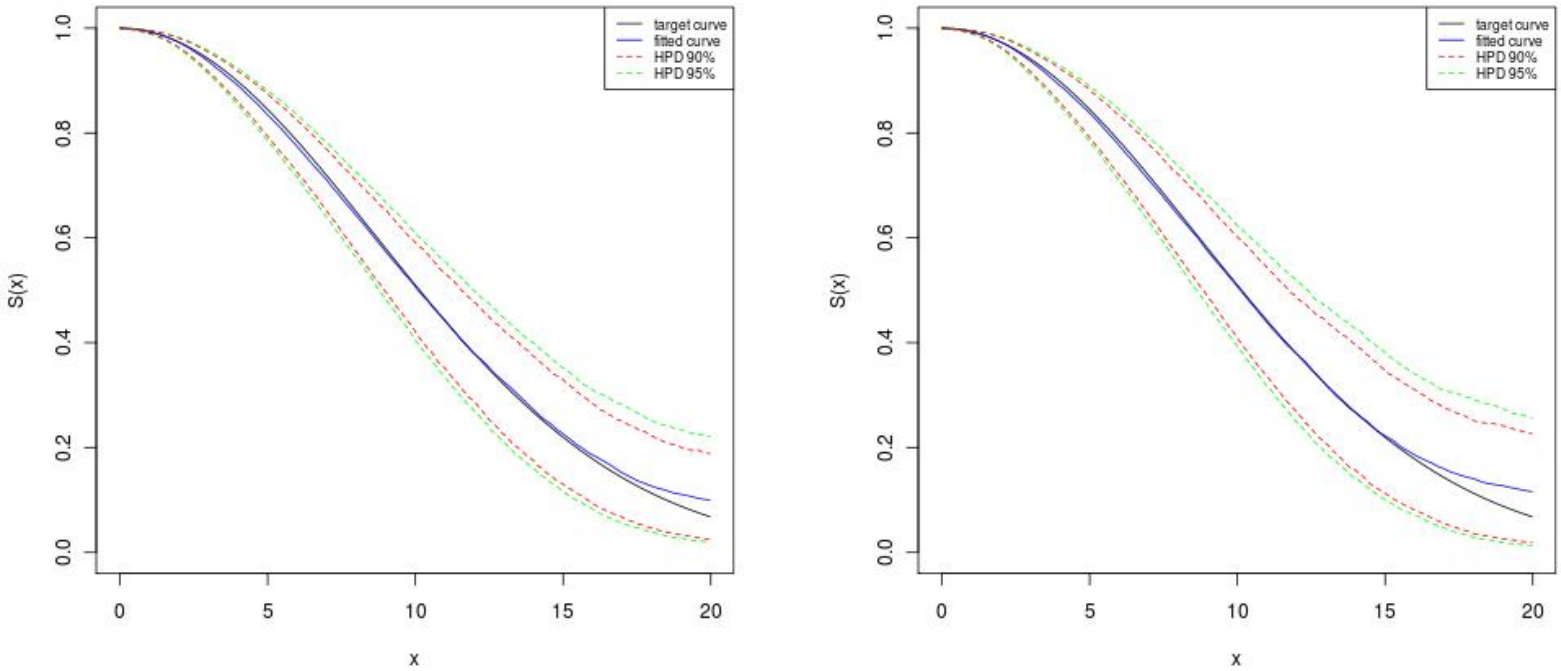


FIGURE: 40% of cured subjects. Left : Without right censoring ; Right : With right censoring

Further work

- Simulations using a bimodal distribution for the baseline distribution.
- Applying model (1) on a real data set.
- Can we find a covariates structure for $F(t|\mathbf{z})$ such that model (1) is identifiable without a sufficient follow up when $\mathbf{x}$ and $\mathbf{z}$ share some components ?
- Generalization to interval censored data.
- Generalization to hierarchical data.

Some references

[1] Chen, M.-H., Ibrahim, J.G. and Sinha, D. (1999), "A New Bayesian Model for Survival Data with a Surviving Fraction." *Journal of the American Statistical Association.* 94, 909-919.

[2] Cooner, F., Banerjee, S., Carlin, B.P. and Sinha, D. (2007), "Flexible cure rate modelling under latent activation schemes" *Journal of the American Statistical Association,* 102, 360-72.

[3] Eilers, P. H. C. and Marx, B. D. (1996) "Flexible smoothing with B-splines and penalties (with discussion)" *Statistical Sciences,* 11, 89-121.

[4] Lang, S., Brezger, A., 2004. "Bayesian P-splines." *J. Comput. Graphical Statist.*, 13, 183-212.

[5] Lambert, P. 2007. "Archimedean copula estimation using Bayesian splines smoothing techniques." *Computational Statistics and data Analysis*, 51,6307-6320.