Bayesian Nonparametric Modelling of Recurrent Event Processes

Maria De Iorio

Yale-NUS College

Dept of Statistical Science, University College London

Collaborators: Alessandra Guglielmi (Polimi, Milano, Italy)
Marta Talarita (UCL, London, UK), Giorgio Paulon (UT, Austin, US)
Francesca Ieva (Polimi, Milano, Italy) and James Malone-Lee (UCL, London, UK)

21th September 2018

Bayesian Computation for High-Dimensional Statistical Models
Outline

- Models for recurrent events
- AR(\(p\)) models for Gap Times
- Choosing order \(p\)
  - Urinary Tract Infection
  - Readmission Data
- Semiparametric Model for recurrent event with termination
  - Administrative Data Application ➔ HF
- Extensions and Conclusions
Recurrent Events

- events repeat over time
- large number of individual processes, small number of recurrent events (recurrent infections, asthma attacks, blood donations)

Examples:

- Initial event
- First recurrence
- Second recurrence
- ... $n_i - 1^{th}$ recurrence
- Censoring event
Example: UTI

Best clinical marker if UTI is pyuria.

Recurrent events for two patients: the last waiting time of the patient on the left is observed, while that of the patient on the right is censored. Red circles denote zero WBC at the visit while green circles denote WBC greater than 0.
Modelling Recurrent Events
(Cook and Lawless, 2007)

For a single process:

- event count process
- gap/waiting times between successive events
Modelling Event Counts

• event counts \( \{N(t), t \geq 0\} \), \( N(t) \) = number of events in \([0, t]\).

**Intensity function:**

\[
\lambda(t|H(t)) = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}(N(t + \Delta t^-) - N(t^-) = 1|H(t))}{\Delta t}
\]

\( H(t) = \{N(s), 0 < s < t\} \) history of the process at time \( t \)

\( \lambda(t|H(t)) \) instantaneous probability of an event occurring at \( t \), conditional on the process history

• useful when individuals frequently experience the events of interest, and the events are incidental (occurrence does not materially alter the process itself)
Modelling Gap Times

• gap/waiting times between successive events $W_1, W_2, W_3, \ldots$

\[ W_j = T_j - T_{j-1} \]

useful:
- when events are relatively infrequent
- when some type of individual renewal occurs after an event
- when prediction of the next time is of interest

**Interpretation:** $W_1, W_2, W_3, \ldots$ can be interpreted as observations on a time series at equally spaced time points $j = 1, 2, 3, \ldots$

Ex: recurrent infections (when an individual returns to a similar state after the infection has been cleared), recurrent hospitalizations
Models for Gap Times based on generalized regressions

For a single process/individual:

\[ L(W_1, W_2, \ldots, W_n) = \]
\[ L(W_1) \times L(W_2|W_1) \times L(W_3|W_1, W_2) \times \ldots \times L(W_n|W_1, \ldots, W_{n-1}) \]
Models for Gap Times based on generalized regressions

For a single process/individual:

\[
\mathcal{L}(W_1, W_2, \ldots, W_n) = \\
\mathcal{L}(W_1) \times \mathcal{L}(W_2 | W_1) \times \mathcal{L}(W_3 | W_1, W_2) \times \ldots \times \mathcal{L}(W_n | W_1, \ldots, W_{n-1})
\]

Model \(\mathcal{L}(W_j | W_1, \ldots, W_{j-1})\) for all \(j\):

- AFT models, e.g. \(\log W_j\) is Gaussian distributed, with mean depending on covariates, previous gap times, individual random effect
- PH models, where the hazards of \(W_j\) depend on covariates, previous gap times, individual random effect
Models for Gap Times

t = 0: start of all event processes (e.g. first infection);

[0, \tau_i]: time window

\( n_i - 1 \) observed events

\[ 0 := t_{i0} < t_{i1} < t_{i2} < \ldots < t_{in_i-1}, \quad w_{ij} := t_{ij} - t_{ij-1} \]

\[ w_{ini} := \tau_i - t_{ini-1} \quad n_i\text{-th event is censored} \]

