Modelling Extreme Taiwan Earthquake Occurrences using the Linear Hawkes Process and the Generalized Pareto Distribution

A Bayesian Approach

MATH616 - Independent Studies in Time Series Analysis II
Christian Aaris Lærkedahl Ørnskov

Department of Mathematics and Computer Science
University of Southern Denmark

Department of Applied Mathematics
National Sun Yat-Sen University

2024.06.05

Contents

1 Introduction

During my time as exchange student in Taiwan here at the National Sun Yat-Sen university, I got to experience an earthquake of magnitude 7.4 which was something new coming from Denmark. And even though I managed to sleep through it, I got the idea to try an model the earthquakes happening in Taiwan. Mainly, because earthquakes are more common in Taiwan and accurately modelling them can be helpful both for insurance companies needing to calculate premiums for extreme events and also to diminish the cost of damages of earthquakes.

The model I chose to employ is a mixture of the models and theory I covered in the class "MATH616 - Independent Studies in Time Series Analysis II". The model uses the Linear Hawkes Process to model the arrival times and the Generalized Pareto Distribution to model the magnitudes. This is all implemented in a Bayesian framework to incorporate uncertainty in the estimated parameters in the form of the posterior distribution.

2 Data overview

The data is taken from the "United States Geological Survey" website using their Search Earthquake Catalog Feature. To stay consistent, I chose the catalog "National Earthquake Information Center (NEIC)" which is the national earthquake data archive for the U.S. I restrict the area of interest to be near and around Taiwan, specifically given by the rectangle encompassed by the following longitude and latitude restrictions:

$$\text{North: } 25,78 \qquad \text{East: } 122,61 \qquad \text{South: } 21,36 \qquad \text{West: } 119,38$$

The first few rows of the data frame is shown in table 1 below. The column "arrival" is calculated as the number of days since the first data point.

Table 1: First 6 rows of the earthquake data from NEIC. The column "arrival" was calculated as the number of days since the first data point.

time	latitude	longitude	depth	place	arrival	mag
1973-02-10 03:32:00	24.272	121.751	56	35 km NNE of Hualien City	0.000000	4.8
1973-02-14 00:49:00	22.288	121.55	38	88 km ENE of Hengchun	3.886806	5.9
1973-03-26 22:53:00	24.530	121.972	29	33 km SE of Yilan	44.806250	4.8
1973-04-23 13:29:00	24.005	121.505	42	10 km WNW of Hualien City	72.414583	5.2
1973-06-05 13:29:00	23.208	121.507	33	85 km S of Hualien City	115.414583	4.6
1973-06-07 06:22:00	22.363	121.105	33	54 km NE of Hengchun	117.118056	4.4

The data shown in table 1 is a cleaned version of the entire data file. When applying the Bayesian framework later, only the columns "arrival" and "mag" will be used.

3 Model description

Broadly, the dataset gives rise to the two-dimensional stochastic process given by the arrival times and their corresponding magnitudes $(t_i, w_i), i = 1, \dots, n$. More precisely, this is called a **marked point process** since the arrivals $t_1, ..., t_n$ constitutes a point process and each arrival is associated with a "mark" $w_1, ..., w_n$, in this case the magnitude of the earthquake. To develop an accurate model for the marked point process, we need a joint density $f(t, w|\mathcal{H}_t)$ for the model, where $\mathcal{H}_t$ represents the history of points up until time t, but not including t. Using the rules of conditional probability we can write:

$$f(t, w|\mathcal{H}_t) = \lambda(t|\mathcal{H}_t) \cdot g(w|t, \mathcal{H}_t)$$

Here $\lambda(t|\mathcal{H}_t)$ represents the density of the arrivals conditioned on the history of previous points while $g(w|t, \mathcal{H}_t)$ represents the density of the marks conditioned on the current arrival time and the history of previous points. Generally, $\lambda(t|\mathcal{H}_t)$ is called the **conditional intensity function**. For simplicity, we employ the assumption that the magnitudes is not dependent on the time of arrival nor the history of previous arrivals. Then the joint density becomes:

$$f(t, w|\mathcal{H}_t) = \lambda(t|\mathcal{H}_t) \cdot g(w)$$

While the restriction that the magnitudes is not dependent on the history is quite restrictive, it allows us to model the two dimensions somewhat separately. To capture the nature of aftershocks in earthquakes we choose a self-exciting conditional intensity function, meaning that an arrival increases the probability of new arrivals shortly after. These type of processes is also known as Hawkes Processes. Additionally, as we are interested in the extreme magnitudes of earthquakes, we choose the generalized pareto distribution (GPD) to model the magnitudes above some threshold u. This is also referred to as a **Peak Over Threshold** (POT) model.

