

Gamma- and Beta-Tilted Beta Processes

BNP 9, Amsterdam, June 2013

Nils Lid Hjort

University of Oslo

- * **Beta processes** are versatile components for Bayesian non-parametrics model building – survival and event history analysis, growth functions, Indian Buffet Processes, etc.
- * They may be constructed from different perspectives – fine limits of cumsums of small Beta increments, Lévy processes, Poisson measures, etc. Each of these angles lends itself to **potential extensions**.
- * Here I introduce and examine **tilted Beta processes**, with $A = \{A(t): t \geq 0\}$ the limit of

$$A_m(t) = \sum_{i/m \leq t} H(p_i, \gamma) = \sum_{i/m \leq t} p_i^\gamma,$$

with $p_i \sim \text{Beta}[ca_0(i/m)(1/m), c\{1 - a_0(i/m)(1/m)\}]$.

PLAN:

A. Mini-review of Beta processes:

Construction; Lévy measure; jumps; prior and posterior for cumulative hazard rates; Bernshteĭn–von Mises theorems; random growth curves; other random curves; mixtures; shock models; Indian buffet processes.

B. Tilting the Beta processes:

To what extent can some or all of the above constructions and properties be followed through, generalised, and implemented?
How useful is **the extension**

from $\text{Beta}(c, A_0)$ to $\text{TiltBeta}(c, A_0, \gamma)$?

Does this invite further models and further uses?

C. Concluding remarks and some pointers to more research.

A: Mini-review of Beta processes

(NLH, 1990; review paper NLH and Kim, 2013)

1. Construction: Cumsum process of many small Betas:

$$A_m(t) = \sum_{i/m \leq t} p_i,$$

where $p_i \sim \text{Beta}[ca_0(i/m)(1/m), c\{1 - a_0(i/m)1/m\}]$,

$$\mathbb{E} p_i = a_0 \left(\frac{i}{m} \right) \frac{1}{m}, \quad \text{Var } p_i = \frac{1}{c+1} a_0 \left(\frac{i}{m} \right) \frac{1}{m} \left\{ 1 - a_0 \left(\frac{i}{m} \right) \frac{1}{m} \right\}.$$

We have $A_m \rightarrow_d A$ in the space of non-decreasing right-continuous functions with all jumps $\Delta A(t) \leq 1$. Also,

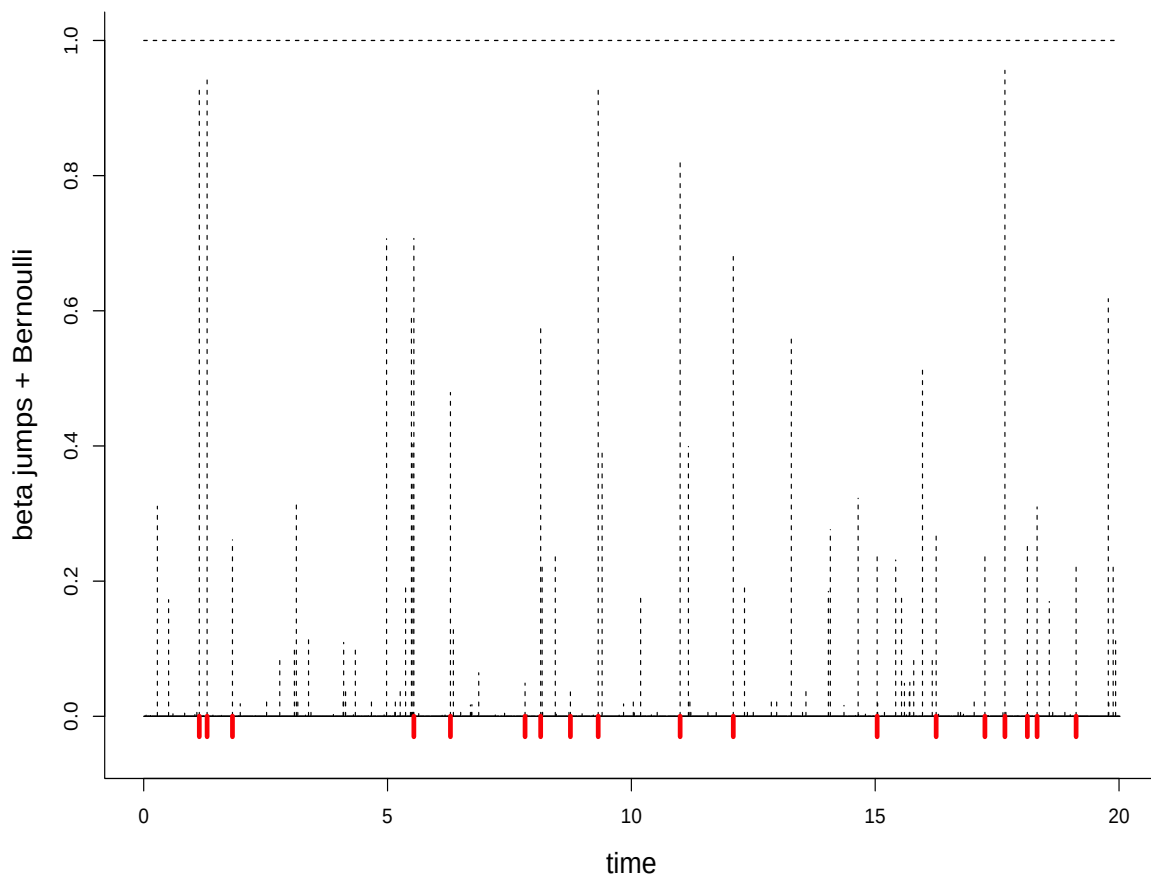
$$\mathbb{E} A(t) = A_0(t), \quad \text{Var } A(t) = \int_0^t \frac{dA_0(s)}{c+1}.$$