Number of events per individual can be different.
Likelihood for $m$ individuals

\[
\mathcal{L} = \prod_{i=1}^{m} \left\{ \left( \prod_{j=1}^{n_i-1} f_j(w_{ij}\mid z_{ij}) \right) S_{n_i}(w_{in_i}\mid z_{in_i}) \right\}
\]

where $z_{ij} = (x_{ij}, w_{i1}, \ldots, w_{ij-1})$
The model - likelihood

\[ Y_{ij} := \log(W_{ij}), \quad x_{ij} \]  

covariate vector at recurrence \( j, \quad j = 1, \ldots, n_i, \)  

\[ i = 1, \ldots, n \]

\[ Y_{ij} | x_{ij}, \beta_j, \alpha_i, \sigma^2 \sim \mathcal{N}(x_{ij}^T \beta_j + \alpha_{ij}, \sigma^2) \]

The likelihood can be rewritten by considering the joint distribution of \( Y_i = (Y_{i1}, \ldots, Y_{in_i}) \)

\[ Y_i | \alpha_i, \beta_1, \ldots, \beta_J, x_i, \sigma^2 \sim \text{ind. } \mathcal{N}_{n_i}(m_i, \Sigma_i) \]

- Conditional independence between different patients
- \( W_{ij} \) has generalized AFT distribution, i.e. \( \log W_{ij} = z_{ij} \beta + \sigma \varepsilon_j \), where \( z_{ij} = (x_{ij}, 1) \)
- frailty term : \( \alpha_{ij} \) depends on the recurrence occasion \( j \), but also on the item \( i \)
- the regression parameters \( \beta_j \), expressing the effect of covariates \( x_{ij} \), depend on the recurrence occasion \( j \)
Nonparametric AR(p) Frailty Model

Autoregressive parameters $\alpha_i = (a_{i1}, \ldots, a_{in_i})$

$$\alpha_{ij} \mid m_{i0}, m_{i1}, \ldots, m_{ip}, \tau \overset{\text{ind.}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il} Y_{ij-l}, \tau^2)$$

$$j = p + 1, \ldots, n_i + 1$$

$$(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid G \overset{\text{iid}}{\sim} G$$

$$G \overset{\sim}{\sim} \text{DP}(M, G_0)$$

nonparametric prior

The distribution of $\alpha_{ij}$ for $j \leq p$, depends only on the available past observations as in any AR(p) model.
Why Bayesian Nonparametrics

• model flexibility and robustness, as parametric models are often impose unrealistic constraints.

• models we consider are in fact richly parametric (formally, using an infinite-dimensional parameter space) rather than nonparametric.

• Bayesian nonparametric models involve placing prior distributions on broad families of probability distributions; e.g. Mixtures of Polya trees (MPT) and Dirichlet Processes mixtures (DPM).
  
  ★ nice theoretical properties
  ★ ease of computation
  ★ ease of interpretability
  ★ extensions to more complex setups
  ★ accommodates for heterogeneity in population, over-dispersion of the observations, outliers
  ★ automatic clustering of individuals/identification of subgroups
• **Bayesian Nonparametrics (BNP) deals with infinite dimensional parameter spaces**
• Traditional prior distributions becomes **stochastic process priors**
• Focus on stochastic process on **discrete probability measures**

\[ G = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}, \]

Assign flexible prior to \( G \), e.g. DP, PT, Pytman-Yor, ...
Dirichlet Process (DP)

Probability model on distributions:

\[ G \sim DP(\alpha, G_0) \]

- Two parameters:
  - \( G_0 \): distribution on \((\Theta, A)\)
  - \( \alpha \): precision on \(\mathbb{R}^+\)

- \( G \) is a.s. discrete
Sethuraman’s stick breaking representation

\[ G = \sum_{h=1}^{\infty} w_h \delta_{\theta_h} \]

\[ \xi_h \sim \text{Beta}(1, \alpha) \]

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

\[ \theta_h \overset{iid}{\sim} G_0, \quad h = 1, 2, \ldots \]

where \( \delta(x) \) denotes a point mass at \( x \), \( \psi_h \) are weights of point masses at locations \( \theta_h \).
Dirichlet Process Mixtures (DPM)

In many data analysis applications the discreteness is inappropriate.