3.1 The Linear Hawkes Process

The Linear Hawkes Process is a self-exciting point process. This is especially interesting for modelling earthquakes, since it allows us to incorporate the nature of aftershocks in earthquakes. The conditional intensity of the Linear Hawkes is generally a combination of background intensity and a triggering function given by:

$$\lambda(t|\mathcal{H}_t) = \mu(t) + \sum_{i:t_i \leq t} \nu(t|(t_i, w_i))$$

where t_i are the previous observed points. In this analysis, the following trigger function is used:

$$\nu(t|(t_i, w_i)) = c(w_i - w_u) \cdot g(t - t_i) = e^{\beta \cdot (w_i - w_u)} \cdot \frac{k}{(t - t_i + c)^p}$$

For each previous marked arrival, $c(w_i)$ incorporates the effect of the size of the mark w_i above thethreshold w_u and $g(t - t_i)$ models the effect of time since the arrival. This is based on the paper from Yosihiko Ogata's [3]. Also, the background intensity is assumed to be constant. In total the conditional intensity function is then given by:

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{i:t_i < t} e^{\beta \cdot (w_i - w_u)} \cdot \frac{k}{(t - t_i + c)^p}$$

The parameters and their explanation are as follows:

- μ is the background intensity or minimum "probability" of earthquakes.
- β is a scaling factor for the effect of the magnitude of the earthquake. In total the effect of the magnitude size is exponential.
- k is a scaling factor for the time since arrival.
- c models the delay before the effect of an earthquake decays. This also ensures the effect of aftershocks does not approach infinity as $t - t_i$ goes to 0.
- p models the rate of decay of the aftershocks.

Another important part of point processes is the **intensity measure** $\Lambda(\tau)$, also called the compensator function:

$$\begin{aligned}\Lambda(\tau) &= \int_0^\tau \lambda(t)dt \\ &= \int_0^\tau \mu + \sum_{i:t_i<t} e^{\beta\cdot(w_i-w_u)} \cdot \frac{k}{(t-t_i+c)^p}dt \\ &= \mu\tau + \int_0^\tau \sum_{i:t_i<t} e^{\beta\cdot(w_i-w_u)} \cdot \frac{k}{(t-t_i+c)^p}dt\end{aligned}$$

As the intensity will be important in both calculating the likelihood and in the goodness of fit testing, we derive an expression of it below by computing the integral.

Since the sum loops over a finite set of observations and the integral is of finite length we can switch the order of summation and integration. Additionally, for each individual point t_i, we need only consider the time in between τ and t_i as we are considering the effect of t_i on the current time τ. This gives:

$$\Lambda(\tau) = \mu\tau + \sum_{i:t_i\le\tau} e^{\beta\cdot(w_i-w_u)} \cdot k\int_{t_i}^\tau \frac{1}{(t-t_i+c)^p}dt$$

We use u-substitution and set $u = t-t_i+c$, then $du = dt$. For $t = t_i$, $u = c$ and for $t = \tau$, $u = \tau - t_i + c$. Thus, the integral becomes:

$$\int_c^{\tau-t_i+c} u^{-p}du = \left.\frac{u^{1-p}}{1-p}\right|_c^{\tau-t_i+c} = \frac{1}{1-p}\left((\tau-t_i+c)^{1-p} - c^{1-p}\right), \qquad p \neq 1$$

So that the intensity measure becomes:

$$\Lambda(\tau) = \mu\tau + \frac{k}{1-p}\sum_{i:t_i\le\tau} e^{\beta\cdot(w_i-w_u)}\left((\tau-t_i+c)^{1-p} - c^{1-p}\right), \qquad p \neq 1$$

In the special case of $p = 1$, the integral of u^{-1} is $\log(u)$ so that the intensity measure becomes:

$$\Lambda(\tau) = \mu\tau + \sum_{i:t_i\le\tau} e^{\beta\cdot(w_i-w_u)} k\log\left(\frac{\tau-t_i+c}{c}\right)$$

3.2 The Generalized Pareto Distribution

The Generalized Pareto Distribution (GPD) is typically used to model excesses over a threshold u. The density of the GPD, $g(w)$ is determined by the parameters σ and ξ and is given by:

$$g(w|\sigma,\xi) = \frac{1}{\sigma}\left(1+\xi\frac{z-u}{\sigma}\right)^{-\frac{1}{\xi}-1}$$

3.3 The Likelihood

The loglikelihood is inspired by the model by McGill and Chavez-Demoulin [1]. The likelihood consists of the product of joint densities but also needs to account for the probability of not observing points in between. This is given by the exponential of the negative intensity measure at time $\tau = T$, where T is the last arrival. Let $\theta = (\sigma, \xi, \mu, \beta, k, c, p)$, then the entire likelihood is given by:

$$\begin{aligned}L\left(\theta|(t_1,w_1),\dots,(t_n,w_n)\right) &= \exp\left\{-\Lambda(T)\right\}\prod_{i=1}^n f(t_i,w_i|\mathcal{H}_t) \\ &= \exp\left\{-\int_0^T \lambda(t|\mathcal{H}_t)\mathrm{d}t\right\}\cdot\prod_{i=1}^n \lambda(t_i|\mathcal{H}_{t_i})\cdot g(w_i|\sigma,\xi)\end{aligned}$$

