Statistical Modelling for Earthquakes Accounting for Measurement Error

Kushagra Agarwal[1] Wanchen Yue[2]

[1]University of Warwick [2]STOR-i, Lancaster University

excellence with impact

1. Motivation & Background

- Groningen gas field was founded in 1959 and is one of the largest gas fields in the world.
- Extraction of gas from the field began in 1963 by a joint venture by Shell and Exxon.
- The extraction of gas has led to a number of **human-induced earthquakes** (Hi-Quakes) since 1991 in the province of Groningen, and this remains a problem to this day, with the most recent earthquake occurring on 13th August, 2024.
- The induced seismic activity has caused damage to buildings and infrastructure in the region, leading to financial losses and safety concerns.

Figure 1: Reinforcement of some properties in Groningen due to earthquake risk, while others are deemed too dangerous to inhabit, illustrating the impact of induced seismic activity on local infrastructure.

2. Hawkes Processes

A Hawkes process is a point process that is widely used in earthquake modelling and financial analysis. The defining characteristic of this process is that it **self-excites**, meaning that each arrival increases the rate of future arrivals for some period of time.

The intensity function is history dependent on $\{t_1, t_2, \ldots, t_k\}$ and is given by

$$\lambda(t \mid \mathcal{H}_t) = \mu + \sum_{i:t_i<t} \gamma(t - t_i)$$

where μ is the background intensity and γ is the excitation function. A common choice for the excitation function is the exponential decay, that is, $\gamma(t) = \alpha e^{-\beta t}$ which is parameterised by constants α, $\beta > 0$ which gives us

$$\lambda(t \mid \mathcal{H}_t) =: \lambda^*(t) = \mu + \sum_{i:t_i<t} \alpha e^{-\beta(t-t_i)}$$