To remove discreteness: convolution with a continuous kernel

\[
f(y \mid \theta) = p(y \mid \theta) \\
\theta \mid G \sim G \\
G \sim \text{DP}(\alpha, G_0)
\]

\[\Rightarrow\] sampling model is

\[
f(y) = \int p(y \mid \theta)dG(\theta) \\
G \sim \text{DP}(\alpha, G_0)
\]
Dirichlet Process Mixtures (DPM)

... or with latent variables $\theta_i$

$$G \sim \text{DP}(\alpha, G_0)$$

$$\theta_i \sim G$$

$$f(y_i) = p(y_i \mid \theta_i)$$

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

Often $p(y \mid \theta) = N(\beta, \sigma^2) \longrightarrow f(y) = \sum_{h=1}^{\infty} \psi_h N(\beta_h, \sigma^2)$
Comment

Under $G : p(\theta_i = \theta_{i'}) > 0$

Observations share the same $\theta \Rightarrow$ belong to the same cluster

$\Rightarrow$ DPM induces a random partition of the observations \{1, \ldots, n\}

**Note:** in this case the clustering of observations depends only on the distribution of $y$. 
Going back to Recurrent Events model....

Gap times $Y_{ij} := \log(W_{ij}), j = 1, \ldots, n_i, i = 1, \ldots, n$

\[
Y_{ij} | x_{ij}, \beta_j, \alpha_i, \sigma^2 \sim \mathcal{N}(x_{ij}^T \beta_j + \alpha_{ij}, \sigma^2)
\]

Autoregressive parameters $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{in_i})$

\[
\alpha_{ij} | m_{i0}, m_{i1}, \ldots, m_{ip}, \tau \text{ ind.} \sim \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il} Y_{ij-l}, \tau^2)
\]

Random effect distribution:

\[
(m_{i0}, m_{i1}, \ldots, m_{ip}) | G \overset{iid}{\sim} G
\]

\[
G \sim DP(M, G_0)
\]

The model for $\alpha_i$ is DPM $\implies$ clustering based on $m_i$. 
Hyperpriors

\[ \beta_j \overset{iid}{\sim} \mathcal{N}_q(0, \beta_0^2 I_q) \]
\[ \sigma^2 \sim \text{Inv-Gamma}(a_\sigma, b_\sigma) \]
\[ \tau^2 \sim \text{Inv-Gamma}(a_\tau, b_\tau) \]
\[ M \sim \mathcal{U}(0, M_0) \]
\[ G_0 = \mathcal{N}(0, \sigma_g^2) \times \underbrace{\text{TBeta}(a_Z, b_Z) \times \cdots \times \text{TBeta}(a_Z, b_Z)}_{p \text{ times}}. \]

assuming a priori independence among the different parameters.

\text{TBeta}(a_Z, b_Z) \] denotes the translated Beta distribution defined on the interval \((-1, 1)\) with density proportional to
\[ (y + 1)^{a_Z - 1}(1 - y)^{b_Z - 1}1_{(-1,1)}(y). \]
Testing for the Order of Dependence

The order of dependence on past observations, $p$ is often unknown in applications, and it needs to be estimated.

Two approaches:

- Include in the base measure of the DP a spike and slab distribution on the autoregressive coefficient, leading to Spiked Dirichlet process prior introduced by Kim et al. (2009)

- specify directly a prior on $p$, and then, conditional on $p$, specify the prior distribution for the remaining parameters. The dimension of the parameter vector $(m_{i0}, m_{i1}, \ldots, m_{ip})$ changes according to $p$ and consequently the dimension of the space where the Dirichlet process measure is defined (Quintana & Müller (2012)).
306 female patients
28% censoring
975 gap times
include for each patient a 5-dimensional vector of time-varying covariates: the standardized age of the patient and symptoms (urgency, pain, stress incontinence and voiding symptoms).
Posterior Inference on $p$

Predictive marginal distributions of $m_i0$, $m_i1$, $m_i2$ and $m_i3$ obtained with spike and slab variable selection. No dependence on previous gap times.

Specifying a prior directly on $p$ leads to a posterior distribution with mode on 0.
Example: Hospital Readmissions

• we consider the *readmission* dataset in the R package *frailtypack*.

• The dataset contains rehospitalization times (in days) after surgery in patients diagnosed with colorectal cancer.

• The origin of the time axis is the date of the surgical procedure

• consider data on *N* = 197 patients, for a total number of 495 recurrent events.

• 60% censoring

• we include covariates:
  ✓ *chemo*: variable indicating if the patient received chemotherapy
  ✓ *sex*: gender of the patient
  ✓ *dukes*: classification of the colorectal cancer, ordinal (A-B, C, D)
  ✓ *charlson*: Charlson comorbidity index, ordinal (0, 1-2, 3), time-varying
Posterior Inference on $p$

- Predictive marginal distributions of $m_{i0}$, $m_{i1}$, $m_{i2}$ and $m_{i3}$ obtained with spike and slab variable selection.

