A Dirichlet Process Mixture Regression Model for the Analysis of Competing Risk Events

Francesco Ungolo
School of Risk & Actuarial Studies, UNSW Business School, Sydney
Centre of Excellence in Population Ageing Research

joint work with Edwin R. van den Heuvel

Insurance Data Science conference
Stockholm, 17 June 2024

Introduction

- The analysis of competing risk events is common place in statistical analysis. It concerns the exposure of an individual to simultaneous events, the occurrence of one of these is absorbing for the others. For example:

Introduction

- The analysis of competing risk events is common place in statistical analysis. It concerns the exposure of an individual to simultaneous events, the occurrence of one of these is absorbing for the others. For example:
  - Death due to several different causes;

Introduction

- The analysis of competing risk events is common place in statistical analysis. It concerns the exposure of an individual to simultaneous events, the occurrence of one of these is absorbing for the others. For example:
  - Death due to several different causes;
  - Policyholders' lapse of an insurance contract, which may occur following termination of the contract due to surrending, death or default on premium payment.

Introduction

- The analysis of competing risk events is common place in statistical analysis. It concerns the exposure of an individual to simultaneous events, the occurrence of one of these is absorbing for the others. For example:
  - Death due to several different causes;
  - Policyholders' lapse of an insurance contract, which may occur following termination of the contract due to surrending, death or default on premium payment.
- Some (or all) events can be dependent, e.g. the lapse of an insurance policy to cash the value of the policy to cover the medical expenses.

Mathematical problem

- **Aim**: estimate the joint distribution of the time to M competing events $(T_1, \ldots, T_M)$. However, the researcher can only observe $T = \min(T_1, \ldots, T_M)$, and the corresponding cause of decrement C;

Mathematical problem

- **Aim**: estimate the joint distribution of the time to M competing events $(T_1, ..., T_M)$. However, the researcher can only observe $T = \min(T_1, ..., T_M)$, and the corresponding cause of decrement C;
- **Problem**: given the available data this joint distribution cannot be identified;

Mathematical problem

- **Aim**: estimate the joint distribution of the time to M competing events $(T_1, ..., T_M)$. However, the researcher can only observe $T = \min(T_1, ..., T_M)$, and the corresponding cause of decrement C;
- **Problem**: given the available data this joint distribution cannot be identified;
- $\Longrightarrow$ Further point identifying assumptions are thus needed, for example:
  - $T_1, ..., T_M$ are pairwise independent;
  - Specify a copula model with known dependence parameter;
  - Subdistribution approach.

Mathematical problem

- **Aim**: estimate the joint distribution of the time to M competing events $(T_1, ..., T_M)$. However, the researcher can only observe $T = \min(T_1, ..., T_M)$, and the corresponding cause of decrement C;
- **Problem**: given the available data this joint distribution cannot be identified;
- $\Longrightarrow$ Further point identifying assumptions are thus needed, for example:
  - $T_1, ..., T_M$ are pairwise independent;
  - Specify a copula model with known dependence parameter;
  - Subdistribution approach.
- There is an increasing need for of flexible and adaptable statistical models capable to handle complex problems, which can be interpretable at the same time.

The conditional independence model

- We assume that the time to each event are pairwise conditionally independent given a latent vector θ for the ith unit. We obtain the following joint density for $(T_1, ..., T_M)$:

The conditional independence model

- We assume that the time to each event are pairwise conditionally independent given a latent vector θ for the ith unit. We obtain the following joint density for $(T_1, ..., T_M)$:

$$
\begin{aligned}
f\left(t_{1,i}, \ldots, t_{M,i}\right) &= \int_{\Omega_\theta} \prod_{c=1}^{M} f_c\left(t_{c,i}; \theta_i\right) \mathrm{d}P\left(\theta_i\right) \qquad (1) \\
&= \int_{\Omega_\theta} \left[\prod_{c=1}^{M} f_c\left(t_{c,i}; \theta_{c,i}\right)\right] \mathrm{d}P\left(\theta_{1,i}, \ldots, \theta_{M,i}\right)
\end{aligned}
$$

The conditional independence model

- We assume that the time to each event are pairwise conditionally independent given a latent vector θ for the ith unit. We obtain the following joint density for $(T_1, ..., T_M)$:

$$f\left(t_{1,i},\dots,t_{M,i}\right)=\int_{\Omega_\theta}\prod_{c=1}^{M} f_c\left(t_{c,i};\theta_i\right)\mathrm{d}P\left(\theta_i\right) \tag{1}$$
$$=\int_{\Omega_\theta}\left[\prod_{c=1}^{M} f_c\left(t_{c,i};\theta_{c,i}\right)\right]\mathrm{d}P\left(\theta_{1,i},\dots,\theta_{M,i}\right)$$

- The latent multivariate random vector $\theta=(\theta_1,\dots,\theta_M)$ induces dependence among $T_1,\dots,T_M$. Its aim is to capture those latent features which are hidden, whilst affecting the joint occurrence of the M risks.

