Processes With Steadily Rarer But Steadily Bigger Shocks

Nils Lid Hjort / Stability and Change 2022-2023, CAS

Bo Lindqvist Emeritus Celebrations, 15/ix/2022

Main themes & two-page summary

(a) I work with a class of stochastic processes, with steadily rarer but steadily bigger shocks:

$$Z_n(t) = \sum_{i \le [nt]} J_i c(i/n)^\alpha U_i,$$

where the U_i are i.i.d., the J_i are Bernoullis with $p_i = \min(1, 1/(ci))$. There is a clear limit process,

$$Z_n(t) \to_d Z(t),$$

with independent increments. There are nice special cases; favourite case has the U_i i.i.d. exponential.

(b) The $Z(t)$ can be useful by itself – but here I focus on time-to-reach-threshold:

$$T = \min\{t > 0\colon Z(t) \ge k\}.$$

I study density $f(t, c, \alpha, k)$, survival function $S(t, c, \alpha, k)$, etc.; these have power-law tails: $F(t) \doteq 1 - d/t^{1/c}$ for growing t.

(c) With survival (or other) data $T_1, \ldots, T_n$, how can we estimate parameters c, α, k? Various non-trivial technicalities; I develop estimating techniques different from maximum likelihood.

(d) The model gives a good fit to CoW data, battle deaths in major wars 1823-to-present. Better than the three-parameter model of Cunen, Hjort, Nygård (JPR, 2020)? Don't know (yet).

(e) How useful are such analyses? Don't know (yet) – but parameters are interpretable; we may test for constancy over time vs. change points; covariates may be introduced in the model; etc.

(f) Monitoring processes for assessing goodness-of-fit (quite a bit of work).

(g) I've only tried with the CoW data, so far – would be of interest to try survival data where time to event might be l-o-o-o-n-g, and to violence data sets, to see how the power-laws can be assessed (and 'explained').

A: Let's begin: the Z_n processes

With $Z_n(t) = \sum_{i \le [nt]} J_i c(i/n)^\alpha U_i$, let $\mathrm{E}\, U_i = \xi$, $\mathrm{Var}\, U_i = \sigma^2$.

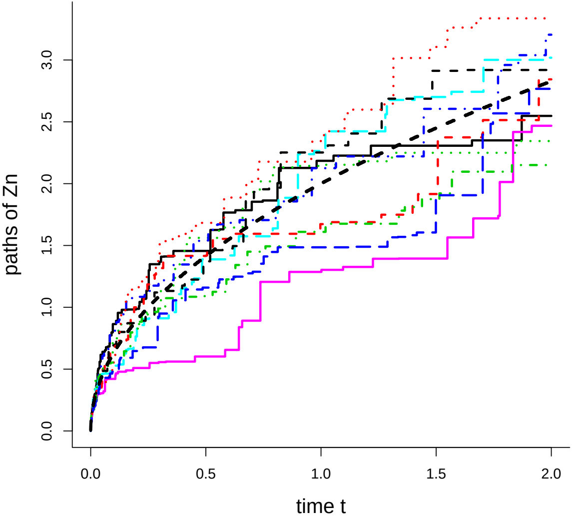
Note: infinitely many $J_i = 1$, with $p_i = \min(1, 1/(ci))$, from Borel–Cantelli.

Also:

$$\mathrm{E}\, Z_n(t) = \sum_{i \le [nt]} p_i c(i/n)^\alpha \xi \doteq \frac{[nt]^\alpha}{n^\alpha}(1/\alpha)\xi \to (1/\alpha)\xi t^\alpha,$$

$$\begin{aligned}\mathrm{Var}\, Z_n(t) &= \sum_{i \le [nt]} c^2 (i/n)^{2\alpha}\{p_i(\xi^2+\sigma^2) - (p_i\xi)^2\} \\ &\to c/(2\alpha)(\xi^2+\sigma^2)t^{2\alpha}.\end{aligned}$$

Note that $\mathrm{Var}\, Z(t)/\{\mathrm{E}\, Z(t)\}^2 \to \mathrm{constant}$.



