Hawkes processes

Estimates of a Hawkes process

Due date : June 5th midnight

Instructions

You must give a Jupyter Notebook in Python/Sage/Julia (the choice is yours). The criteria by which the score is given are

1. the correctness of the results
    2. the overall quality of the presentation
    3. well-written answers to theoretical questions
2. comment on the code in the spirit of literate programming.

An example of what is expected is this kind of post

https://medium.com/analytics-vidhya/credit-risk-modelling-in-python-3ab4b00f6505

1. If you feel very inspired by a topic don't hesitate to add some additional perspectives

Context

In Cyber security, attacks often occur by bursts. The file

https://nextcloud.r2.enst.fr/nextcloud/index.php/s/DzfPE8Axx2H2afM

contains the dates of plenty of attacks, sorted by category and date of appearance. See the file

https://nextcloud.r2.enst.fr/nextcloud/index.php/s/xW2b6FF9ikPR8Gs

for a complete description of its content and of the motivations behind this analysis. We are going to make a simplified analysis by modeling time of attacks by a Hawkes process with intensity

$$\dot{y}(N,t) = \alpha + \beta \int_0^{t^-} e^{-\gamma(t-s)} \, \mathrm{d}N(s)$$

We recall that from Exercise 3.4.2, we know that log-likelihood is given by

$$\log L(N,t) = \sum_{T_n \leq t} \log\left(\alpha + \beta \sum_{j=1}^{n-1} e^{-\gamma(T_n - T_j)}\right) + (1-\alpha)t - \frac{\beta}{\gamma} \sum_{T_n \leq t} \left(1 - e^{-\gamma(t-T_n)}\right).$$

Your work

- Extract from the file above the attacks of type “HACK”. Here is my code (which probably can be improved)

```
import pandas as pd
from datetime import datetime
from scipy.optimize import minimize,root_scalar
import numpy as np
import matplotlib.pylab as plt
df=pd.read_csv("PRC.csv",sep=';')
dff=df.loc[df['Type of breach']=="HACK",'Date Made Public']
dff=dff.apply(lambda x:datetime.strptime(x,'%m/%d/%Y'))
debut=min(dff)
dff=dff-debut
dff=dff.apply(lambda x:x.days).sort_values()
l=np.asarray(dff)[1:]
plt.scatter(l,1+np.arange(len(l)),s=0.1)
plt.show()
```

  The dates are expressed in days starting from the first event.

- Using the `scipy` optimization library, estimate α, β, γ which maximize the log-likelihood for these data.
- Now that we have an estimate of $y(N,t)$, we know that in theory, the sequence

  $\left(y(N, T_q(N)),\ n \geq 1\right)$

  must have the law of a Poisson process of intensity 1. Compute this sequence and run the test you want on the inter-arrivals to assess the proximity between it and the law of a Poisson process.

Necessary files

The Nextcloud link is broken...