To ease the computational complexity of the analysis and improve the numerical stability, we want to use the log-likelihood instead. Thus, assuming $p \neq 1$ and by taking the logarithm and applying it's properties we get:

$$\begin{aligned}
\ell(\theta|(t_1, w_1), \ldots, (t_n, w_n)) &= -\int_0^T \lambda(t|\mathcal{H}_t)\mathrm{d}t + \sum_{i=1}^n \log\lambda(t_i|\mathcal{H}_{t_i}) + \sum_{i=1}^n \log g(w_i|\sigma, \xi) \\
&= \mu T + \frac{k}{1-p} \sum_{i:t_i \leq T} e^{\beta\cdot(w_i - w_u)} \left((T - t_i + c)^{1-p} - c^{1-p}\right) \\
&\quad - n\log\sigma - \left(1 + \frac{1}{\xi}\right) \log\left\{1 + \xi\frac{w_i - u}{\sigma}\right\} + \sum_{i=1}^n \log\lambda(t_i|\mathcal{H}_{t_i})
\end{aligned}$$

3.4 Goodness of fit testing

The goodness of fit testing is carried out separately for each arrivals and the magnitudes. From the Bayesian framework, a posterior distribution for each parameter is extracted. From this posterior distribution the posterior mean is calculated which will represent the point estimate of the parameters. For the GPD, the point estimates of σ and ξ will be used to create QQ-plots and construct statistical tests for the goodness of fit. The statistical test used will be the Anderson-Darling test of the `goftest` package in `R` [5].

For the Linear Hawkes Process, we'll introduce a theorem without proof.

Theorem 3.1 (*The Random Time Change Theorem* [2])

Let $N(t)$ be a simple point process with a natural filtration $\mathcal{H}_t$, a bounded, strictly positive conditional intensity $\lambda(t) = \lambda(t|\mathcal{H}_t)$ anda compensator function $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$ that is not almost surely bounded. Define $\tau = \Lambda(t)$, then under the random time change $t \longrightarrow \tau$, the transformed process defined as:

$$\tilde{N}(\tau) = N\left(\Lambda^{-1}(\tau)\right) = N(t)$$

is a homogeneous Poisson proces with unit rate.

Essentially, the inter-arrival times of the transformed arrivals $\{\tau_i\} = \{\Lambda(t_i)\}, i = 1, \ldots, n$ will follow a Poisson process with rate $\lambda = 1$, if the fitted model is a suitable fit to the arrival times. Hence, to perform goodness of fit testing on the fitted Linear Hawkes model, we'll transform the arrival times using the posterior means of each parameter and test the inter-arrival times of these transformed arrivals against the exponential distribution with parameter $\lambda = 1$. To this end, QQ-plots will be constructed and a Kolmogorov-Smirnov test will be run.

4 The Bayesian Framework

For the model fitting and parameter analysis, we'll apply a Bayesian framework using MCMC simulation with the Metropolis-Hastings algorithm. To leverage maximum likelihood estimation, we'll employ an empirical Bayes approach by using the MLE parameters as initial value of the chain and as means of the prior distribution.

The parameters of the model is constrained such that ξ must be negative while σ, μ, β, k, c and p must be positive. To circumvent specifying priors that satisfies this and to avoid numerical instabilities for proposals near zero, the model is reparameterized in the following way:

$$\bar{\sigma} = \log \sigma, \quad \bar{\xi} = \log -\xi, \quad \bar{\mu} = \log \mu, \quad \bar{k} = \log k, \quad \bar{c} = \log c, \quad \bar{p} = \log p$$

Thus the priors and proposals is specified in terms of the logarithm of the parameters. This makes them easy to use as non-informative priors, which is good since we have no information on the prior distribution beforehand. For each parameter the prior will be a normal distribution with mean equal to the MLE of that parameter and standard deviation 10. The proposal will be the current value plus an error term ϵ that is normally distributed with mean 0 and standard deviation σ_ϵ, not to be confused with the parameter σ. Denoting the standard deviation of the error term as weights, then the weight of each parameter will be set to:

$$\sigma_{\bar{\sigma}} = 0.1 \qquad \sigma_{\xi} = 0.02, \qquad \sigma_{\mu} = 0.15, \qquad \sigma_{\beta} = 0.1, \qquad \sigma_{k} = 0.1, \qquad \sigma_{c} = 0.01, \qquad \sigma_{p} = 0.05$$

The choice of the weights was chosen empirically to balance the exploration of the parameter space.