Ten simuated paths from the Z_n model, with $c = 0.20$, $\alpha = 0.50$, unit exponentials for the U_i, and $n = 10^5$. The mean curve $(1/\alpha)t^\alpha$ is the dashed curve in the middle.

B: There are clear limits, $Z_n(t) \to_d Z(t)$

Recall $Z_n(t) = \sum_{i \le [nt]} J_i c(i/n)^\alpha U_i$. Behaviour depends on the distribution of the core shocks, the i.i.d. U_i. Let

$$h(s) = \mathrm{E}\,\exp(-sU_i), \quad \text{the Laplace transform.}$$

Theorem: We have $Z_n(t) \to_d Z(t)$, with independent increments, and

$$\mathrm{E}\,\exp\{-\theta Z(t)\} = \exp\Big\{-\frac{1}{c\alpha}\int_0^{c\theta t^\alpha} \frac{1-h(s)}{s}\,\mathrm{d}s\Big\}.$$

Favourite Special Case (so far): the U_i are i.i.d. unit expo. Then

$$h(s) = \frac{1}{1+s} \quad \text{implying} \quad \frac{1-h(s)}{s} = \frac{1}{1+s}.$$

This leads to

$$\mathrm{E}\,\exp\{-\theta Z(t)\} = \frac{1}{(1+c\theta t^\alpha)^{1/(c\alpha)}},$$

which means

$$Z_n(t) \to_d Z(t) \sim \mathrm{Gamma}(1/(c\alpha), 1/(ct^\alpha)).$$

C: A new type of Gamma process

Gamma processes are seen in lots o' models and applications. They are nearly always of the type

$$Z(t) \sim \mathrm{Gamma}(H_1(t), H_2(t)),$$

for increasing $H_1(t)$, and typically constant $H_2(t)$.

The present type is rather different:

$$Z(t) \sim \mathrm{Gamma}(1/(c\alpha), 1/(ct^{\alpha})).$$

Footnote: Somewhat peculiarly:

$Z(t_2) = Z(t_1) + E$, with $Z(t_1)$ and $Z(t_2)$ Gamma, but E not Gamma.

D: Time to reach threshold: power laws

Consider $T = \min\{t > 0\colon Z(t) \geq k\}$. Then

$$\begin{aligned} S(t) = \mathrm{Pr}(T \geq t) &= \mathrm{Pr}\{\mathrm{Gamma}(1/(c\alpha), 1/(ct^\alpha)) < k\} \\ &= \mathrm{Pr}\Big\{\frac{\mathrm{Gamma}(1/(c\alpha), 1/(ct^\alpha))}{ct^\alpha} < \frac{k}{ct^\alpha}\Big\} \\ &= G_0\Big(\frac{k}{ct^\alpha}, \frac{1}{c\alpha}\Big), \end{aligned}$$

with $G_0(t, a)$ the c.d.f. for a $\mathrm{Gamma}(a, 1)$.

Density:

$$f(t) = \frac{\alpha}{\Gamma(1/(c\alpha))}\Big(\frac{k}{c}\Big)^{1/(c\alpha)} \exp\Big(-\frac{k}{ct^\alpha}\Big)\frac{1}{t^{1/c+1}}.$$

It peaks at $t_0 = \{k\alpha/(c+1)\}^{1/\alpha}$, which is typically a low value, then goes slowly to zero in power-law fashion.

Theorem: Distribution of T, given $T \geq t_0$, becomes uniformly close to power-law, proportional to $1/t^{1/c+1}$, as t_0 grows:

$$\sup_{t \geq t_0} \left| \frac{f(t \mid T \geq t_0)}{(1/c) t_0^{1/c} / t^{1/c+1}} - 1 \right| \to 0.$$

So for large data values, only tail index $\gamma = 1/c$ matters:

$$\Pr(T \geq t \mid T \geq t_0) \doteq (t_0/t)^{1/c} \quad \text{for } t \geq t_0.$$