- Note that the spike and slab prior allows to have clusters with different time dependency.

- Specifying a prior directly on $p$ leads to a posterior distribution with mode on 2.
Recurrent Events with Termination

• events repeat over time

• large number of individual processes, small number of recurrent events

• when a recurrent event process is terminated by another event (e.g. death), that is the absorbing state

• the terminating event may be observed or censored
Heart failure (HF)

✓ Heart failure is a **chronic** (progressive) condition in which the heart muscle is unable to pump enough blood through to meet the body’s needs for blood and oxygen.

✓ Symptoms: shortness of breath (dyspnea), cardiac asthma, stasis in the systemic circulation or in the portal circulation, edema, cyanosis, hypertrophy of the heart and unusual fatigue, lack of appetite or nausea, impaired thinking, increased heart rate.

✓ Causes: coronary artery disease including a previous myocardial infarction (heart attack), high blood pressure, atrial fibrillation, valvular heart disease, excess alcohol use, infection, and cardiomyopathy of an unknown cause.

✓ Diagnosis: based on the history of the symptoms and a physical examination with confirmation by echocardiography, and other tests.

✓ Treatment: e.g. pace-maker, drugs.
HF Burden

- HF is one of the main causes of morbidity, hospitalization and death in the western world and the related economic burden is relevant and expected to increase.
- HF prevalence significantly increases with age.
- Individuals are often admitted to hospitals and outpatient care services.
- Multiple readmission are burdensome to the patient and healthcare system.
- For example, it costs about 30 billion dollars each year in US (more than 6 million people in US are affected); about 80% of those costs are due to hospitalization (http://www.heart.org/)
- The average cost a HF-related event in Lombardia (most populated Italian region) is around 6000 Euro
HF data project

Italian Ministry of Health funded a project proposed by the management of the healthcare system in Lombardia

Participants: Politecnico di Milano (Dept of Management, Dept of Mathematics), Regione Lombardia, hospitals (Niguarda, San Carlo, etc)

Aim: to build a large and reliable database on patients hospitalized for HF, which should link data on hospitalizations, outpatient service utilization, and drug prescriptions and could be used for epidemiological purposes, cost analysis, risk prediction, and quality of care evaluation

In Italy there is a universalistic health care system: the regional offices serve as regulators and payers. Reimbursement to healthcare providers is based on a protocol.

Database was originally intended to help policy makers to better manage HF burden of the healthcare costs of Lombardia.
Administrative data are complete, systematic, continuous in time and free usually characterised by large numbers and provide an invaluable source to study prevalence and incidence of major diseases.

Disadvantages of the use of administrative data: lack of accuracy, difficulties in linkage, merging, different coding.

In our application no detailed information on relevant clinical markers is available.

These limitations are particularly relevant when studying conditions/diseases characterized by transition from chronic to acute phases and back, as for HF.
Administrative data are complete, systematic, continuous in time and free usually characterised by large numbers and provide an invaluable source to study prevalence and incidence of major diseases.

Disadvantages of the use of administrative data: lack of accuracy, difficulties in linkage, merging, different coding

In our application no detailed information on relevant clinical markers is available.

These limitations are particularly relevant when studying conditions/diseases characterized by transition from chronic to acute phases and back, as for HF

Need for flexible and robust models:
Administrative data

Administrative data are complete, systematic, continuous in time and free Usually characterised by large numbers and provide an invaluable source to study prevalence and incidence of major diseases.

Disadvantages of the use of administrative data: lack of accuracy, difficulties in linkage, merging, different coding

In our application no detailed information on relevant clinical markers is available.

These limitations are particularly relevant when studying conditions/diseases characterized by transition from chronic to acute phases and back, as for HF

Need for flexible and robust models: BNP models!
Data

**Recurrence process:** hospitalizations for HF of the subject

**Termination:** death of the subject

- identification of hospitalizations for HF (for all adults resident in Lombardia) from all hospitals in Lombardia, hospitalized for the first time from 01/01/2006 to 31/12/2012
- in-hospital deaths collected from the hospitals; out-hospital deaths collected from the vital statistics regional dataset
- huge dataset, more than 217 thousand subjects
- N = 810 patients (with at least 1 occurrence in addition to the first – the origin of time); 44% has right-censored survival time
Aim

• develop a joint model for waiting times between hospitalizations and survival within a BNP framework for data from HF in Regione Lombardia;

• clustering of subjects according to their history of recurrent events and termination;

• assessment of the relationship between event occurrence, survival and covariates; particularly important because administrative data may be inaccurate!

