Modelling Tweets as a Hawkes Process

Vishnu Lokhande

October 24, 2016

1 Description

I have used an R package named **ptproc** [1] to fit the parameters of the Hawkes process model to the data. As we know the conditional intensity of the Hawkes process is given by

$$\lambda(t) = \mu + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)} \tag{1}$$

I have used the inditial parametres for the model as

$$\mu = 0.1, \alpha = 1, \beta = 0.0001 \tag{2}$$

I have analysed for **TaylorSwift** for the year 2012 with and without peak tweets removed. For removing the peak tweets, I used a cut-off as I used previously.

2 Tweets from year 2012

These were the final parameters of the Hawkes model after fitting the model using **ptproc**.

$$\mu = 1.1 * 10^{-05}, \alpha = 1.067, \beta = 6.68 * 10^{-02} \tag{3}$$

The AIC of the fitted model and an Homogeneous Poisson Process is as follows

$$\text{Model AIC } = 73884 \tag{4}$$

$$\text{H. Pois. AIC } = 46577 \tag{5}$$

The following plot, Fig.1, shows the tweets counts in blue and the conditional intensity of the fitted model in green. Conditional intensities were integrated over 1 day. Conditional intensities were evaluated using **evalCIF** function present in the R-package.

It is observed that the peaks in the intensity plot are not as high as the actual tweet count. The two plots have been plotted separately in Fig.2 and Fig3.

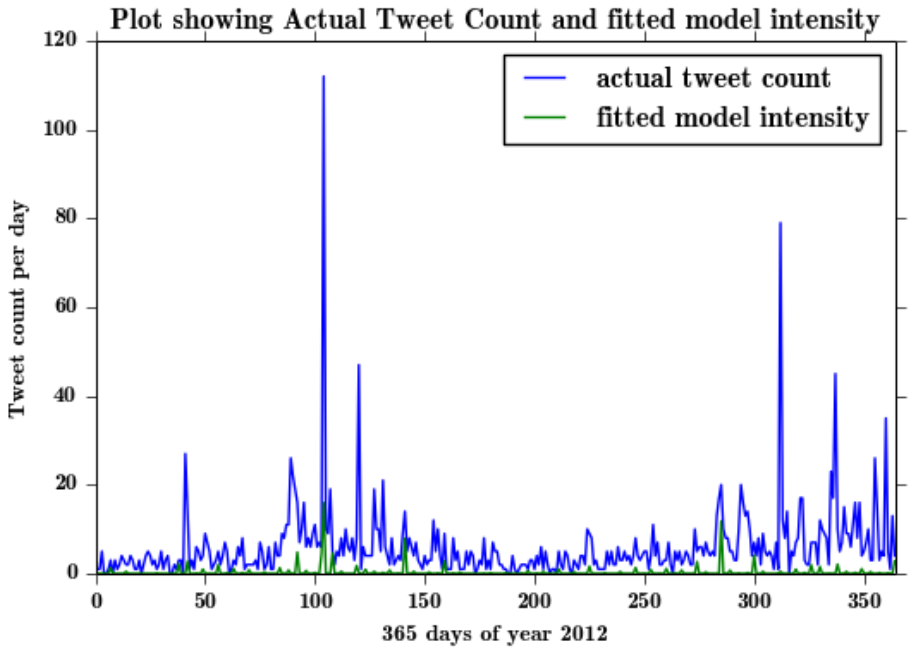

Figure 1

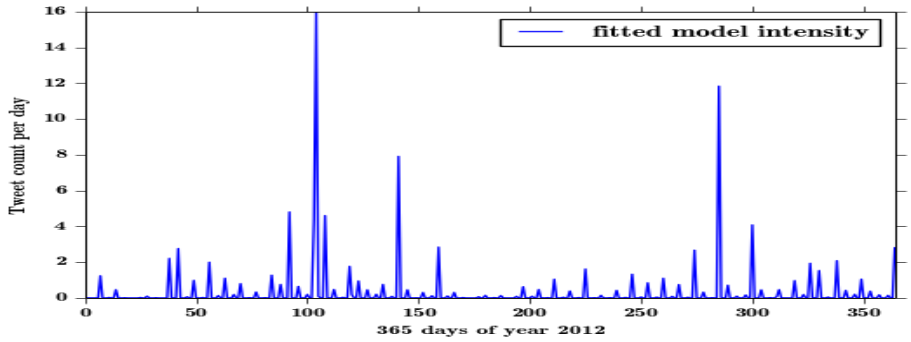

Figure 2: Plot showing fitted model intensity

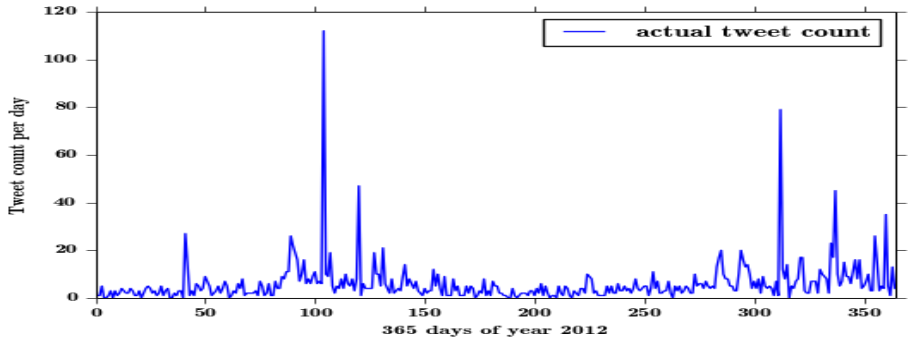

Figure 3: Plot showing Actual Tweet Count

3 Tweets from year 2012 with high peak tweets removed

I have put a cut off of 50 tweets per day. Same parameters were used as in Section.2 for the Hawkes Model. The final parameters of the Hawkes model are