With data above threshold t_0,

$$y_i = \log(t_i/t_0) \quad \text{i.i.d. Expo}(1/c).$$

So may use maximum likelihood for these:

$$c^* = (1/m) \sum_{i=1}^{m} \log(t_i/t_0).$$

E: Setting the threshold

Start from full dataset $t_1, \ldots, t_n$. For each candidate threshold t_0, throw the $t_i \geq t_0$ to a goodness-of-fit machine to see if $F(t) = 1 - (t_0/t)^\gamma$ is ok for $t \geq t_0$. With a suitable such test statistic $W(t_0)$, compute

$$p(t_0) = \mathrm{Pr}\{W^*(t_0) \geq W_{\mathrm{obs}}(t_0)\},$$

where $W^*(t_0)$ is from the null distribution. Result: a p-value plot.

I've transformed to goodness-of-fit to $\mathrm{Expo}(\gamma)$ for $y_i = \log(t_i/t_0)$, and used

$$W_m = \sqrt{m} \int |F_m(y) - F(y, \widehat{\gamma})| \, \mathrm{d}F_m(y) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} |i/m - F(y_{(i)}, \widehat{\gamma})|,$$

with F_m the empirical c.d.f. for $y_{(1)} < \cdots < y_{(m)}$. Also, $F(y, \widehat{\gamma}) = 1 - \exp(-\widehat{\gamma} y)$, with $\widehat{\gamma} = 1/\bar{y}$, for t_i above threshold. Accept as threshold (first) t_0 where power-law is ok.

F: Estimating the three parameters (using all data)

I now wish to use all data $t_1, \ldots, t_n$, e.g. the CoW battle deaths, not merely those above (an estimated) threshold.

(i) Can use ML (a bit troublesome numerically, but it works), using

$$f(t) = \frac{\alpha}{\Gamma(1/(c\alpha))}\Big(\frac{k}{c}\Big)^{1/(c\alpha)} \exp\Big(-\frac{k}{ct^{\alpha}}\Big)\frac{1}{t^{1/c+1}}.$$

(ii) ML gives each datum equal importance. Here might wish to give more emphasis on higher values. The quantile function is

$$F^{-1}(q, c, \alpha, k) = \Big\{\frac{k}{cG_0^{-1}(1-q_j, 1/(c\alpha))}\Big\}^{1/\alpha}.$$

For a set of quantiles, can minimise

$$Q_n(c, \alpha, k) = \sum_{j=1}^{r} w(q_j)\{F_n^{-1}(q_j) - F^{-1}(q_j, c, \alpha, k)\}^2,$$

and this delivers $\widehat{c}, \widehat{\alpha}, \widehat{k}$.

G: Assessing goodness of fit

Suppose a parametric model with c.d.f. $F(t, \theta)$ is correct at θ_0. Then the ordered $F_{(i)} = F(t_{(i)}, \theta_0)$ are an ordered sample from the uniform, with expected values $1/(n+1), \ldots, n/(n+1)$. The Diagonal Diagnostic Plot is

$$(i/(n+1), \widehat{F}_{(i)}) \quad \text{for } i = 1, \ldots, n,$$

with the ordered version of $\widehat{F}_i = F(t_i, \widehat{c}, \widehat{\alpha}, \widehat{k})$. If model is good, this should produce a plot close to the diagonal.

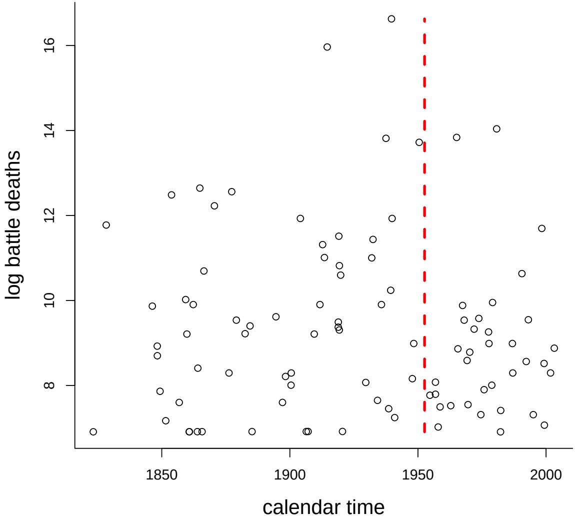
There are various other goodness-of-fit monitoring processes to pursue, also based on the parameter estimation methods used (see my paper-to-be). In particular, comparing Nelson–Aalen to estimated parametric model, versions of

$$\sqrt{n}\{\widehat{A}(t) - A(t, \widehat{c}, \widehat{\alpha}, \widehat{k})\},$$