Dirichlet Process (1): Features

- A flexible specification for $P(\theta)$ is to assume it is a random draw from a Dirichlet Process (DP)

Dirichlet Process (1): Features

- A flexible specification for $P(\theta)$ is to assume it is a random draw from a Dirichlet Process (DP)

$$P(\theta) \sim \mathrm{DP}(\phi, P_0) \tag{2}$$

Dirichlet Process (1): Features

- A flexible specification for $P(\theta)$ is to assume it is a random draw from a Dirichlet Process (DP)

$$P(\theta) \sim \text{DP}(\phi, P_0) \tag{2}$$

where

- P_0 is the *base probability measure*, e.g. $P_0 : N(m_\theta, \Sigma_\theta)$;

Dirichlet Process (1): Features

- A flexible specification for $P(\theta)$ is to assume it is a random draw from a Dirichlet Process (DP)

$$P(\theta) \sim \text{DP}(\phi, P_0) \tag{2}$$

where

- P_0 is the *base probability measure*, e.g. $P_0 : N(m_\theta, \Sigma_\theta)$;
- ϕ is the concentration parameter

Dirichlet Process (DP): Introduction (2)

- A draw from a DP consists of a discrete probability distribution a.s., with the following representation:

Dirichlet Process (DP): Introduction (2)

- A draw from a DP consists of a discrete probability distribution a.s., with the following representation:

$$P(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

where $\{\theta_k\}_{k=1}^{\infty} \overset{\text{iid}}{\sim} P_0$;

Dirichlet Process (DP): Introduction (2)

- A draw from a DP consists of a discrete probability distribution a.s., with the following representation:

$$P(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

where $\{\theta_k\}_{k=1}^{\infty} \overset{\text{iid}}{\sim} P_0$; $\sum_{k=1}^{\infty} \pi_k = 1$

Dirichlet Process (DP): Introduction (2)

- A draw from a DP consists of a discrete probability distribution a.s., with the following representation:

$$P(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

  where $\{\theta_k\}_{k=1}^{\infty} \overset{\text{iid}}{\sim} P_0$; $\sum_{k=1}^{\infty} \pi_k = 1$ and $\delta_{\theta_h}(\cdot)$ is the Dirac delta function.

- The construction is completed through the following characterization of the mixture weights π_1, π_2, ..., as follows:

Dirichlet Process (DP): Introduction (2)

- A draw from a DP consists of a discrete probability distribution a.s., with the following representation:

$$P(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

  where $\{\theta_k\}_{k=1}^{\infty} \stackrel{\text{iid}}{\sim} P_0$; $\sum_{k=1}^{\infty} \pi_k = 1$ and $\delta_{\theta_h}(\cdot)$ is the Dirac delta function.

- The construction is completed through the following characterization of the mixture weights π_1, π_2, ..., as follows:

$$\pi_k = \psi_k \prod_{j=1}^{k-1} (1 - \psi_j); \quad \psi_k \stackrel{\text{iid}}{\sim} \text{Beta}(1, \phi)$$

Dirichlet Process (DP): Introduction (2)

- A draw from a DP consists of a discrete probability distribution a.s., with the following representation:

$$P(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

where $\{\theta_k\}_{k=1}^{\infty} \overset{\text{iid}}{\sim} P_0$; $\sum_{k=1}^{\infty} \pi_k = 1$ and $\delta_{\theta_h}(\cdot)$ is the Dirac delta function.