Additionally, to capture any dependence between the parameters, the acceptance/rejection step is done collectively. Thus, either all new proposals are accepted or non of them are. This also speeds up the computational time as there are fewer calculations of the log-likelihood. To this end, another assumption is employed which is that the complete prior and complete proposal is the product of the individual priors/proposals given by e.g.:

$$p(\theta) = p(\sigma)p(\xi)p(\mu)p(\beta)p(k)p(c)p(p)$$

The Metropolis-Hastings Ratio is given in full as:

$$R_{\theta^*,\theta} = \frac{p(\theta^*) \cdot \ell(\theta^* | \{(t_1, w_1), \ldots, (t_n, w_n)\}) \cdot g(\theta|\theta^*)}{p(\theta) \cdot \ell(\theta | \{(t_1, w_1), \ldots, (t_n, w_n)\}) \cdot g(\theta^*|\theta)}$$

As the proposal distribution is normal, it is symmetrical around it's mean. Thus $g(\theta|\theta^*)$ will cancel out with $g(\theta^*|\theta)$ such that we are left with:

$$R_{\theta^*,\theta} = \frac{p(\theta^*) \cdot \ell(\theta^* | \{(t_1, w_1), \ldots, (t_n, w_n)\})}{p(\theta) \cdot \ell(\theta | \{(t_1, w_1), \ldots, (t_n, w_n)\})}$$

Additionally, as we are calculating the log-likelihood, we also need the log probabilities of the priors so that:

$$p(\theta)_{\log} = p(\sigma)_{\log} + p(\xi)_{\log} + p(\mu)_{\log} + p(\beta)_{\log} + p(k)_{\log} + p(c)_{\log} + p(p)_{\log}$$

Then the Metropolis-Hasting Ratio is given by:

$$R_{\theta^*,\theta} = \exp\left\{(p(\theta^*)_{\log} + \ell(\theta^* | \{(t_1, w_1), \ldots, (t_n, w_n)\})) - (p(\theta)_{\log} + \ell(\theta | \{(t_1, w_1), \ldots, (t_n, w_n)\}))\right\}$$

Lastly, due to the nature of earthquake magnitudes, we'll introduce one more constraints, and that is the magnitudes cannot be infinite. This implies that the GPD must have an upper-bound so that ξ must be negative. This is forced by the reparametrization of ξ. The benefit of the Bayesian Framework also allows me to estimate the posterior distribution of the return level by using the posterior distribution of σ and ξ by the following calculation:

$$Z_{100}^{(i)} = \text{threshold} + Q_{\text{GPD}}\left(1 - \frac{1}{100} | \sigma^{(i)}, \xi^{(i)}\right), \quad i = 1, \ldots, 25000$$

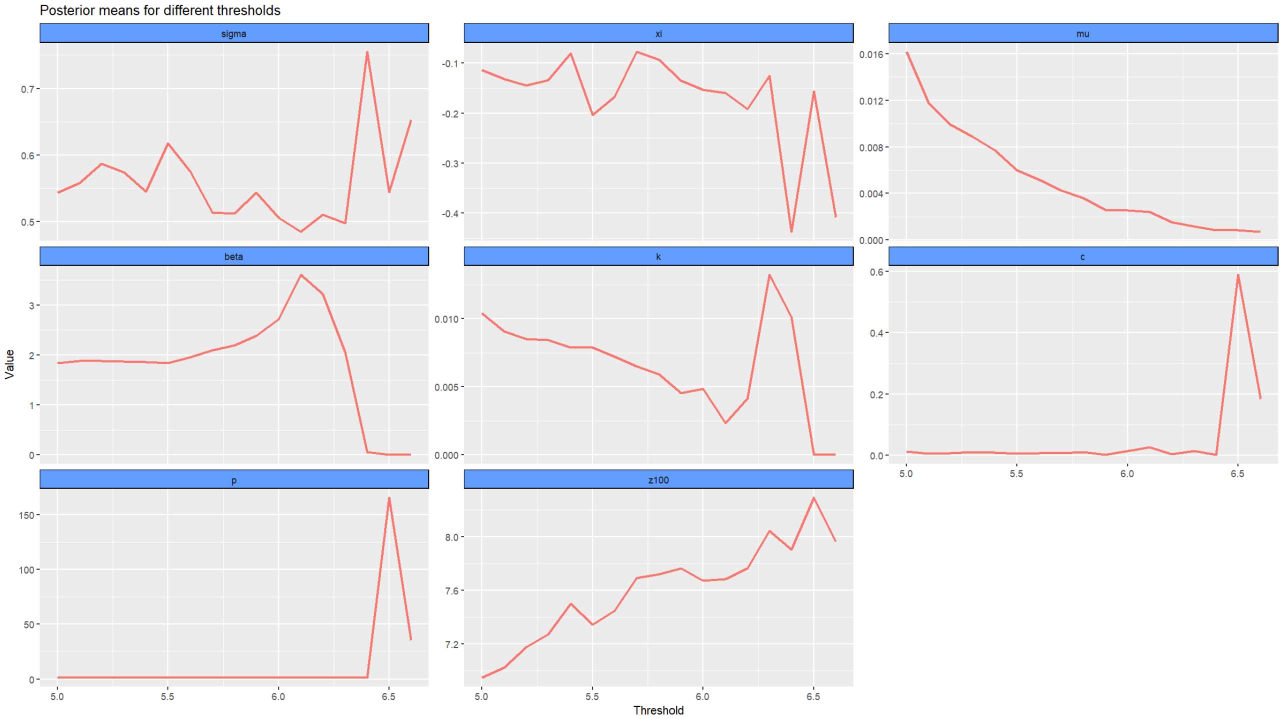

Figure 1: Point estimates for the posterior mean at different thresholds.