2. Lévy measure:

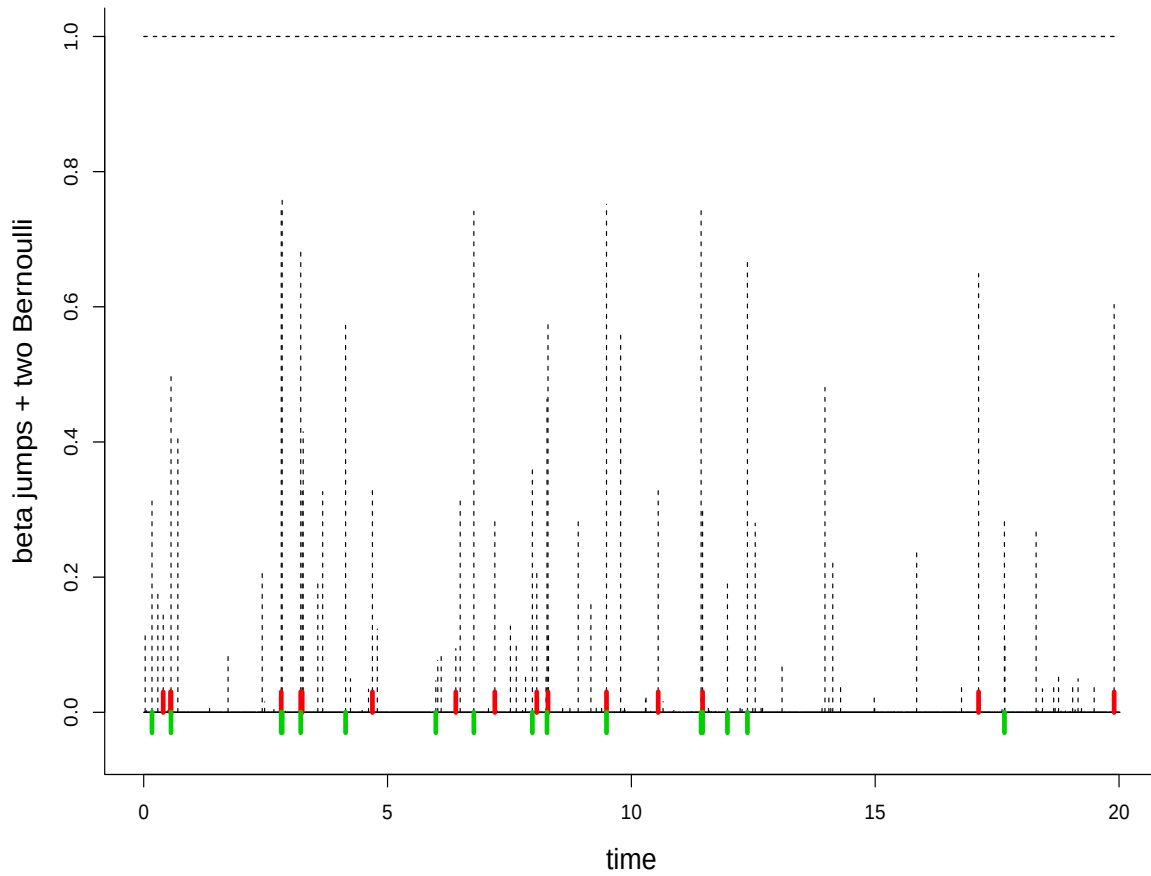
$$\mathbb{E} \exp\{-\theta A_m(t)\} = \prod_{i/m \leq t} \mathbb{E} \exp(-\theta p_i) \rightarrow \exp\{-D(\theta, t)\},$$

$$D(\theta, t) = \int_0^1 (1 - e^{-\theta x}) dL_t(x), \quad dL_t(x) = cx^{-1}(1-x)^{c-1} dx A_0(t).$$

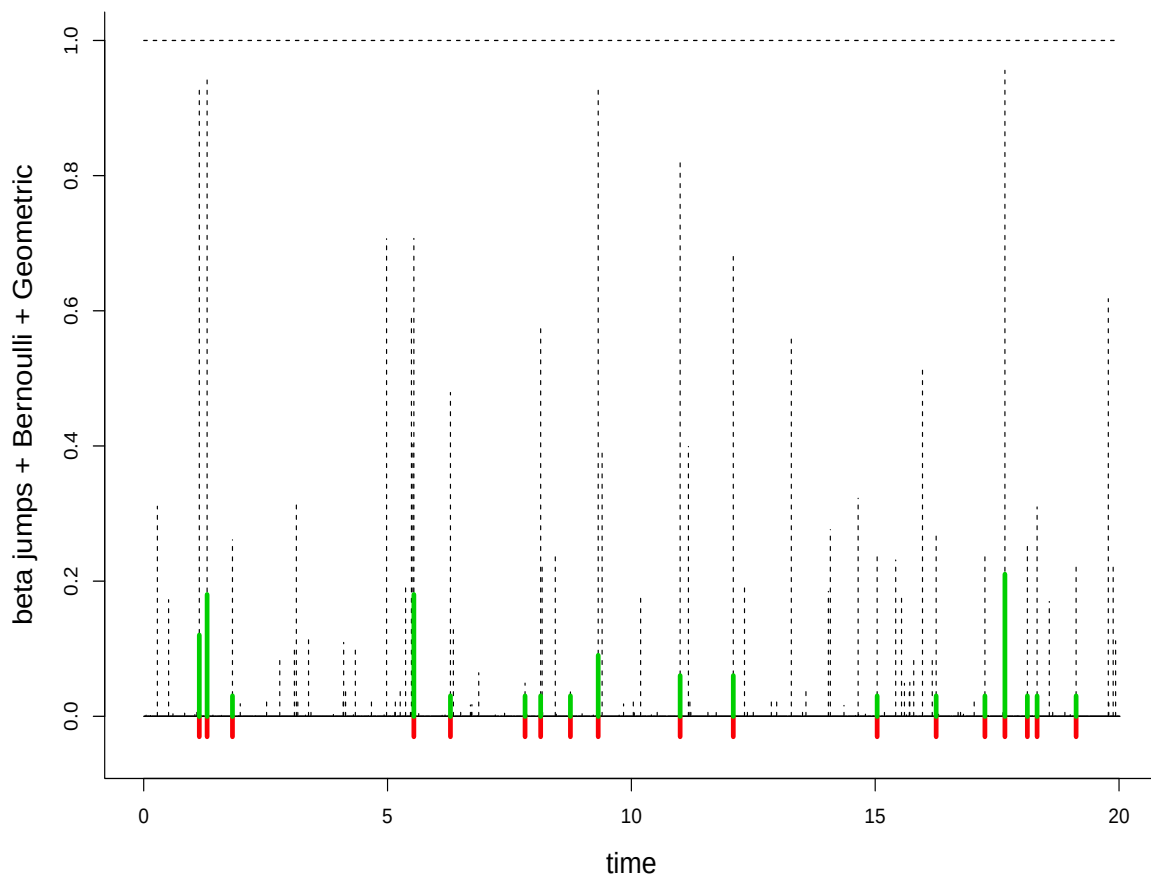
There is a link from this to **Poisson measure**.