- The construction is completed through the following characterization of the mixture weights π_1, π_2, ..., as follows:

$$\pi_k = \psi_k \prod_{j=1}^{k-1} (1 - \psi_j); \quad \psi_k \overset{\text{iid}}{\sim} \text{Beta}(1, \phi)$$

This is also known as *stick-breaking process* (SBP, Sethuraman (1994))

Dirichlet Process (2): The Polya urn representation

- The DP implies that the distribution of θ_n given $\theta^{(-n)} = (\theta_1, \ldots, \theta_{n-1})$ can be characterized in terms of the Polya urn scheme (Blackwell-MacQueen (1973)):

$$p\left(\theta_n \mid \theta_{1:n-1}; \phi, P_0\right) \propto \frac{1}{\phi + n - 1} \sum_{i=1}^{n-1} \delta_{\theta_i}\left(\theta_n\right) + \frac{\phi}{\phi + n - 1} P_0\left(\theta_n\right). \quad (3)$$

where δ_θ is the Dirac delta function.

Dirichlet Process (2): The Polya urn representation

If we write $\theta^*_{1:n-1} = (\theta^*_1, \ldots, \theta^*_J)$, $J < n$, as the set of unique values of $\theta_{1:n-1}$

$$\theta_n \mid \theta_1, \ldots, \theta_{n-1} \sim \begin{cases} \text{Discrete}\left(\theta^*_1, \ldots, \theta^*_J\right) & \text{w.p.} \quad \frac{n_j}{n-1+\phi}, \quad \text{j=1,...,J} \\ P_0 & \text{w.p.} \quad \frac{\phi}{n-1+\phi} \end{cases} \tag{4}$$

where n_j are such that $\sum_j n_j = n - 1$, and denote the number of observations which are equal to θ^*_j.

The Dirichlet Process Mixture

- The density for the *i*th unit can be written as a Dirichlet Process Mixture:

$$
\begin{aligned}
f\left(t_{1,i},\ldots,t_{M,i}\right) &= \int \prod_{c=1}^{M} f_c\left(t_{c,i};\theta_i\right)\mathrm{d}P\left(\theta_i\right) \\
&= \sum_{k=1}^{\infty}\pi_k\left[\prod_{c=1}^{M} f_c\left(t_{c.i};\theta^*_{c,k}\right)\right]
\end{aligned}
\tag{5}
$$

The Dirichlet Process Mixture

- The density for the ith unit can be written as a Dirichlet Process Mixture:

$$f\left(t_{1,i},\ldots,t_{M,i}\right) = \int \prod_{c=1}^{M} f_c\left(t_{c,i};\theta_i\right)\mathrm{d}P\left(\theta_i\right) = \sum_{k=1}^{\infty}\pi_k\left[\prod_{c=1}^{M} f_c\left(t_{c.i};\theta^*_{c,k}\right)\right] \tag{5}$$

In this way, we aim at carrying out statistical inferences which are robust with respect to misspecification;

The Dirichlet Process Mixture

- The density for the *i*th unit can be written as a Dirichlet Process Mixture:

$$
\begin{aligned}
f\left(t_{1,i}, \ldots, t_{M,i}\right) &= \int \prod_{c=1}^{M} f_c\left(t_{c,i}; \theta_i\right) \mathrm{d}P\left(\theta_i\right) \\
&= \sum_{k=1}^{\infty} \pi_k \left[\prod_{c=1}^{M} f_c\left(t_{c.i}; \theta^*_{c,k}\right) \right]
\end{aligned} \tag{5}
$$

  In this way, we aim at carrying out statistical inferences which are robust with respect to misspecification;

- Unit cause-specific covariates $x_{c,i}$ can be easily included:

$$
f\left(t_{1,i}, \ldots, t_{M,i} \mid x_{1,i}, \ldots, x_{M,i}\right) = \sum_{k=1}^{\infty} \pi_k \left[\prod_{c=1}^{M} f_c\left(t_{c,i} \mid x_{c,i}; \theta^*_{c,k}\right) \right] \tag{6}
$$

Data - The `uslapseagent` data set

- 29,317 Whole life insurance policies (75% training set, 25% test set), from the R package CASdatasets (Dutang and Charpentier (2020));

Data - The `uslapseagent` data set

- 29,317 Whole life insurance policies (75% training set, 25% test set), from the R package CASdatasets (Dutang and Charpentier (2020));
- Observation period: 1st January 1995- 31th December 2008;

Data - The `uslapseagent` data set

- 29,317 Whole life insurance policies (75% training set, 25% test set), from the R package CASdatasets (Dutang and Charpentier (2020));
- Observation period: 1st January 1995- 31th December 2008;
- Time to event in quarters for $M = 3$ causes of decrement:
  - Surrending ($C = 1$);
  - Death ($C = 2$);
  - Other ($C = 3$)

Model

Milhaud and Dutang (2018) emphasize the need to focus on the time to surrending as this is relevant to recover the initial expenses.