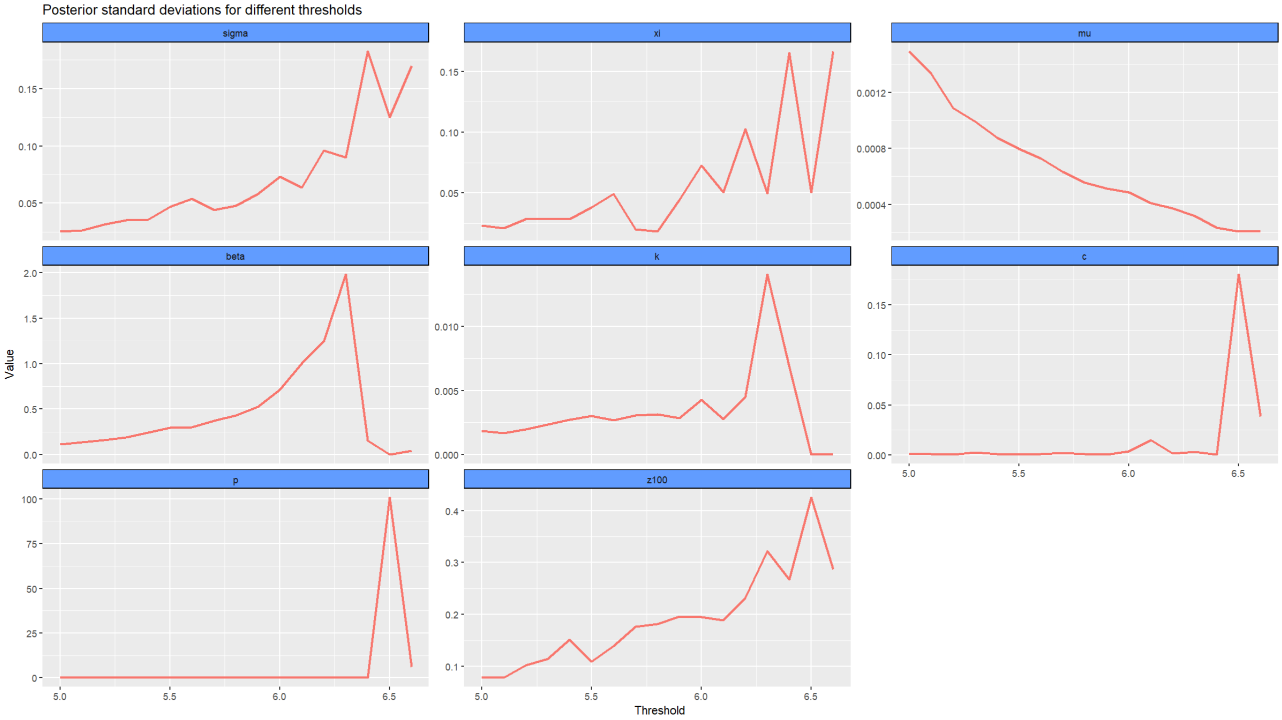

Figure 2: Point estimates for the posterior standard deviation at different thresholds.

5 Results

The Metropolis Hasting was run for the thresholds $5.0, 5.1, 5.2, \ldots, 6.5, 6.6$. Each chain was allowed to run for 25.000 steps with a burn-in period of 10.000. This was due to some parameters being more unstable than others and hence needed a longer burn-in and a larger sample to get accurate point estimates. This happened mostly for higher thresholds where the number of exceedances was too low, resulting in the variance of the posterior distribution being too high. Figure 1 and 2 shows the point estimates of the mean and standard deviation at the different thresholds. Accounting for the uncertainty in the estimation, the estimates seems relatively stable up until around 6.1 or 6.2.