• time-to-event is the main clinical outcome of interest, but also the dynamics of the hospitalization process itself; controlling/reducing the costs of hospitalizations is important;

• model the relationship between survival times and recurrence of events; several re-hospitalizations are often related to risk of death and are likely to affect it;

• a better understanding of how recurrences affect survival may lead to a more effective planning of healthcare resources.
Joint Model for gap times and termination

\[ 0 =: T_{i0} < T_{i1} < T_{i2} < \cdots < T_{in} < T_i \leq \zeta_i \quad (= 31/06/2012) \]

Recurrent Events

Time to Event

**Gap Times:** \( W_{ij} = T_{ij} - T_{ij-1}, j = 1, \ldots, n_i \)

Last gap \( W_{in_i+1} \) time censored by \( \zeta_i \)

**Joint Model:** \( W_{i1}, \ldots, W_{in_i}, W_{in_i+1}, \mathcal{T}_i \)

**Covariates:** \( x_{ij} \) fixed-time and time-varying covariates influencing \( W_{ij} \)

\( z_i \) fixed predictors of \( T_i \)

**dependent censoring of gap times by termination:** semi-competing risks model

**Idea:** common frailty parameter shared by \( \mathcal{T}_i \) and \( W_{i1}, \ldots, W_{in_i} \) as in Huang and Liu (2007), and the prior is nonparametric (Brown and Ibrahim, 2003, BNP model for a longitudinal outcome used as predictor for the time-to-event)
A BNP semi-competing risks model: likelihood + BNP prior component

Model:

\[
\begin{align*}
\log W_{ij} | \beta_j, x_{ij}, \alpha_i, \sigma_i^2 & \sim \mathcal{N}(x_{ij}\beta_j + \alpha_i, \sigma_i^2), \quad j = 1, \ldots, n_i + 1; \\
\log T_i | z_i, \gamma, \alpha_i, \psi, \eta_i^2 & \sim \mathcal{N}(z_i\gamma + \psi\alpha_i, \eta_i^2); \quad i = 1, \ldots, N \\
(\alpha_i, \sigma_i^2, \eta_i^2) | G & \sim G, \quad G \sim \text{DP}(M, G_0)
\end{align*}
\]

\(x_{ij}\): \(p\) fixed-time and \(q\) time-varying covariates

\(z_i\): \(r\) fixed-time covariates

- The subject-specific random effect \(\alpha_i\) takes into account the dependent censoring of gap times by termination and the correlation between different gap times of the same subject \(i\)

- \(\alpha_i\) allows the clustering to depend on both gap times trajectories and survival outcome
Parametric Prior Component

Prior independence among parameters $\beta_0$, $(\beta_1, \ldots, \beta_J)$, $\gamma$, $\psi$, 
\{$(\alpha_i, \sigma_i^2, \eta_i^2)$\} 

$\beta_0 \sim \mathcal{N}_p(0, \beta_0^2 I_p)$

$\beta_1, \ldots, \beta_J \mid \mu := (\mu_1 \ldots, \mu_q)^T$, $\Sigma := \text{diag}(\tau_1^2, \ldots, \tau_q^2) \sim \mathcal{N}_q(\mu, \Sigma)$

$\mu_1, \ldots, \mu_q \sim \mathcal{N}(0, \sigma_\mu^2)$, $\tau_1^2, \ldots, \tau_q^2 \sim \text{Inv-Gamma}(a_\tau, b_\tau)$

$\gamma \sim \mathcal{N}_r(0, \gamma_0^2 I_r)$

$\psi \sim \mathcal{N}(0, \psi_0^2)$

$G_0 = \mathcal{N}(0, \alpha_0^2) \times \text{inv-Gamma}(a_\sigma, b_\sigma) \times \text{inv-Gamma}(a_\eta, b_\eta)$

$M \sim \mathcal{U}(a_M, b_M)$. 

Several variations, e.g. $\text{Var}(Y_{ij}|\text{par}) = \sigma_{ij}$ or $\eta_i^2 = \eta^2 \sim$ parametric marginal prior, different distributional assumptions
Application to HF dataset

\( N = 810 \) patients (with at least 1 occurrence in addition to the first - the origin of time); 44\% has right-censored survival time.