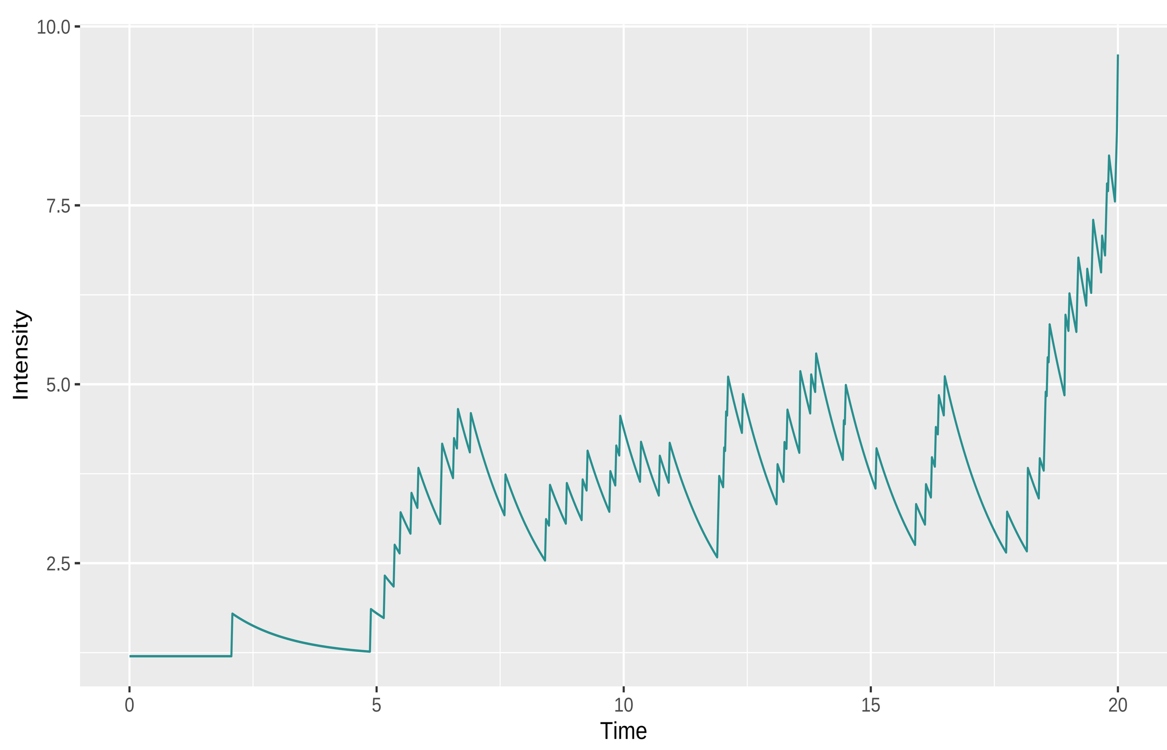

Figure 2: Conditional intensity function of a Hawkes process with parameters $\mu = 1.2$, $\alpha = 0.6$, and $\beta = 0.8$.

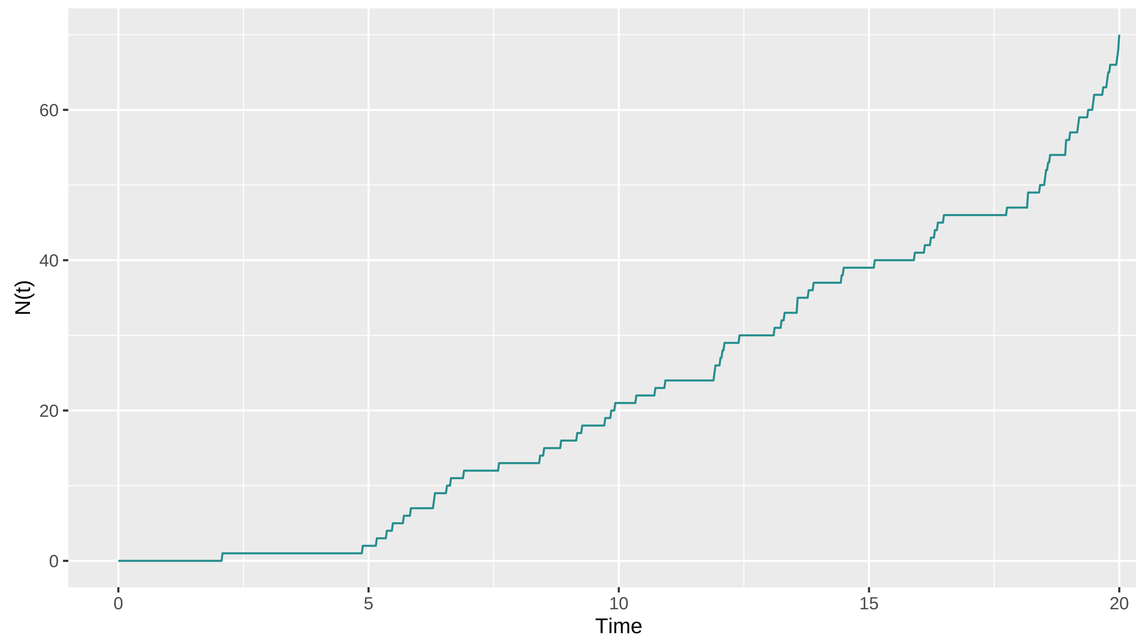

Figure 3: And the corresponding counting process.

3. Extreme Value Theory

Modelling of extremes can be done in the following two ways:

1. Block Maxima
   - If $X_1, \ldots, X_n$ are i.i.d random variables, then there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that $\frac{M_n - b_n}{a_n} \dot\sim \text{GEV}(\mu, \sigma, \xi)$ as $n \to \infty$, where GEV is the **Generalized Extreme Value distribution** with parameters $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$, and $M_n := \max\{X_1, \ldots, X_n\}$.
   - The CDF of the GEV distribution is given by:

$$G(x) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, & \xi = 0 \end{cases}$$