Beta + Bernoulli = Poisson: *Beta process jumps p_i (dotted lines) with $Y_i \sim \text{Bin}(1, p_i)$.*



Beta + two Bernoullis = two dependent Poissons: *Beta process jumps p_i (dotted lines) with $X_i \sim \text{Bin}(1, p_i)$ and $Y_i \sim \text{Bin}(1, p_i)$; two Poisson processes, with correlation $1/(c + 1)$.*



Beta + Bernoulli = Poisson, Beta + Geometric: *Beta process jumps p_i (dotted lines) with $Z_i \sim (1 - p_i)p_i^z$ and $Y_i = I(Z_i \geq 1)$.*

3. Prior and posterior for cumulative hazard rates: Consider survival data (t_i, δ_i) from a distribution with cumulative hazard rate

$$A(t) = \int_0^t \frac{dF(s)}{F[s, \infty)}.$$

If $A \sim \text{Beta}(c(\cdot), A_0(\cdot))$, then $A \mid \text{data}$ is $\text{Beta}(c + Y, \hat{A})$, with

$$\hat{A}(t) = \int_0^t \frac{c dA_0 + dN}{c + Y}.$$

Here

$$dN(s) = \#\{\text{transitions observed in } [s, s + ds]\},$$

$$Y(s) = \#\{\text{at risk at time } s\}.$$

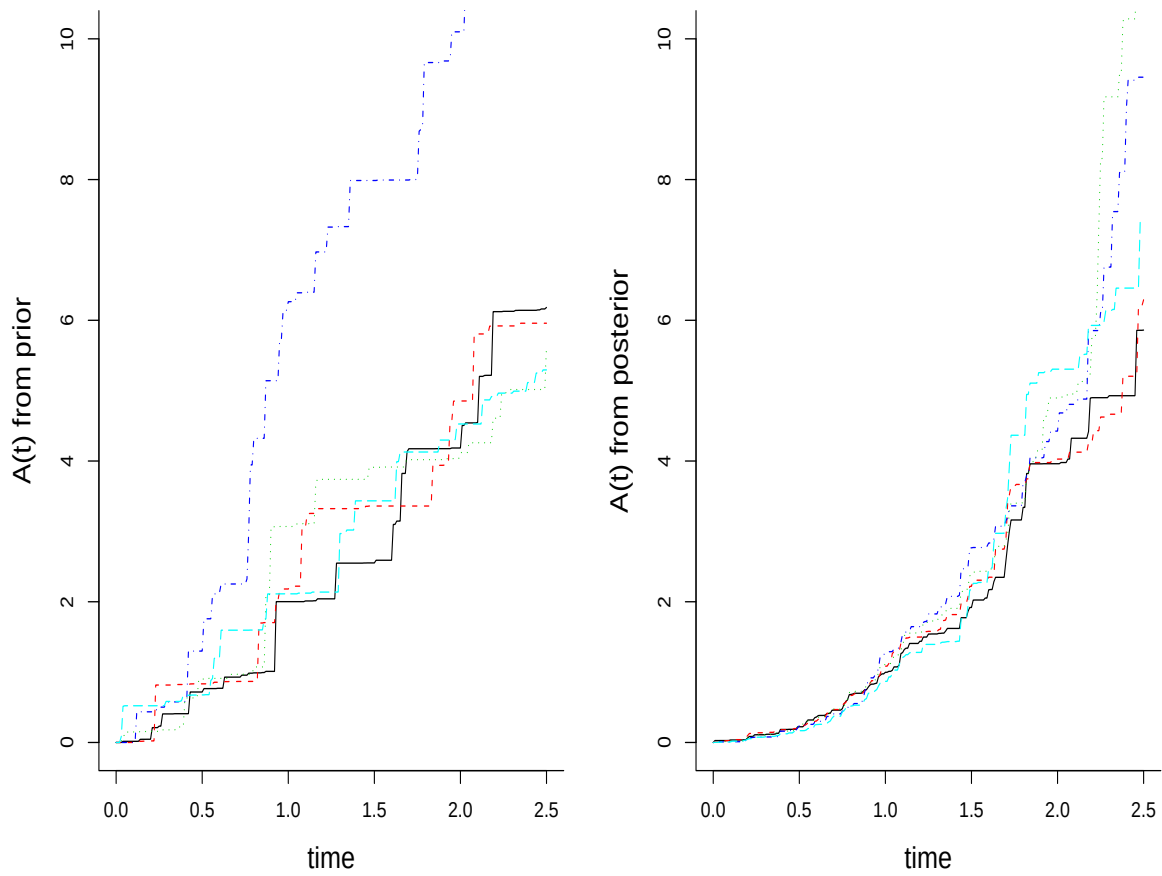
With c small we recover the **Nelson–Aalen** and **Kaplan–Meier** estimators.

Similar results for the bigger world of **event history analysis** (e.g. non-parametric Markov chains).

Canonical ‘Bayesian Cox regression’:

$$1 - dA_i(s) = \{1 - dA(s)\}^{\exp(x_i^t \beta)},$$

semiparametric prior for (A, β) . **It works.**



Five simulations of Beta processes A , from prior and from posterior, $n = 100$.

4. Bernshtein–von Mises theorems: First, $\sqrt{n}(\hat{A} - \hat{A}_{\text{freq}}) \rightarrow_{\text{pr}} 0$ and $\sqrt{n}(\hat{F} - \hat{F}_{\text{freq}}) \rightarrow_{\text{pr}} 0$. Even better:

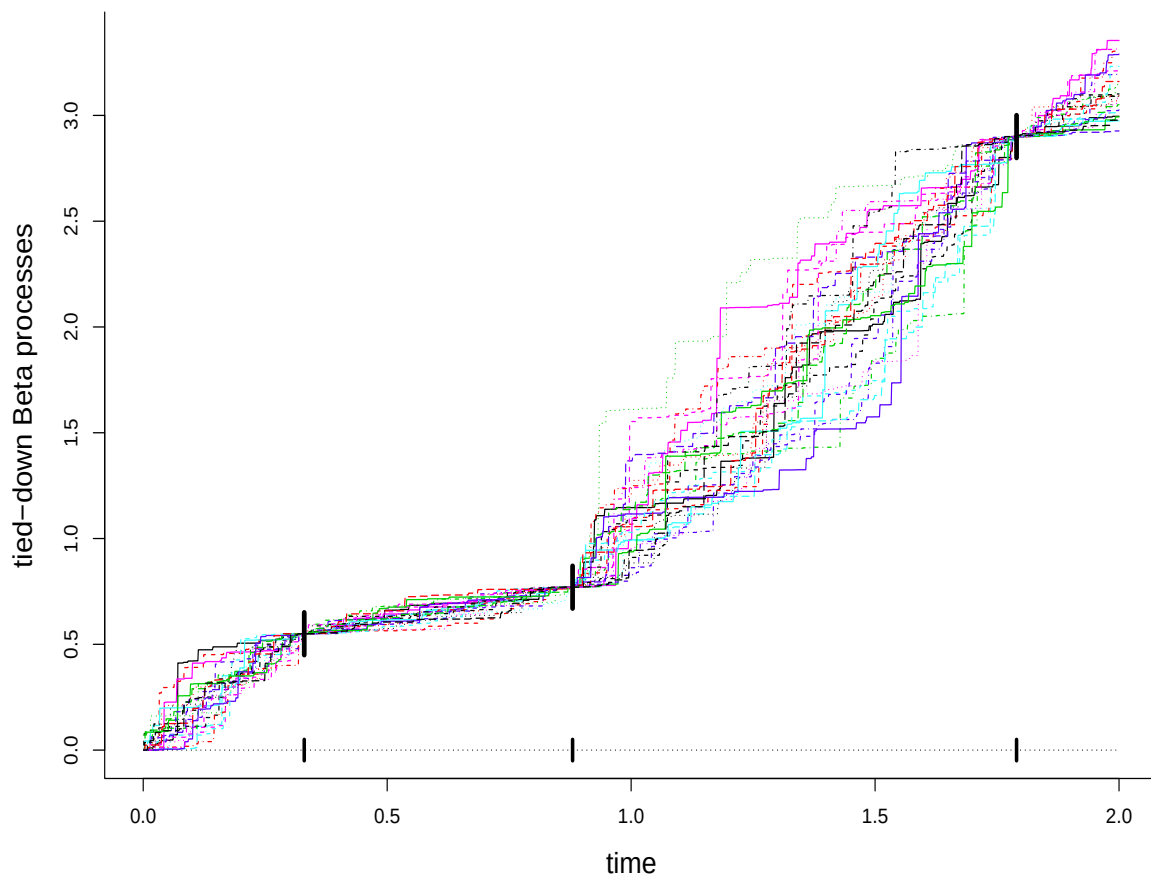
$$\sqrt{n}(A - \hat{A}) \mid \text{data} \rightarrow_d W\left(\int_0^\cdot (\alpha_{\text{true}}/y) \, ds\right),$$

the same Brownian motion limit as for $\sqrt{n}(\hat{A}_{\text{freq}} - A_{\text{true}})$. Such results also hold for ‘Bayesian Cox regression’ etc.

5. Growth curves: May use a Beta process as a prior for growth curves (a single such, or for an ensemble), and perhaps with further parameters inside c and/or A_0 . **Easy to update**, A given $A(t_j) = z_j$ at certain locations. (Gaussian processes: in trouble.)

With n individuals growing we could e.g. take $A_i \sim \text{Beta}(c, \lambda_i A_0)$, with inference for the λ_i based on observed A_i values at some locations.

With a distribution for λ_i we have **mixtures of Beta processes** (there are several classes of such; Kim 2001, NLH and Kim 2013).



25 simulations of Beta processes A , conditional on values 0.55, 0.77, 2.90 at positions 0.33, 0.88, 1.79.

6. Random curves: May take

$$Z(t) = \xi + \sigma(c+1)^{1/2} \frac{A(t) - A_0(t)}{A_0(t)^{1/2}}$$

as a prior for an unknown curve. It is stationary with covariance function

$$\text{cov}\{Z(s), Z(t)\} = \sigma^2 \exp\{-\tfrac{1}{2}|A_0(t) - A_0(s)|\}.$$

This is a Beta process based alternative to Ornstein–Uhlenbeck processes, **allowing skewness**. Again: easy to update.

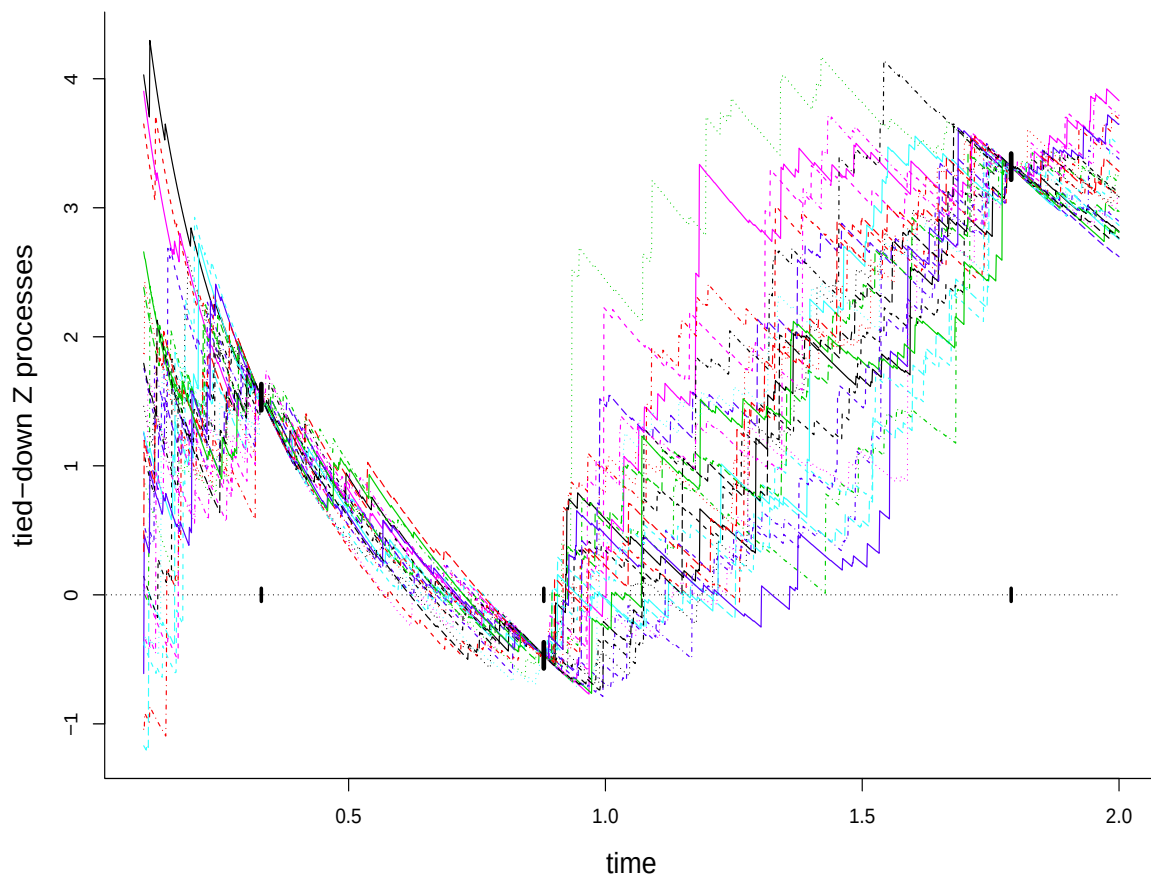
Application (ongoing, with a PhD student): Oil geologists need posterior distribution for a certain **fault line**, which has