Model

Milhaud and Dutang (2018) emphasize the need to focus on the time to surrending as this is relevant to recover the initial expenses.

- For each cause $c = 1, \ldots, M$, we assume

$$Y_c = \ln T_c = \beta_c x_{c,i} + \theta_{c,i} + \epsilon_{c,i} \quad \epsilon_{c,i} \sim N\left(0, \sigma_c^2\right), \tag{7}$$

  where $x_{c,i}$ is the vector of individual cause-specific covariates, β_c is the vector of regression coefficients and

Model

Milhaud and Dutang (2018) emphasize the need to focus on the time to surrending as this is relevant to recover the initial expenses.

- For each cause $c = 1, \ldots, M$, we assume

$$Y_c = \ln T_c = \beta_c x_{c,i} + \theta_{c,i} + \epsilon_{c,i} \quad \epsilon_{c,i} \sim N\left(0, \sigma_c^2\right), \tag{7}$$

  where $x_{c,i}$ is the vector of individual cause-specific covariates, β_c is the vector of regression coefficients and

$$\begin{aligned} \theta_i &= (\theta_{1,i}, \ldots, \theta_{M,i}) \sim P \\ P &\sim \mathrm{DP}\left(\phi, \mathrm{MVN}\left(m_\theta, \Sigma_\theta\right)\right) \end{aligned} \tag{8}$$

Model - Set of covariates

Surrending	Death	Others
Annual Premium (std)	Gender	Annual Premium
Accidental D. Rider (Yes)	UW Age	Premium frequency
DJIA	Living place	Accidental D. Rider
UW Age (0-54/55+)	Smoking	
Gender		
Premium Frequency (Ann/Infr.)		

Full Bayesian specification

- Furthermore, to complete the full Bayesian specification:

$$
\begin{aligned}
\phi &\sim \mathrm{Ga}\left(1,1\right) && (9)\\
m_\theta &\sim \mathrm{MVN}\left(\Lambda_1, 9I\right); && (10)\\
\Sigma_\theta &\sim \text{Inv-Wish}\left(\Lambda_3, \Lambda_4\right) && (11)\\
\beta_{c,p} &\sim \mathrm{N}\left(0,9\right) \quad c = 1,\ldots,M; \quad p = 1,\ldots,p_c && (12)\\
\sigma_c^2 &\sim \text{Inv-Gamma}\left(1,1\right) && (13)
\end{aligned}
$$

Full Bayesian specification

- Furthermore, to complete the full Bayesian specification:

$$\phi \sim \text{Ga}(1,1) \quad (9)$$

$$m_\theta \sim \text{MVN}(\Lambda_1, 9I); \quad (10)$$

$$\Sigma_\theta \sim \text{Inv-Wish}(\Lambda_3, \Lambda_4) \quad (11)$$

$$\beta_{c,p} \sim \text{N}(0,9) \quad c = 1, \ldots, M; \quad p = 1, \ldots, p_c \quad (12)$$

$$\sigma_c^2 \sim \text{Inv-Gamma}(1,1) \quad (13)$$

- Therefore parameters can be easily drawn in closed form within a Data Augmentation (to cope with censoring of the other failure times) MCMC scheme.

Model analysis - Regression coefficients

Table 1: Posterior summaries of β_{surr}.

Parameter	Mean	95% CI
Annual Premium	-0.0411	$(-0.0514; -0.0306)$
DJIA	-0.6450	$(-0.6563; -0.6351)$
Acc. death rider (Yes)	0.0534	(0.0233; 0.0883)
Gender (Female)	0.0310	(0.0095; 0.0501)
Paym. freq. (Annual and over)	0.0744	(0.0540; 0.0964)
Underwriting Age (less than 55yo)	-0.1036	$(-0.1379; -0.0733)$

Model analysis - Dependent time to event

We sample a large number of log-time to events Y, conditional on sampled mixture allocation component S, using the posterior mean of the parameters.

Table 2: Linear, Spearman and Kendall correlation coefficients between Y_{Surr}, Y_{Death} and Y_{Other}.