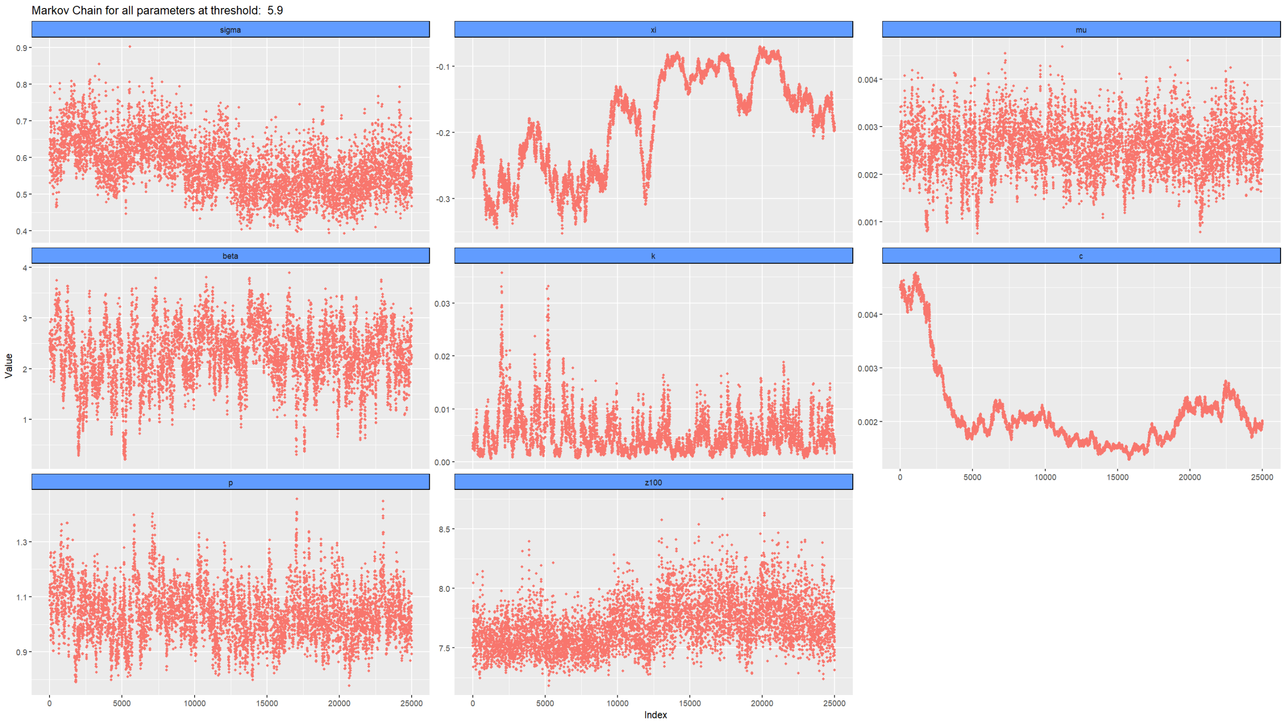

Figure 3: Markov Chains of all parameters. From top left and to the right each row goes σ , ξ, μ, then β, k, c, lastly p and the 100-year return level. Especially ξ and c needs longer time to converge.

The fit of each model can be tested with the described goodness of fit methods. Specifically table 2 shows the p-values of the Kolmogorov-Smirnov tests for the transformed arrival times and the Anderson-Darling test of the magnitudes, at different thresholds. Interestingly, the p-value of the Anderson-Darling test increases with higher thresholds. The reason for this is because of two things. First, with increasing thresholds the number of observations decreases and the accuracy of the test therefore decreases. Additionally, for lower thresholds many observations have the same magnitude, as they are only stored with 1 digit after the decimal. Thus, the observations comes in discrete magnitudes while the GPD models continuous values and the lower the threshold the more points are equal. This effect can be seen on the QQ-plot in figure 4 which shows a good fit if you take into account the discrete values. Based on table 2 and figure 1 & 2, I decide to continue my analysis on the Markov chains generated by threshold 5.9.

The Markov chains of each parameter at threshold 5.9 is shown in figure 3 along with the chain for the estimated 100-observation return level. It can also be seen from figure 3 that especially ξ and c takes a long time to converge and is constantly more unstable than the other parameters. Based on the plots, I decided to increase the burn-in period to 13.000 when calculating the mean and standard deviation. Figure 3 is also a good example of why the posterior distribution in some cases can be tough to find analytically. Especially, the posterior distribution of k is highly skewed.

The updated point estimates of the parameters at threshold 5.9 is presented in table 3. Based on the standard deviation in the table the point estimate of the mean seems to be precise. To confirm this, we look at the QQ-plot of the GPD model in figure 4 and the QQ-plot of the inter-arrival times of the transformed arrival times in figure 5.

Table 2: P-values of Kolmogorov-Smirnov (KS) tests for the transformed arrival times and Anderson-Darling (AD) tests for the magnitudes at different thresholds.

Threshold	**KS p-value**	**AD p-value**
5.0	0.582	0.00000104
5.1	0.745	0.0000363
5.2	0.837	0.000648
5.3	0.482	0.00202
5.4	0.847	0.00275
5.5	0.867	0.0144
5.6	0.982	0.0217
5.7	0.875	0.0155
5.8	0.993	0.0277
5.9	0.955	0.0128
6.0	0.998	0.0587
6.1	0.512	0.0845
6.2	0.248	0.122
6.3	0.161	0.303
6.4	0.761	0.885
6.5	$<$ 2e-16	0.603
6.6	2.06e-12	0.889

Table 3: Posterior means and standard deviations at threshold 5.9 with burn-in period of 13.000

Statistic	σ	ξ	μ	β	k	c	p	z_{100}
Posterior Mean	0.5360	-0.1224	0.0026	2.3449	0.0048	0.0019	1.0286	7.7881
Posterior SD	0.0565	0.0324	0.0005	0.5368	0.0029	0.0004	0.0912	0.1957