$c(t) \leq U(t) \leq d(t)$ for known boundaries and for which one knows $U(t_j) = u_j$ at a few positions (for 10^7 dollars per data point).

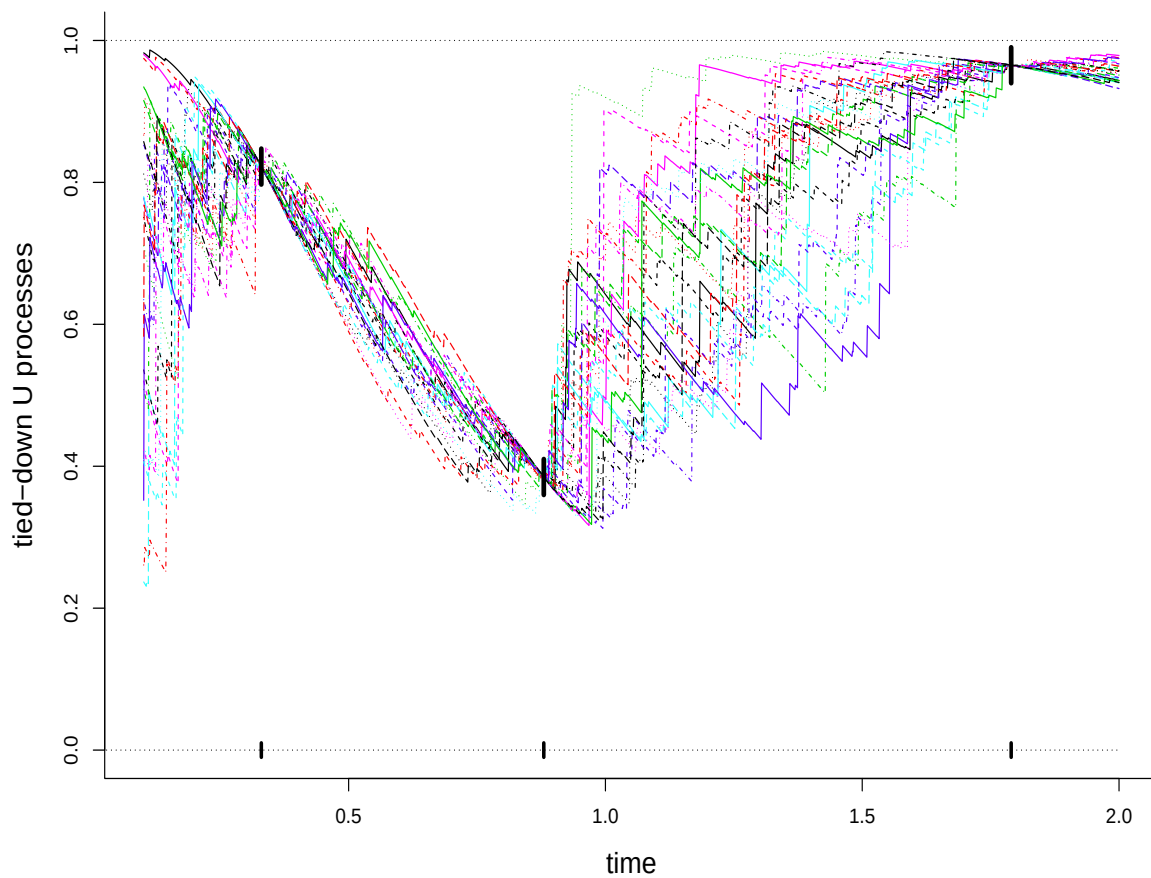
Our solution:

$$U(t) = c(t) + H(Z(t))\{d(t) - c(t)\},$$

with H the logistic transform. **Ongoing:** Lifting this to dimensions two and three, e.g. via $A(s, t) = A_1(s) + A_2(t) + A_3(|s - t|)$ etc.



25 simulations of processes Z , conditional on values 1.53, -0.47 , 3.32 at positions 0.33, 0.88, 1.79.



25 simulations of processes U , known boundaries 0 and 1, conditional on values 0.82, 0.38, 0.97 at positions 0.33, 0.88, 1.79.

7. Jumps: The Beta process has infinitely many jumps (at infinitely many locations) over each interval. May look for ‘Sethuraman–Tiwari representations’ (stick-breaking), e.g. of the form

$$A(t) = \sum_{j=1}^{\infty} J_j I\{\xi_j \leq t\} \quad \text{for } t \in [0, t_0]$$

with **ordered jumps** $J_1 > J_2 > \dots$. Simplest if $c(s) = c$ constant on $[0, t_0]$: (i) locations ξ_j are i.i.d. $a_0(\cdot)/A_0(t_0)$; (ii) positions and jumps are independent; (iii) the ordered jumps can be viewed as

$$\Lambda(J_j) = (V_1 + \dots + V_j)/A_0(t_0) \text{ in terms of } V_1, V_2, \dots \text{ i.i.d. } \sim \text{Expo}(1)$$

where

$$\Lambda(w) = \int_w^1 x^{-1} c(1-x)^{c-1} dx.$$

Particularly simple if $c = 1$:

$$J_j = e^{-(V_1 + \dots + V_j)/A_0(t_0)} = B_1 \dots B_j, \quad \text{with } B_j \sim \text{Beta}(A_0(t_0), 1).$$

Easy recipe for simulation also for $c \neq 1$.

8. Derived hazard models: Rather than simply ‘inventing’ hazard rate models (like Cox regression) one should (also) **derive** good models from **biologically plausible** background models.