![Diagram showing number of patients versus number of gap times](image-url)
Fixed Time Covariates

measured at entrance in the study

- **gender** of the patient (1 if female): proportion of women for each $j$ is $\approx 50\%$

- **age** [years] of the patient at each hospitalization; empirical means:

<table>
<thead>
<tr>
<th>Women</th>
<th>Men</th>
<th>TOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>78.44</td>
<td>73.14</td>
<td>75.77</td>
</tr>
</tbody>
</table>

- **group**: clinical classification of the patient according to worldwide indicators (Mazzali et al., 2016)
  - G1: patient having HF as the cause of admission or complicating another cardiac disease (standard HF)
  - G2: patient with myocardial or cardiopulmonary diseases
  - G3: patient with Acute HF as a complication of other diseases or for whom HF is reported as comorbidity + G4 (3 subjects only).
  - **group**: binary covariate, =0 if G1 (560 patients), and 1 otherwise (250).
Time-varying covariates

- **rehab**: binary, =1 if any time during the hospitalization is spent in a rehabilitation unit; 11.78% of the hospitalizations, corresponding to 29.01% of the patients.

- **ic**: binary, =1 if at least a part of the hospitalization is spent in an intensive care unit; 11.95% of the hospitalizations, corresponding to 31.48% of patients.

- **n_com**: total number of comorbidities (e.g. dementia, weight loss, alcohol abuse, tumor, arrhythmia, hypertension, pulmonary disease, diabetes, psychosis, HIV/AIDS) for each hospitalization. Sample means for all patients observed at the $j$-th gap times range from 2.19 to 5.09, increasing with $j$.

- **n_pro**: total number of surgical procedures (e.g. CABG - Coronary Artery Bypass Grafting, PTCA - angioplasty) for each patient hospitalization. Sample means for all patients observed at the $j$-th gap times range from 0.03 to 0.15.

- summary of time-varying covariates included in the survival regression
95% Marginal posterior credible intervals of the regression coefficients $\beta_0$ (left) in the gap time and $\gamma$ (right) survival time likelihood components of the time-homogeneous covariates under our model.
Summary

• there is no effect of gender on the gap times, but a weak effect is detectable on the survival times (with a negative effect on the survival time for women)

• the group variable seems to be weakly relevant for the gap times (patients with non-standard pathology, i.e. not in group G1, have larger gap times); moreover, it has a negative influence on the survival time;

• the age variable is a predictor of both gap times and survival: in particular, there is evidence that the average time between hospitalizations and the survival times are shorter for older patients.
Time-varying Covariates

Rehabilitation status

Intensive care unit status

# of comorbidities

# of surgical procedures
Summary

• Time-varying covariates are reported at the end of each gap time.

• In general time-varying covariates do not appear to have a strong effect on the recurrence process, except for few early gap times.

• As expected, the uncertainty on the effect estimates increases over time, due to the smaller number of available observations.

• \( ic \) has an effect on the distribution of waiting times, as patients with \( ic \) equal 1 show shorter gap times.

• Both a large number of comorbidities and a large number of surgical procedures yield frequent hospitalizations at the beginning of the study, followed by an opposite effect for later gap times.
Posterior Inference on $\psi$

- $\psi$ reflects the strength of the relationship between the two processes.
- Posterior of $\psi$ is centred away from zero, on the positive axis, as the time between hospitalizations widens, the probability of survival increases as well.
- Consistent with the fact that rehospitalization is common among patients affected by HF and that the course of this disease is often characterized by repeated hospital admissions at relatively short intervals and a limited prognosis for survival.

![Marginal posterior density of $\psi$.](image)
Cluster model

\( \alpha_1, \ldots, \alpha_n \) are a sample from a DP, i.e. a sample with ties:

\[ i \text{ and } j \text{ are in the same cluster } \iff \alpha_i = \alpha_j \]

If \( \rho = \{ C_1, \ldots, C_k \} \) is a partition of the data index set \{1, \ldots, n\}, then the model induces a prior for \( \rho \):

\[ \pi(\rho) = \mathbb{P}(\rho = \{ C_1, \ldots, C_k \}) = f(\#C_1, \ldots, \#C_k), \]

where \( f \) is an infinite exchangeable partition probability function (DP in this case).