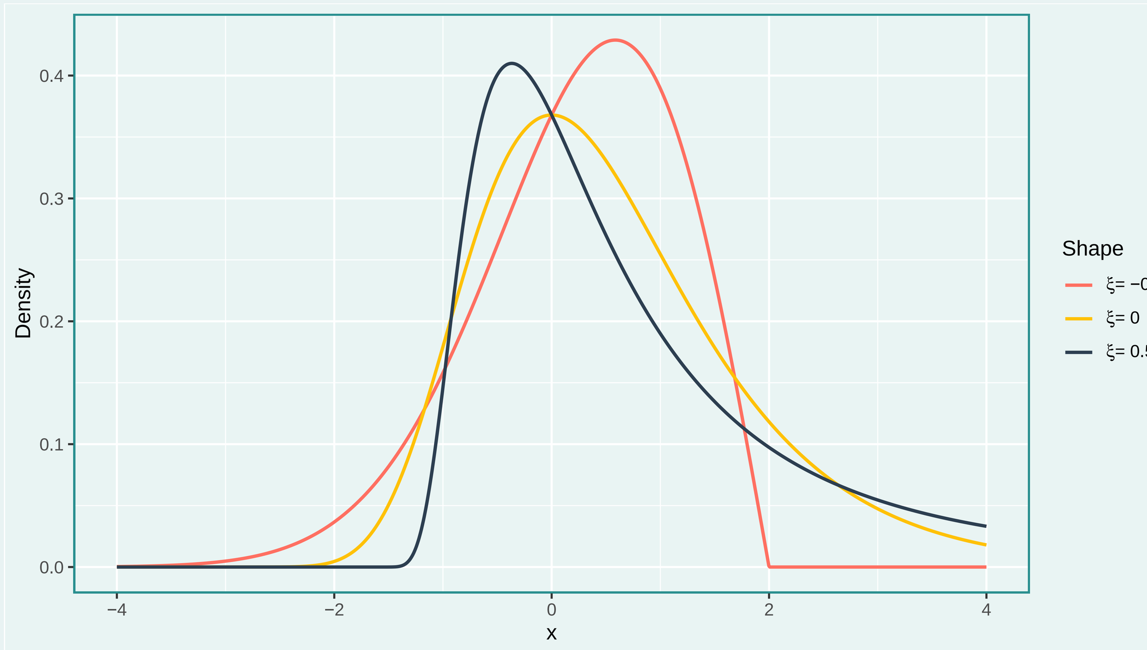

Figure 4: Probability density function for GEV distribution with $\mu = 0$, $\sigma = 1$ and different values for ξ.

2. Threshold Excess Modelling
   - However, using only the maximum is a wasteful of data. Using the same setting as above, for large enough u and conditional on $X_i > u$, $X_i - u \dot\sim \text{GPD}(u, \sigma, \xi)$ as $n \to \infty$, where GPD is the **Generalised Pareto distribution**.
   - The CDF of the GPD distribution is given by:

$$H(x) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left[-\left(\frac{x-u}{\sigma}\right)\right], & \xi = 0 \end{cases}$$