Example 1: A Beta process governs your life. You die the first time its jump is bigger than threshold z . Your survival distribution is then

$$\Pr\{T > t\} = \Pr\{J_1 \leq z\} = \exp\{-A_0(t)\Lambda^{-1}(z)\},$$

so **your hazard rate** is $h(t) = a_0(t)\Lambda^{-1}(z)$. This rhymes with proportional hazards – different patients carry the same type of Beta processes, but tolerate different levels z_i ; Cox regression says $\Lambda^{-1}(z_i) = \exp(x_i^t\beta)$. **Other biologically plausible models** also emerge – different patients have different $(c_i, A_{0,i})$; mixing over frailty distributions; you live until the k th shock exceeds threshold; etc.

Example 2: You are exposed to two types of risks, modelled with two independent Beta processes, and die when jump type 1 is $\geq z_1$ or jump type 2 is $\geq z_2$. Your hazard rate is then of type

$$h(t) = a_{0,1}(t)x_1 + a_{0,2}(t)x_2,$$

with $x_j = \Lambda_j^{-1}(z_j)$ for $j = 1, 2$. This is the **Aalen model**.

9. Bayesian nonparametrics for time series: A Gaussian stationary time series is governed by the **covariance function**

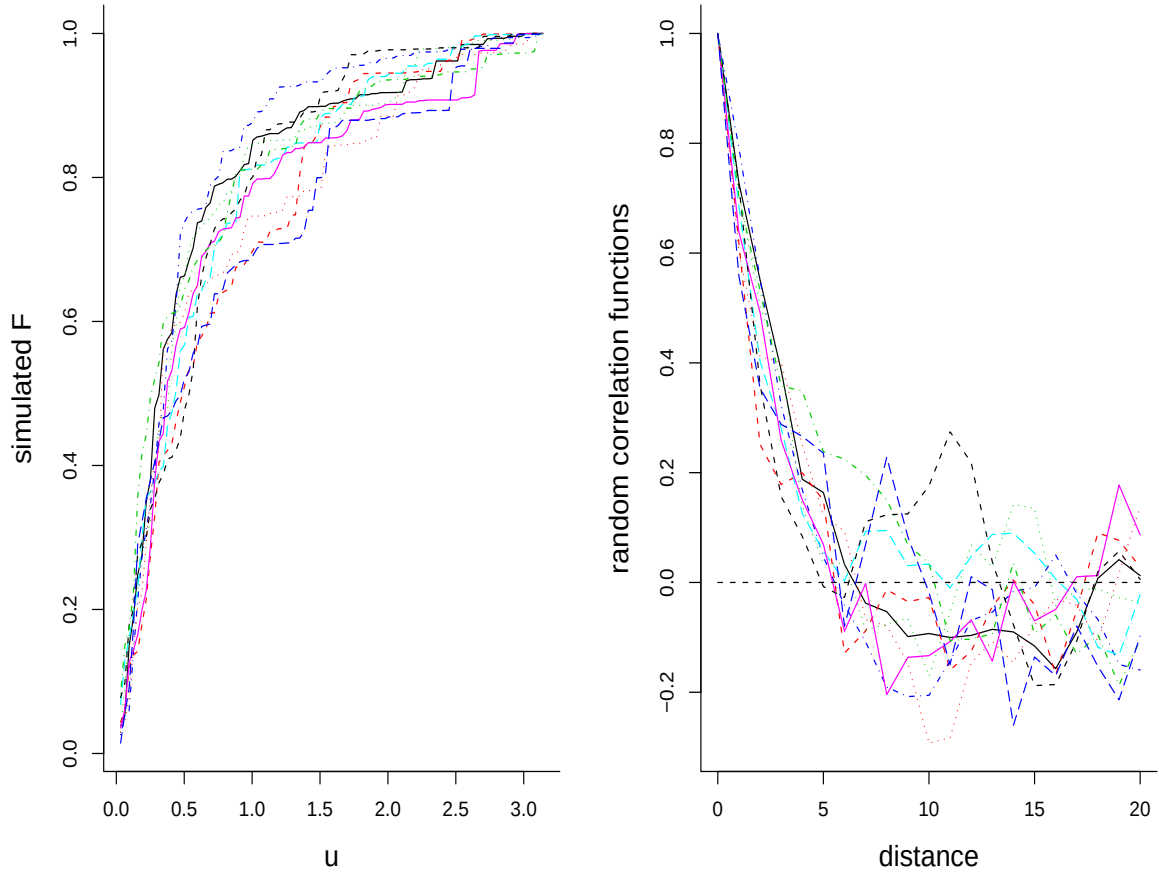
$$C(h) = \text{cov}(Y_i, Y_{i+h}) = 2 \int_0^\pi \cos(h\omega) \, dA(\omega),$$

where A is the **spectral measure** on $[0, \pi]$. Letting A be a $\text{Beta}(c, A_0)$ yields a nonparametric time series model ‘around’ $C_0(h)$ associated with A_0 .

We have **random correlation functions** via

$$\text{corr}(h) = 2 \int_0^\pi \cos(h\omega) \frac{dA(\omega)}{A(\pi)}.$$

[Hermansen and Hjort (2013), also with Bernshtein–von Mises theorems.]



Left: simulated normalised Beta processes, $F = A/A(\pi)$ on $[0, \pi]$; right: the random correlation functions $C(h) = 2 \int_0^\pi \cos(hu) dF(u)$.

10. Random shapes: Consider

$$(X(t), Y(t)) = R(t)(\cos(2\pi t), \sin(2\pi t)) \quad \text{for } 0 \leq t \leq 1.$$

This is a random shape model as long as the stochastic radius process $R(t)$ is positive, continuous, and has $R(0) = R(1)$.