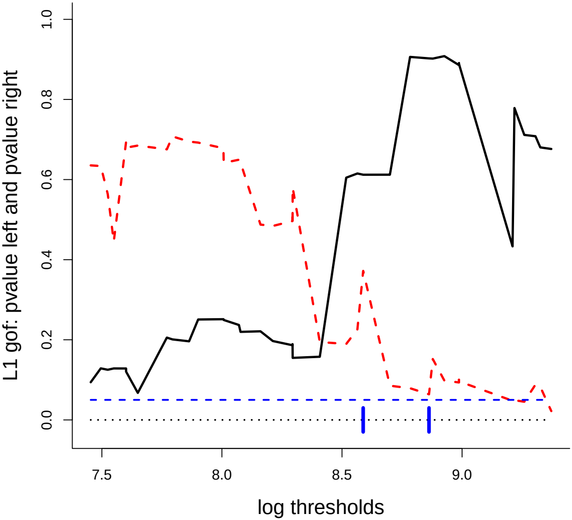
suitable for survival type data.

H: Battle deaths from the CoW database

The log of battle death counts for all $n = 95$ major interstate wars, from 1823 to the present; Korea 1950 tentative change point.

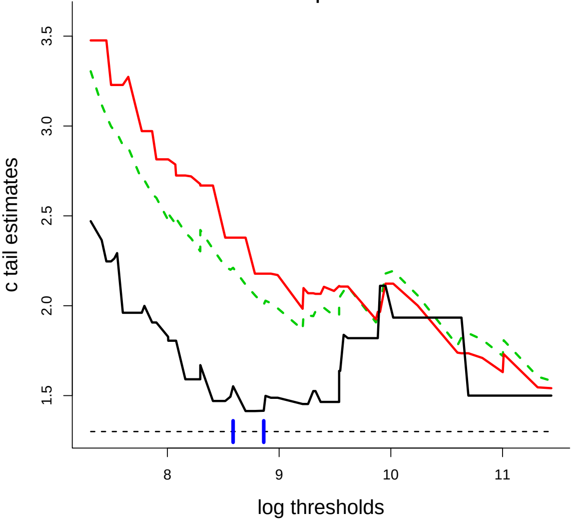


p-value plots for testing the power-law tail models $F(t) = 1 - (t_0/t)^\gamma$, for data above threshold t_0. Left of Korea: red; right of Korea: black, Blue marks on the log-thresholds scale are threshold values 5368 (sensible here), and 7061 (Clauset, 2018).

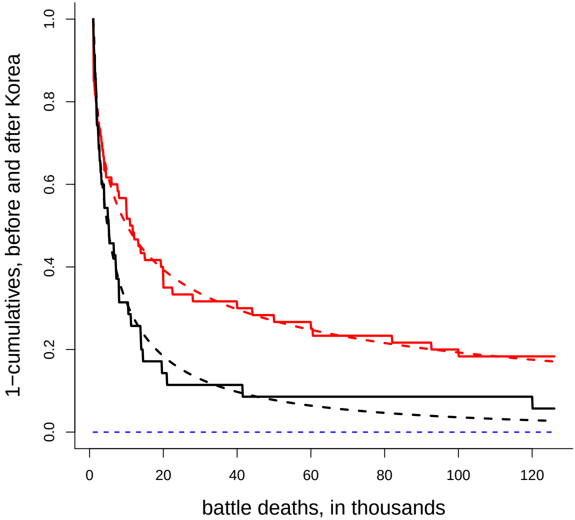


Plots of estimated tail power index $c^*(t_0)$, computed for data above threshold t_0. All wars (green, dashed line), for wars before 1950 (red), and after 1950 (black). Blue marks are for thresholds 5368 and 7061, where estimates (c_L^*, c_R^*) are (2.379, 1.552) and (2.178, 1.499).