The return level is of most interest and the posterior mean in table 3 suggests that on average, the value $7.78 \approx 7.8$ will be exceeded once every 100 earthquakes in Taiwan. In total there were 76 observations with a magnitude larger than 5.9 in a period over almost 48.9 years which is 64.3 years for 100-observations. If we call this the 65-year return level. Then a crude extrapolation is that Taiwan on average would experience an earthquake of size 7.8 once every 65 years. To improve on this we use the density of the entire chain after the burn-in period. The density of the 100-observation return level is shown in figure 6. The 2.5% quantile and the 97.5% quantile was estimated to be 7.44574 and 8.19775 respectively. Therefore, the true underlying 65-year return level is with high probability in between 7.45 and 8.2. Additionally, the 155-observation return level corresponds to a 100-year return level and the point estimate in this case is approx 7.9 with the 2.5% quantile being 7.55 and the 97.5% quantile being 8.37.

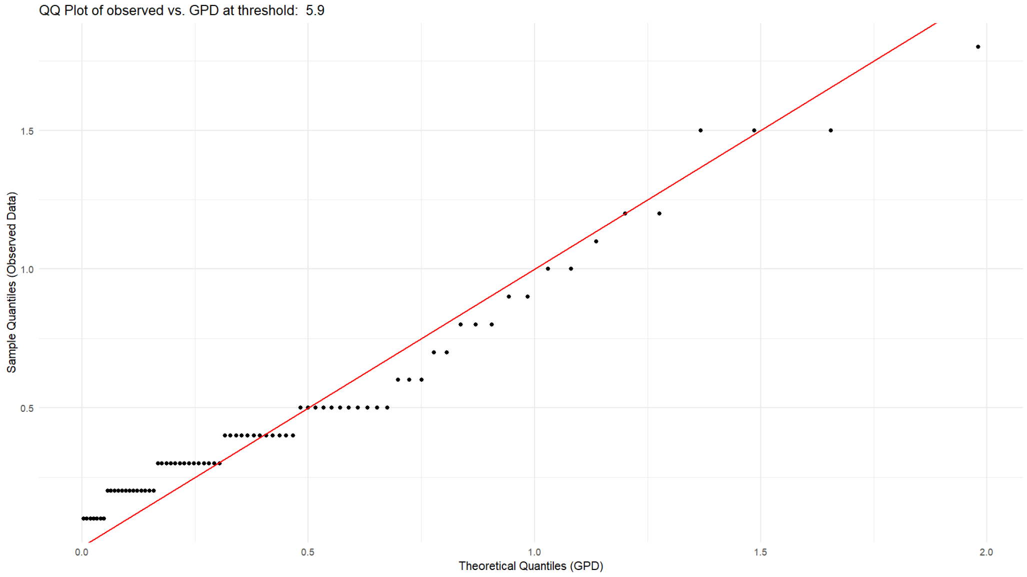

Figure 4: QQ-plot of the GPd distribution for threshold 5.9. Notice the discrete values, which is the reason for the low p-values in the AD-test.

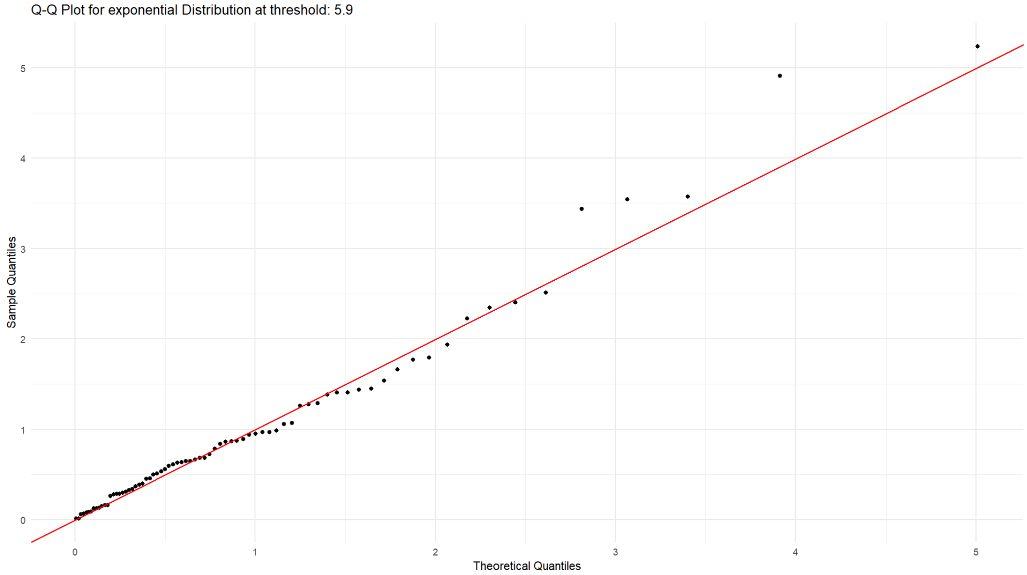

Figure 5: QQ-plot for the inter-arrival times of the transformed arrival times for threshold 5.9.