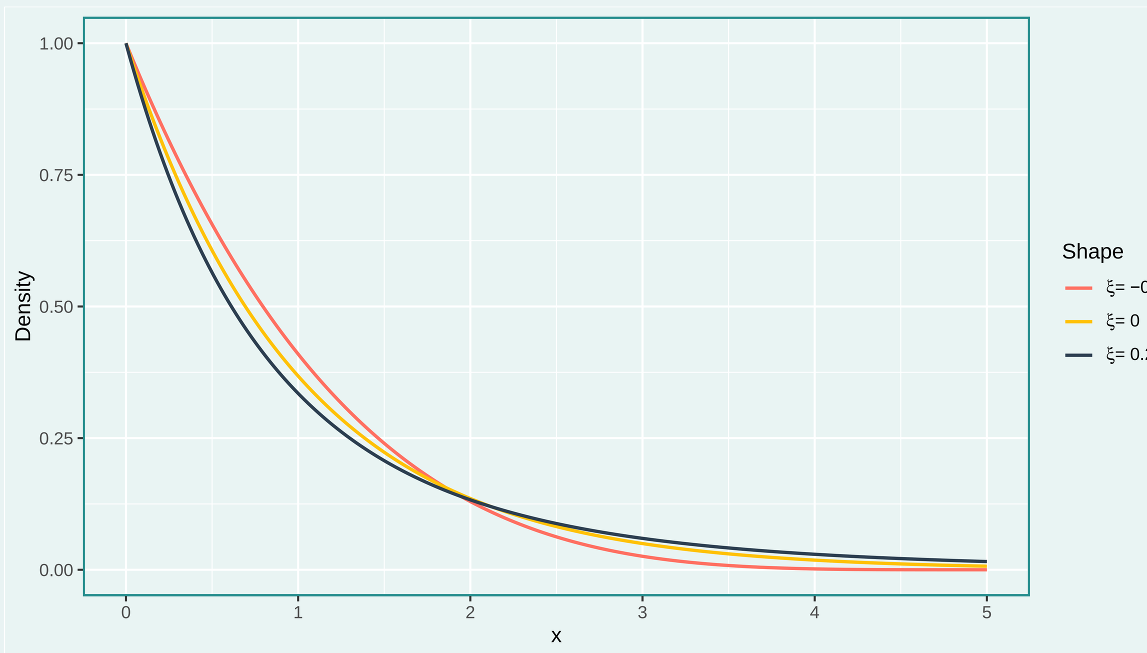

Figure 5: Probability density function for GPD distribution with $u = 0$, $\sigma = 1$ and different values for ξ.

4. Adding Measurement Errors

From here on, we model the magnitudes of earthquakes as GPD distribution and assume that measurement errors are i.i.d $\mathcal{N}(0, \sigma_N^2)$, where σ_N is known and the threshold u is fixed.

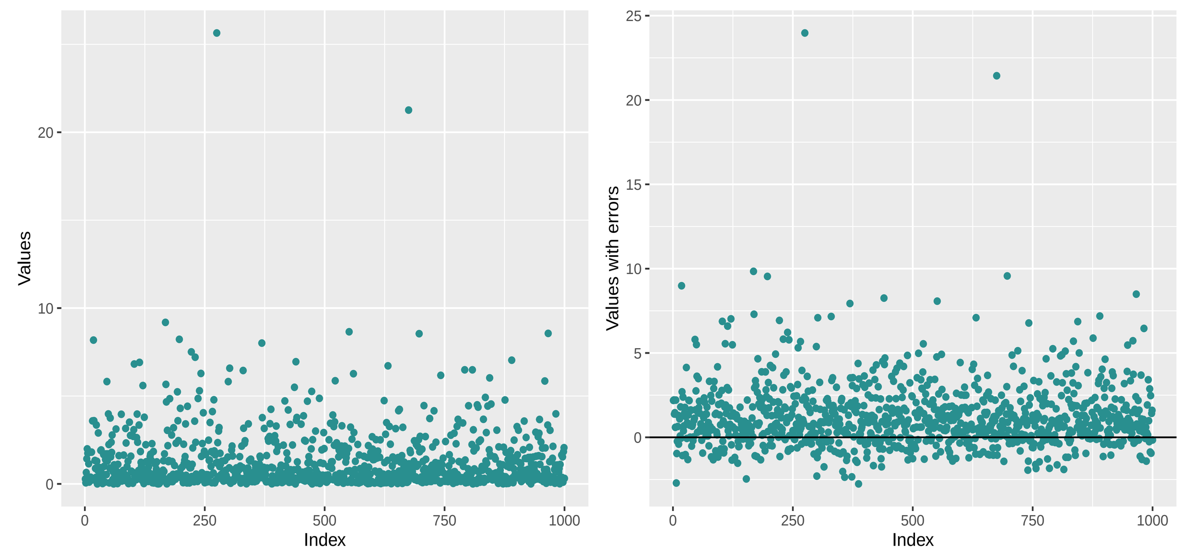

Figure 6: Simulated data from a GPD distribution with threshold $u = 0$, $\sigma = 1$, and $\xi = 0.2$, before (left) and after (right) adding $\mathcal{N}(0, 1)$ errors.