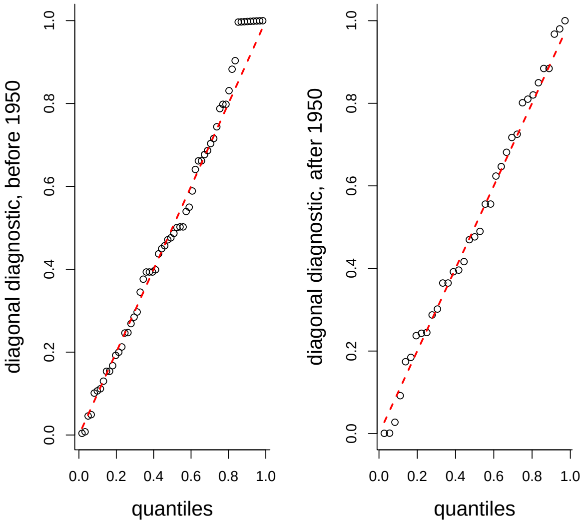
That c is smaller now than in the past is Good News.



Empirical and fitted 1 minus cumulatives, using the three-parameter $F(t, c, \alpha, k)$ model, for battle deaths before (red) and after (black) 1950, counted in thousands.



Diagonal Diagnostic Plots, for the three-parameter $F(t, c, \alpha, k)$ model, for data left and right of Korea 1950.



I: Discussion

- My favourite model, so far, from time-to-threshold, says

$$F(t) = G_0(k/(c(t-1000)^\alpha), 1/(c\alpha)) \quad \text{for } t \geq 1000$$

for the CoW battle deaths data. Can make covariates part of the game, e.g.

$$k_i = k \exp(-\beta\, \mathrm{dem}_i),$$

with dem_i average democracy index for the two warring parties just prior to war.

- For the CoW series, Céline and Nils (JPR, 2020) invented the model

$$F(t) = \Big[\frac{\{(t-1000)/\mu\}^\theta}{1+\{(t-1000)/\mu\}^\theta}\Big]^\alpha \quad \text{for } t \geq 1000.$$

It has tails coming close to power-law; it works well; we used it to spot Korea 1950 as changepoint; we could incorporate covariates. However, it's a bit ad hoc, whereas this talk's model is derived for a plausible interpretable background model. – More comparisons needed.

♯ Changes, discontinuities, trends take many forms: Based on similar ideas in Cunen, Hjort, Nygård (JPR 2020), wish to look for things like

$$k_i = k \exp(-\beta\,\mathrm{dem}_i),$$

with perhaps $\beta \approx 0$ in the past, but $\beta > 0$ now.

♯ Might attempt to apply these new Gamma processes to ‘time to certain events is sometimes very-very long’ phenomena in medical statistics or biology.

♯ Might adjust models to take on board a cure fraction, individuals never experiencing the event.

A few references

- C. Cunen, N.L. Hjort, H.M. Nygård (2020). Statistical Sightings of Better Angels. *Journal of Peace Research*.
- C. Cunen, N.L. Hjort (2022). Combining information across diverse sources: the II-FF-CC paradigm. *Scandinavian Journal of Statistics*.
- K. von Clausewitz (MCMXXXV, original from 1832): *Vom Kriege*. Im Insel-Verlag zu Leipzig; preface by A. Hitler;
- N.P. Gleditsch (2020). Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences (edited volume).
- N.L. Hjort (2018). Towards a More Peaceful World [insert '!' or '?' here]. FocuStat Blog Post.
- N.L. Hjort (2022). Processes swith bigger shocks at rarer rates.
- S. Pinker (2011). *The Better Angels of Our Nature.*
- S. Pinker (2018). *Enlightenment Now.*
- L. Tolstoy (1869). Война и мир.