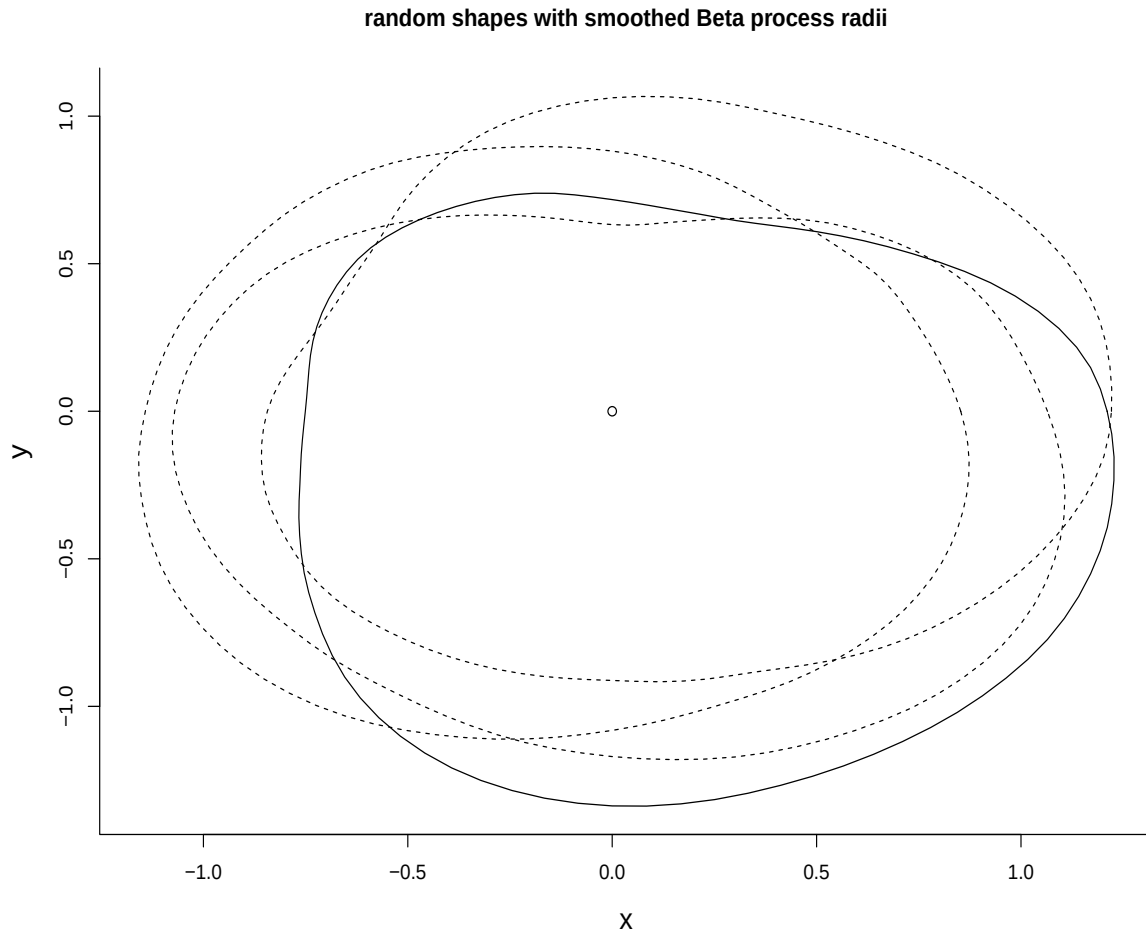
A class of such models:

$$R(t) = \int K_b(t - u) \frac{dA_c(u)}{A(1)},$$

with $A_c \sim \text{Beta}(c, A_0)$, e.g. with $A_0(u) = u$ to generate **featureless shapes** (no ‘landmarks’).

Here integration is ‘modulo the circle’, and $K_b(u) = b^{-1}K(b^{-1}u)$, with K the Yepanechnikov kernel on $[-\frac{1}{2}, \frac{1}{2}]$. **Observed shapes** $\implies$ inference for (b, c) .

There are papers similar in spirit to the above but which use Gaussian radius processes (which is troublesome).



Simulated random shapes via $R(t) = \int K_b(t - u) dA_c(u)$ for a Beta process (c, A_0) . This is the (b, c) shape model, here with $(b, c) = (0.35, 100)$.

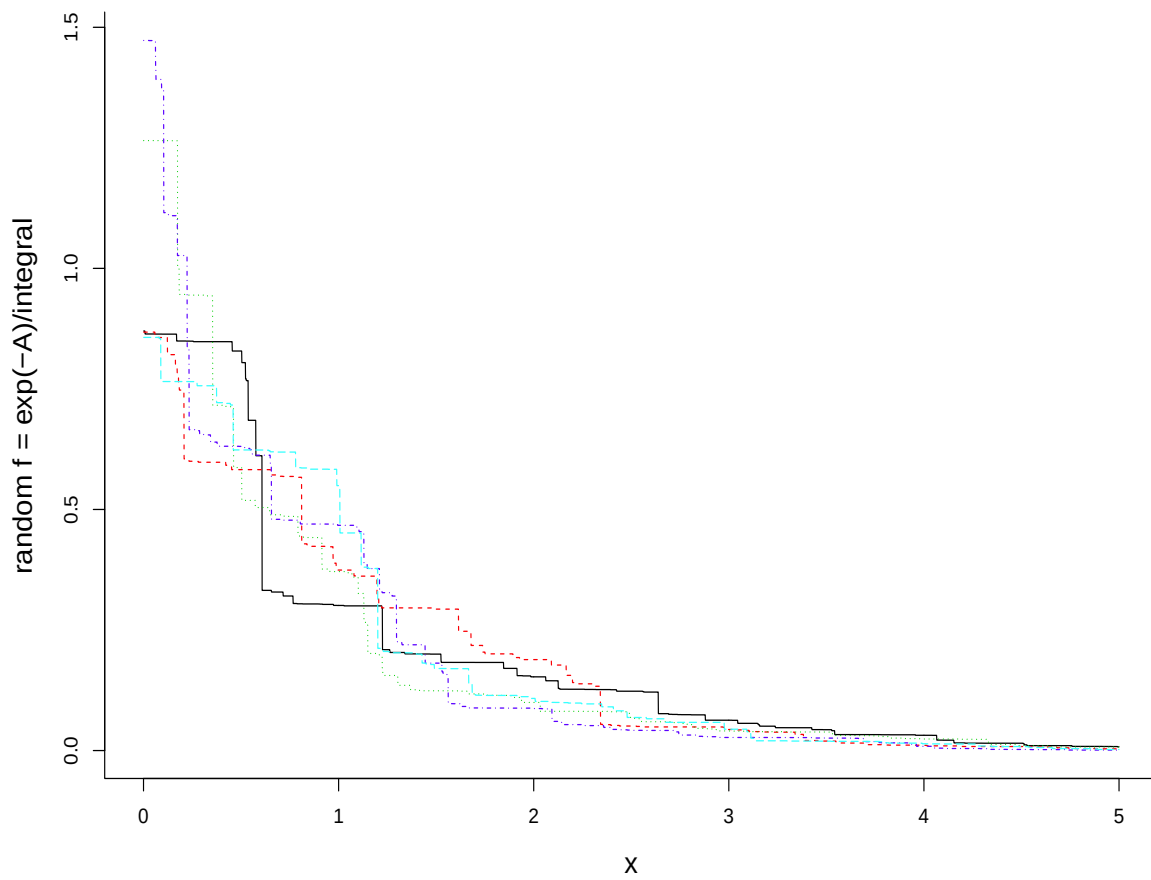
11. Bayesian Grenander: For data $x_1, \dots, x_n$ from a nonincreasing density f on $[0, \infty)$ the **Grenander estimator** (1956) maximises the likelihood $\prod_{i=1}^n f(x_i)$ (with no smoothing parameter or similar). It has $n^{1/3}$ convergence to f_{true} .