5. Likelihood for $Y = X + \epsilon$

We want to inference on the parameters σ and ξ after adding the normal errors. One way to do this is to use MLE approach for log-likelihood for $Y = X + \epsilon$, where X is the original data and ϵ are errors.

Using convolution, the PDF for Y is given by:

$$f_Y(y \mid \sigma, \xi) = \int_u^\infty f_X(x \mid \sigma, \xi) \cdot \frac{1}{\sqrt{2\pi\sigma_N^2}} e^{\frac{-(y-x)^2}{2\sigma_N^2}} \, dx$$

where X $\sim$ $GPD(u, \sigma, \xi)$.

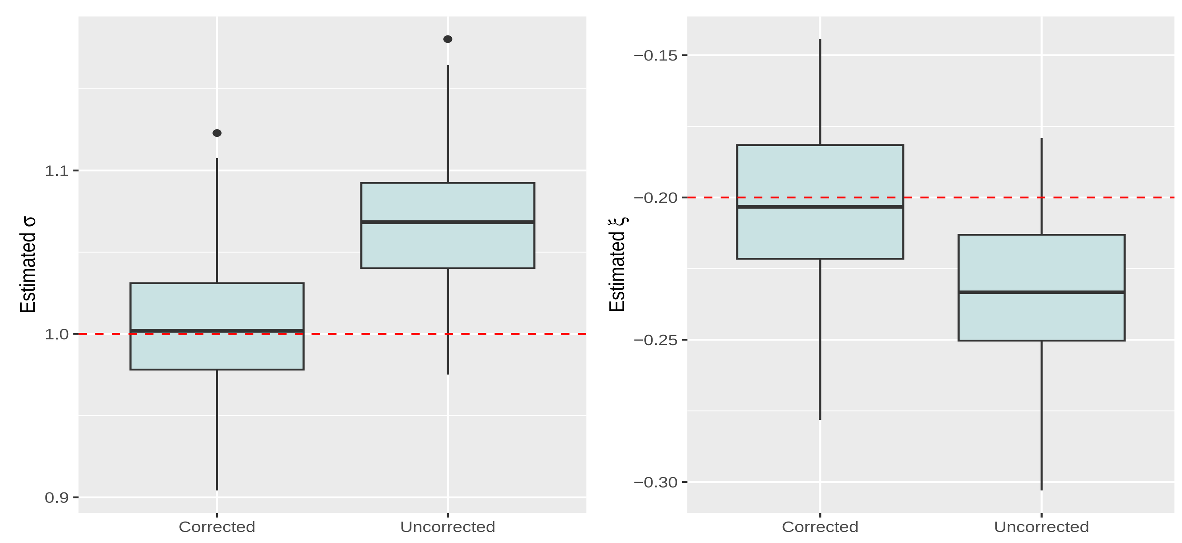

Figure 7: Corrected figure shows the estimates gotten using the log-likelihood for Y, while uncorrected uses the log-likelihood for X. The red dotted line shows the true values of the parameters. We have fixed threshold $u = 0$ and $\sigma_N = 0.1$.

6. ETAS Model

Epidemic Type Aftershock Sequence (ETAS) model is an extension of Hawkes process that is mainly used to model seismic occurrences and their aftershocks.

The modified Omori formula describes the rate of aftershocks as a power law decay function. So, the intensity function for ETAS model is given by

$$\lambda^*(t) = \mu + \sum_{i:t_i<t} \frac{K_0}{(t - t_i + c)^p} \cdot e^{\alpha(m_i - M_r)}$$

where K_0, c, p, and α are parameters, μ is the background intensity, m_i are magnitudes of earthquakes, and M_r is the pre-determined cut-off magnitude.

7. Further Research

- If σ_N **is unknown**, simply using it as an additional parameter in the log-likelihood for Y increases computational demands and inaccuracies in estimating parameters when ξ is positive due to the heavy-tailed nature of the distribution.
- Inferencing on the parameters of the ETAS model is also very computationally intensive and better techniques using Bayesian analysis are being developed to reduce computation time.

The 2011 Japan earthquake shortened the day by about 1.8 microseconds.

kushagra.agarwal@warwick.ac.uk