	Linear	**Spearman**	**Kendall**
Surrending - Death	0.325	0.325	0.217
Surrending - Other	0.297	0.297	0.197
Death - Other	0.889	0.726	0.532

Posterior predictive density for time to surrending

$$f\left(\tilde{t}_{\text{surr}} \mid \tilde{x}; \text{data}\right) = \int \underbrace{f\left(\tilde{t}_{\text{surr}} \mid \tilde{x}; \Delta\right)}_{\text{kernel}} \underbrace{p\left(\Delta \mid \text{data}\right)}_{\text{posterior}} \mathrm{d}\Delta \tag{14}$$

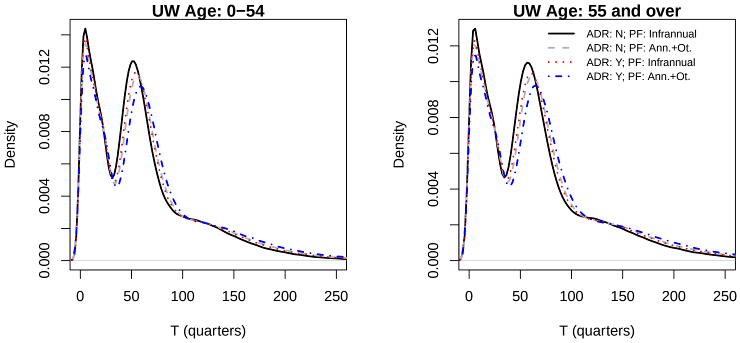


Figure 1: Posterior predictive density

Prediction of surrending rates (by quarter)

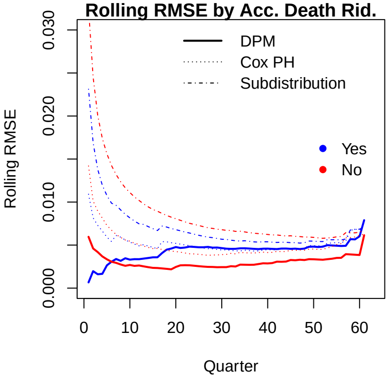
Surrending rate estimation by quarter s_q (using posterior mean of parameters):

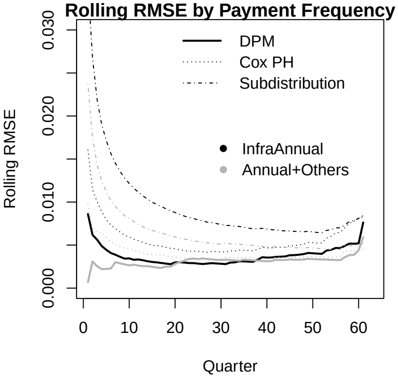
$$\hat{r}_q = \frac{1}{n_{s_q}} \sum_{i \in \mathcal{R}_{s_q}} \widehat{\Pr}\left(s_q < T_{1,i} \leq s_{q+1}, C_i = 1 \mid T_{1,i} > s_q, X_1, \ldots, X_3\right) \qquad (15)$$

where n_{s_q} is the size of at-risk population $\mathcal{R}_{s_q}$.

Prediction of surrending rates (Rolling RMSE by quarter)

$$\text{R-RMSE}_Q = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{r}_q^{\text{Model}} - r_q^{\text{Empirical}}\right)^2} \tag{16}$$





Composition of classes

- It can be possible to investigate the probability of the classes S_i, by using the Bayes' formula. Using posterior mean of the parameters, we have the individual posterior probability of each class

Composition of classes

- It can be possible to investigate the probability of the classes S_i, by using the Bayes' formula. Using posterior mean of the parameters, we have the individual posterior probability of each class

$$q_{i,k} = \Pr\left(S_i = k \mid t_i, x_{1,i}, \ldots, x_{M,i};\right) \tag{17}$$

$$= \frac{\pi_k \prod_{c=1}^{M} f_c\left(t_i \mid x_{c,i}; \theta^*_{c,k}\right)^{d_{c,i}} \left(1 - F_c\left(t_i \mid x_{c,i}; \theta^*_{c,k}\right)\right)^{1-d_{c,i}}}{\sum_{j=1}^{K} \left[\pi_j \prod_{c=1}^{M} f_c\left(t_i \mid x_{c,i}; \theta^*_{c,j}\right)^{d_{c,i}} \left(1 - F_c\left(t_i \mid x_{c,i}; \theta^*_{c,j}\right)\right)^{1-d_{c,i}}\right]}$$