A Bayesian attempt:

$$f(x) = \frac{\exp\{-A(x)\}}{\int_0^\infty \exp\{-A(x')\} dx'},$$

with A a Beta process.

NLH and Walker (‘initiated project’, 2013): (i) it can be done (clever trick for the normalising constant); (ii) we prove consistency; (iii) we have not yet been able to prove convergence of $n^{1/3}(f - \hat{f}_G) \mid \text{data}$.



Simulated decreasing densities $f = \exp(-A) / \int_0^\infty \exp(-A) dx$, with $A \sim \text{Beta}(c, A_0)$.

12. Indian Buffet Processes: Model for exchangeable sequence of 0–1 vectors,

$$p(z_1, \dots, z_n) = \int \left\{ \prod_{i=1}^n P(z_i \mid A) \right\} d\mathcal{P}(A) \quad \text{for all } n.$$

The **de Finetti measure** $\mathcal{P}$ leading to the IBF is a Beta process – see other talks and posters at this conference.

Extended Beta processes $\implies$ extended IBFs.

B: Tilting the Beta processes

Cumsum process with tilted, independent, small Betas:

$$A_m(t) = \sum_{i/m \leq t} p_i^\gamma,$$

where

$$\mathbb{E} p_i = a_0 \left(\frac{i}{m} \right) \frac{1}{m} \quad \text{and} \quad \text{Var } p_i \doteq \frac{1}{c+1} a_0 \left(\frac{i}{m} \right) \frac{1}{m}.$$

There is convergence $A_m \rightarrow_d A$ to a well-defined $\text{TiltBeta}(c, A_0, \gamma)$ (in the space of cumulative hazards, all jumps ≤ 1).

Proof: (i) **Laplace transform convergence**

$$\mathbb{E} \exp\{-\theta A_m(t)\} = \prod_{i/m \leq t} \mathbb{E} \exp(-\theta p_i^\gamma) \rightarrow \exp\{-D(\theta, t)\},$$

$$D(\theta, t) = \int_0^1 (1 - e^{-\theta x}) dL_t(x), \quad dL_t(x) = \gamma^{-1} c x^{-1} (1 - x^{1/\gamma})^{c-1} A_0(t).$$

(ii) **Tightness**, via bounds on $\mathbb{E} \{A_m(t) - A_m(s)\} \{A_m(u) - A_m(t)\}$ etc.

Moments (also used in the proof, to identify limiting Laplace transform and the Lévy measure):

$$\mathbb{E} p_i^{j\gamma} \doteq \frac{\Gamma(c+1)\Gamma(j\gamma)}{\Gamma(c+j\gamma)} a_0 \left(\frac{i}{m} \right) \frac{1}{m}.$$

With $k_j(c, \gamma) = \Gamma(c+1)\Gamma(j\gamma)/\Gamma(c+j\gamma)$, we have

$$\begin{aligned} \mathbb{E} A(t) &= k_1(c, \gamma) A_0(t), \\ \text{Var } A(t) &= k_2(c, \gamma) A_0(t), \\ \text{skew}\{A(t)\} &= \frac{k_3(c, \gamma)}{k_2(c, \gamma)^{3/2}} \frac{1}{A_0(t)^{1/2}}. \end{aligned}$$

The extra γ affords **greater flexibility** regarding **variance-to-mean** ratio, **skewness**, kurtosis, &cetera.

Note: For the special case $c = 1$, $\text{TiltBeta}(c, A_0, \gamma)$ is the same as $\text{Beta}(c, A_0/\gamma)$.

Most of the results and methods and most of the properties listed earlier for the classic Beta process **may be lifted** to the $\text{TiltBeta}(c, A_0, \gamma)$.

Consistency and Bernshteĭn–von Mises for tilted Beta process: I have proven this, but it's rather harder than for classic Beta (i.e. for $\gamma = 1$).

Results of NLH (1990) give a complete description of $A \mid \text{data}$. I use this to prove (i) that the mass between observed transitions is small,

$$(A_{\text{between}} \mid \text{data}) = O_{\text{pr}}(1/n),$$

and then (ii) that the posterior mass $\Delta A(t_i)$ at an observed data point is close to $\text{Beta}(1, Y(t_i))$.

This yields

$$\sqrt{n}(A - \hat{A}) \mid \text{data} \rightarrow_d W\left(\int_0^\cdot (\alpha_{\text{true}}/y) \, ds\right).$$

C: Concluding comments

1. Can work with Beta-Tilted Beta Process:

$$A_m(t) = \sum_{i/m \leq t} H(p_i, a, b),$$

with $H(x, a, b)$ the Beta cdf, e.g. increments $1 - (1 - p_i)^\gamma$. I can prove existence of limit A , along with various properties, but [so far] I don't have a formula for the Lévy measure $dL_t(x)$.

2. Various 'biologically plausible' hazard rate models emerge from background Beta processes. Gamma-tilting leads to **extended hazard rate models** (Cox regression with an extra γ parameter, etc.).

3. The $\text{TiltBeta}(c, A_0, \gamma)$ should lead to a **tilted Indian Buffet process** – **either** by tilting in its original construction **or** via one of various other characterisations.

4. The extra γ affords greater flexibility for modelling random growth curves as well as transformed stationary curves. Such models are **underused** (since the GPs are so good?).

5. For the ‘Bayesian Grenander’, with $f = \exp(-A) / \int_0^\infty \exp(-A) dx$, let

$$\hat{f}_B(x) = \mathbb{E}\{f(x) \mid \text{data}\}.$$

NLH and Walker (project, 2013) are hoping to prove convergence of $n^{1/3}(f - \hat{f}_B) \mid \text{data}$, but we do **not** expect a Bernshtein–von Mises result: limit is likely to be different from $n^{1/3}(\hat{f}_G - f_{\text{true}})$.

6. ‘All extensions are interesting but some extensions are more interesting than others.’ I have not yet sorted out which ‘tilting applications’ are more crucial than others.