$$\mu = 1.1 * 10^{-05}, \alpha = 1.067, \beta = 6.68 * 10^{-02} \tag{6}$$

The AIC of the fitted model and an Homogeneous Poisson Process is as follows

$$\text{Model AIC} = 68681 \tag{7}$$
$$\text{H. Pois. AIC} = 42905 \tag{8}$$

Fig.4 shows the tweet counts in blue and the conditional intensity of the fitted model in green. Conditional intensities were integrated over 1 day.
Fig.5 and Fig.6 shows the plots separately.

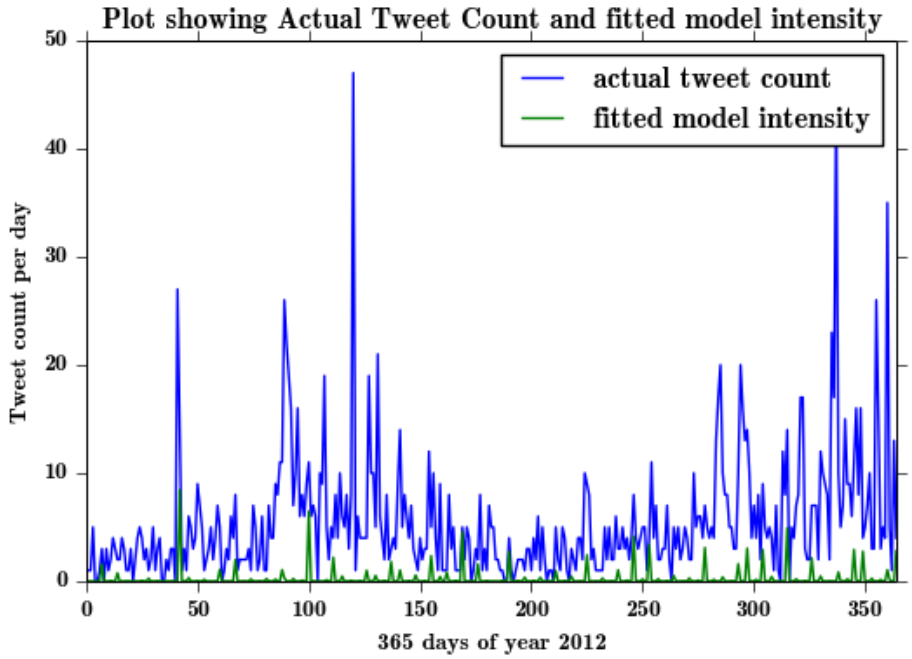

Figure 4

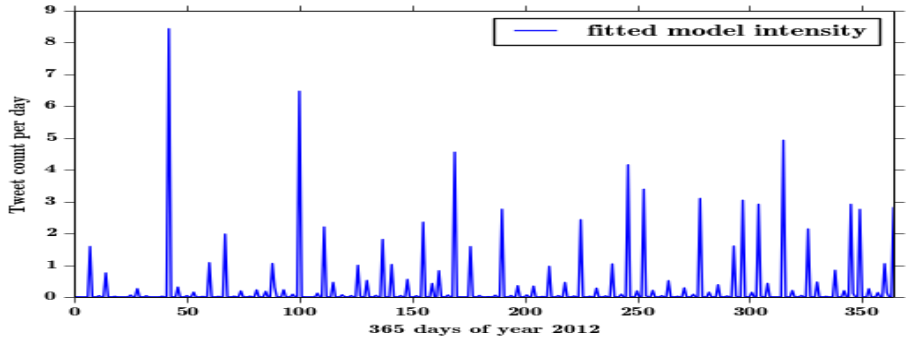

Figure 5: Plot showing fitted model intensity

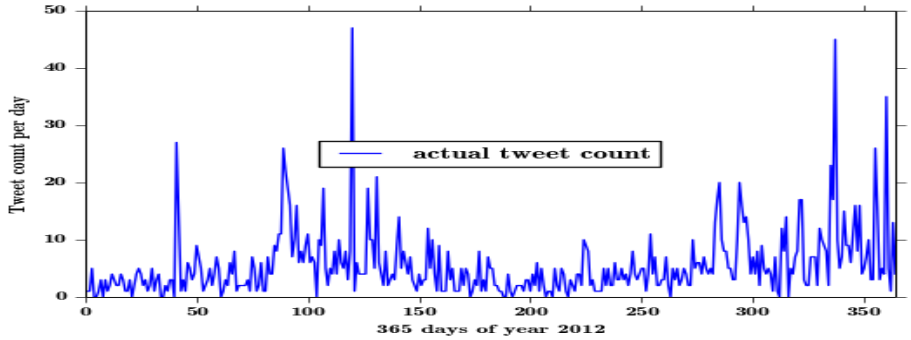

Figure 6: Plot showing Actual Tweet Count

4 Observations

It is observed that the peaks are more aligned between the fitted model and the actual tweet count when there is no cut-off used i.e., when the high peaks are not removed (section2). However in section2, the peaks in the fitted model intensity plot are not as high as the actual sequence.
The R^2 measure for Section.2 is 0.0919 and for Section.3 is 0.1108.

References

[1] Roger D Peng. Multi-dimensional point process models in r. *Department of Statistics, UCLA*, 2002.