Cluster estimates based on the estimate of \( \pi(\rho|\text{data}) \) minimizing posterior expected value of the Binder’s loss function - with equal weights (Lau-Green, 2007).

- clustering of the individuals in the sample is based on the frailty parameter \( \alpha_i \)'s, i.e. based on the whole trajectories and the survival time.
Clustering Inference

α

Kaplan Meier Estimates

\[ p < 0.0001 \]
Summary

✓ The largest cluster of patients (56.85% of the patients, cluster 2) is characterized by large survival times.

✓ Other 2 clusters (3 and 5) include mostly censored observations and are characterized by large survival times and long gap-time trajectories. However, Cluster 3 shows shorter gap times with a large average number of hospitalizations, whereas Cluster 5 includes young patients with the largest waiting times but with only few recurrent events.

✓ Clusters 1, 4 and 6 present shorter time intervals between hospitalizations compared to the others clusters as well as shorter survival times.

✓ The percentage of patients with standard pathology (i.e. group=0) is similar to the overall rate (≈ 70%) in each cluster but in Cluster 5, where it is 47%.
Extensions

Include autoregressive term in the model:

\[
\log W_{ij} \mid \text{all the rest} \overset{\text{ind}}{\sim} \mathcal{N} \left( x_{ij} \beta_j + \alpha_i + \sum_{k=1}^{r} \delta_{ik} \log W_{ij-k}, \sigma_i^2 \right)
\]

\[
\log T_i \mid \text{all the rest} \overset{\text{ind}}{\sim} \mathcal{N} \left( z_{i} \gamma + \psi \alpha_i, \eta_i^2 \right); \quad i = 1, \ldots, N
\]

with parametric or nonparametric prior on the autoregressive coefficients \( \delta_{ik} \)

Extend the BNP Autoregressive Frailty Model to include a survival outcome:

\[
T_i \mid \gamma_i, z_i^*, \rho_i \sim f(t_i \mid z_i^* \gamma_i, \rho_i)
\]

\[
\log W_{ij} \mid x_{ij}, \beta_j, \alpha_i, \sigma^2 \sim \mathcal{N}(x_{ij}^T \beta_j + \alpha_{ij}, \sigma^2)
\]

\[
\alpha_{ij} \mid m_{i0}, m_{i1}, \ldots, m_{ip}, \tau \overset{\text{ind.}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il} \log W_{ij-l}, \tau^2)
\]

\[
(\rho_i, m_{i0}, m_{i1}, \ldots, m_{ip}) \mid G \overset{iid}{\sim} G, \quad G \sim DP(M, G_0)
\]

We link survival and recurrence process by setting a joint nonparametric prior. This is in the spirit of the Random Partition Model with Covariates (Müller & Quintana, 2010)

Model time course of longitudinal markers
Conclusions

✔ HF among major causes of hospitalization and death in population and the main reason for hospital admission in patients ≥ 65 in western countries; urgent need for methods supporting healthcare management.

✔ We have highlighted important determinants of repeated hospitalization as well as survival. For example, advanced age, morbidity load, ic and rehabilitation. The variable pathology has an effect on gap times and reflects different HF protocols.

✔ Profiling patients according their risk profile should help the implementation of appropriate intervention strategies.

✔ BNP models, which are flexible and robust, can be useful because administrative data lack accuracy.

✔ Joint semiparametric model for recurrent HF hospitalizations and time to death; our approach jointly models survival and the hospitalizations times, specifying a DP as random effect distribution of the frailty parameter that links the survival and gap time trajectories.
Wrap-up and Acknowledgements

✓ models for gap times, since events are relatively infrequent and prediction is of interest

✓ recurrent events: rehospitalizations of patients due to Chronic Heart Failure, UTIs, readmission after surgery for cancer patients

✓ easy to generalize PH or AFT models to more general models with $AR(p)$ dependence on the gap times or on the random effects

✓ BNP priors in order to model more flexible $AR(p)$ parameter distribution and to achieve cluster estimates of patients

✓ BNP prior to model shared frailty between survival and recurrence process

✓ other Bayesian nonparametric priors could be OK, at the cost of more expensive computations

✓ future work: much richer dataset, i.e. wider sample size and deeper understanding of new and old explanatory variables, spatial information, longitudinal markers; parallel algorithms for inference for the complete dataset; or BNP covariate-driven clustering models

✓ Support received by the Royal Society (UK) through the International Exchanges Scheme 2014/R2 (inc CNRS).