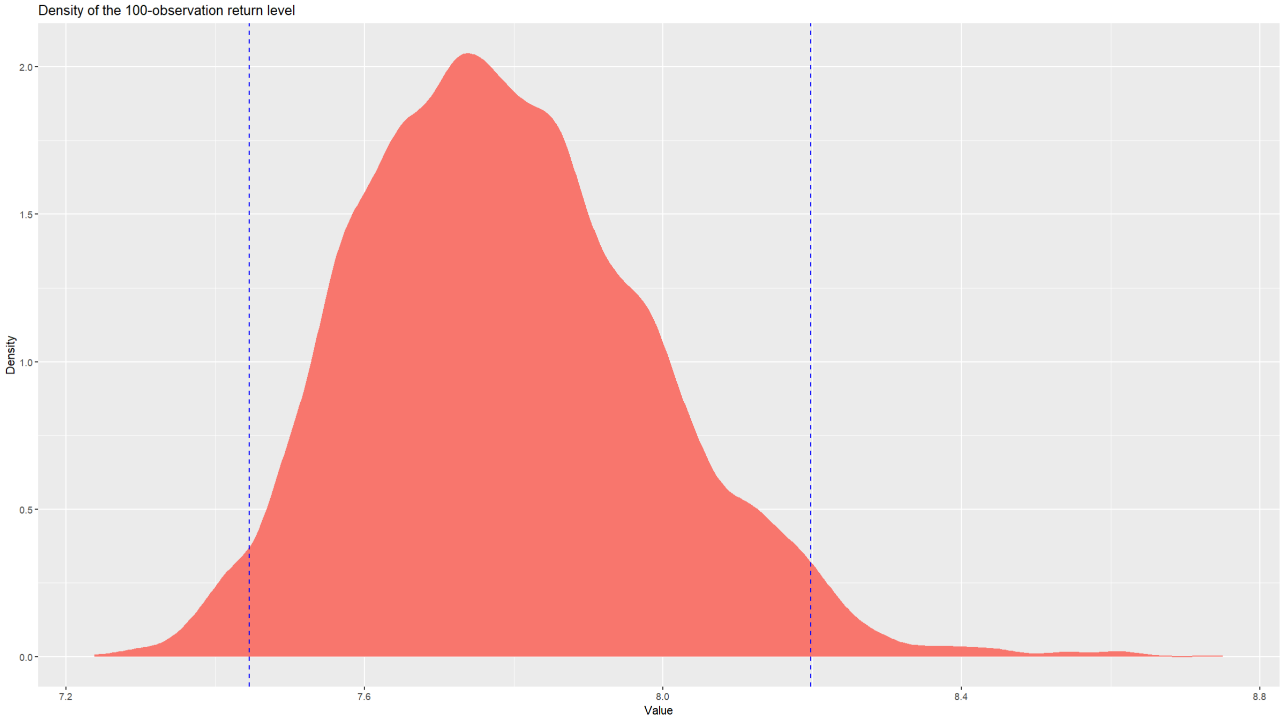

Figure 6: Density of the return level. The lower 2.5 % quantile is 7.45 and the upper 97.5% quantile is 8.2.

6 Discussion

All in all, the framework developed and applied in this paper seems to applicable in most cases. It can clearly be seen the the models faults when the threshold is set to high and the number of observations becomes. However, for valid thresholds, the model is consistent and the point estimates are stable. The power of the Bayesian framework comes to light in the posterior distribution of the return level, since it gives a complete sample of the distribution of the return level, where with MLE every value of the distribution must be estimated. It was found that the 100-observation return level, corresponding to a 65-year return level was in between 7.45 and 8.2 with point estimate at 7.8. For the 155-observation/100-year return level this was 7.55 and 8.37 with point estimate at 7.9. This seems mainly to be a results of the 1999 earthquake that reached 7.7 in magnitude. All the other earthquakes in the dataset had magnitude 7.4 or lower. It is therefore reasonable to claim that the 1999 earthquake was a 1 time in a century event.

One thing to include in this model, that hasn't been accounted for, is the dependence of the history on the magnitude of an arrival. Currently, the probability of aftershocks increasing after an earthquake is implemented in the conditional intensity, but the size of the aftershocks should mostly be less than main earthquake. This could maybe be implemented by adding more parameters to the model, but the computational complexity of having 8, 9, 10 or more parameters might not be worth. Therefore an entirely different model might be needed to implement this such as a clustering and branching structure model [4].

References

[1] V. Chavez-Demoulin and J.A McGill. High-frequency financial data modeling using hawkes processes. *Journal of Banking & Finance*, 2012.

[2] Yuanda Chen. Goodness-of-fit for fitting real data with hawkes processes. 2016.

[3] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 1988.

[4] Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Department of Mathematical Sciences, Aalborg University. Research Report Series No. R-2011-03*, 2011.

[5] M. A. Stephens and V. Choulakian. Goodness-of-fit tests for the generalized pareto distribution.