Composition of classes

- It can be possible to investigate the probability of the classes S_i, by using the Bayes' formula. Using posterior mean of the parameters, we have the individual posterior probability of each class

$$q_{i,k} = \Pr\left(S_i = k \mid t_i, x_{1,i}, \ldots, x_{M,i};\right) \tag{17}$$
$$= \frac{\pi_k \prod_{c=1}^{M} f_c\left(t_i \mid x_{c,i}; \theta^*_{c,k}\right)^{d_{c,i}} \left(1 - F_c\left(t_i \mid x_{c,i}; \theta^*_{c,k}\right)\right)^{1-d_{c,i}}}{\sum_{j=1}^{K}\left[\pi_j \prod_{c=1}^{M} f_c\left(t_i \mid x_{c,i}; \theta^*_{c,j}\right)^{d_{c,i}} \left(1 - F_c\left(t_i \mid x_{c,i}; \theta^*_{c,j}\right)\right)^{1-d_{c,i}}\right]}$$

- The Bayes' rule hard assigns each individual to a specific class, by setting $s_i = k$ if $q_{i,k} > q_{i,j}$ for $j \neq k$

Analysis of classes

	Group 2	Group 4	Group 11	Group 15	Train. sample
% Composition	57.7	8.2	4.3	8.3	–
Annual Prem. (mean in $)	536.83	648.02	641.64	650.68	560.88
Accidental D. Rider (Yes in %)	17.4	14.0	13.8	12.5	16.4
Pr. Freq. (Ann+Oth in %)	41.8	30.4	34.9	29.8	38.9
UW Age (0-54 in %)	80.4	84.5	84.1	84.8	81.4
Surrending (in %)	14.7	100	100	92.5	38
θ^*_{Surr}	4.11	2.57	1.76	3.19	–
θ^*_{Death}	5.00	4.48	4.03	2.02	–
θ^*_{Other}	4.90	4.39	3.78	3.91	–

Conclusions

- The DPM allows for a flexible specification and identifiable joint model for the time to competing risk events;

Conclusions

- The DPM allows for a flexible specification and identifiable joint model for the time to competing risk events;
- Despite the simplicity of the parametric specification of the distribution of the time to event, this turns out to improve the predictive performance for the held-out dataset and to carry out robust statistical inferences;

Conclusions

- The DPM allows for a flexible specification and identifiable joint model for the time to competing risk events;
- Despite the simplicity of the parametric specification of the distribution of the time to event, this turns out to improve the predictive performance for the held-out dataset and to carry out robust statistical inferences;
  - The simple log-linear model is amenable of further improvements (e.g. a nonlinear regression model, generalized additive model);

Conclusions

- The DPM allows for a flexible specification and identifiable joint model for the time to competing risk events;
- Despite the simplicity of the parametric specification of the distribution of the time to event, this turns out to improve the predictive performance for the held-out dataset and to carry out robust statistical inferences;
  - The simple log-linear model is amenable of further improvements (e.g. a nonlinear regression model, generalized additive model);
  - Induction of a sparse prior to perform variable selection;

Conclusions

- The DPM allows for a flexible specification and identifiable joint model for the time to competing risk events;
- Despite the simplicity of the parametric specification of the distribution of the time to event, this turns out to improve the predictive performance for the held-out dataset and to carry out robust statistical inferences;
  - The simple log-linear model is amenable of further improvements (e.g. a nonlinear regression model, generalized additive model);
  - Induction of a sparse prior to perform variable selection;
- The model can be easily estimated by means of a fully Bayesian analysis which may account for the prior information of the researcher;

Conclusions

- The DPM allows for a flexible specification and identifiable joint model for the time to competing risk events;
- Despite the simplicity of the parametric specification of the distribution of the time to event, this turns out to improve the predictive performance for the held-out dataset and to carry out robust statistical inferences;
  - The simple log-linear model is amenable of further improvements (e.g. a nonlinear regression model, generalized additive model);
  - Induction of a sparse prior to perform variable selection;
- The model can be easily estimated by means of a fully Bayesian analysis which may account for the prior information of the researcher;
- An analysis of grouped units can be obtained as by-product to obtain further insight on the further unobserved sources of heterogeneity.

References

- C. Dutang and A. Charpentier (2020), *CASdatasets R Package*;
- X. Milhaud and C. Dutang (2018), *Lapse tables for lapse risk management in insurance: a competing risk approach*, European Actuarial Journal vol. 8, pag. 97–126 (2018)
- F. Ungolo and E. R. van den Heuvel (2024), *A Dirichlet Process Mixture model for the analysis of competing risk events*, Insurance: Mathematics and Economics, Issue 116, pag. 95-113;