Modeling Tweet Arrival Times using Log-Gaussian Cox Processes

Michal Lukasik,[1] **P.K. Srijith,**[1] **Trevor Cohn,**[2] and **Kalina Bontcheva**[1]
[1]Department of Computer Science,
The University of Sheffield
[2]Department of Computing and Information Systems,
The University of Melbourne
{m.lukasik, pk.srijith, k.bontcheva}@shef.ac.uk
t.cohn@unimelb.edu.au

Abstract

Research on modeling time series text corpora has typically focused on predicting what text will come next, but less well studied is predicting when the next text event will occur. In this paper we address the latter case, framed as modeling continuous inter-arrival times under a log-Gaussian Cox process, a form of inhomogeneous Poisson process which captures the varying rate at which the tweets arrive over time. In an application to rumour modeling of tweets surrounding the 2014 Ferguson riots, we show how inter-arrival times between tweets can be accurately predicted, and that incorporating textual features further improves predictions.

1 Introduction

Twitter is a popular micro-blogging service which provides real-time information on events happening across the world. Evolution of events over time can be monitored there with applications to disaster management, journalism etc. For example, Twitter has been used to detect the occurrence of earthquakes in Japan through user posts (Sakaki et al., 2010). Modeling the temporal dynamics of tweets provides useful information about the evolution of events. Inter-arrival time prediction is a type of such modeling and has application in many settings featuring continuous time streaming text corpora, including journalism for event monitoring, real-time disaster monitoring and advertising on social media. For example, journalists track several rumours related to an event. Predicted arrival times of tweets can be applied for ranking rumours according to their activity and narrow the interest to investigate a rumour with a short inter-arrival time over that of a longer one.

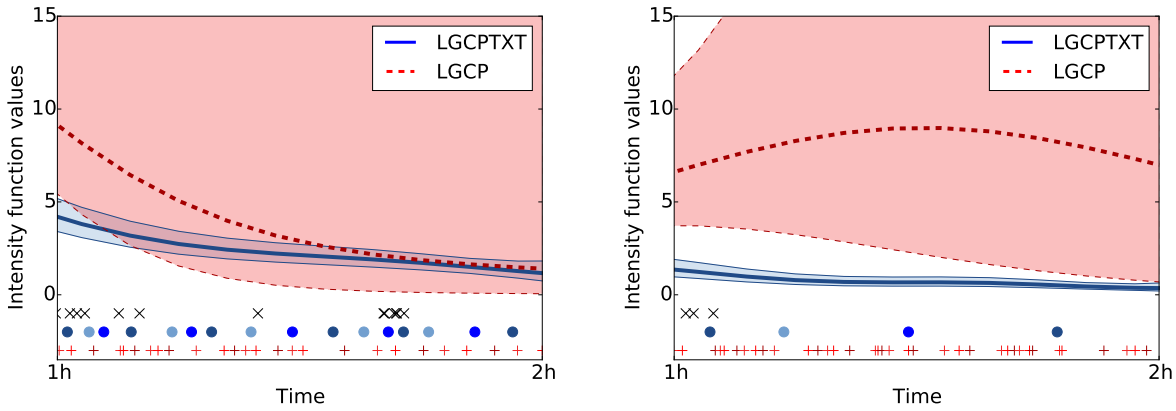
Modeling the inter-arrival time of tweets is a challenging task due to complex temporal patterns exhibited. Tweets associated with an event stream arrive at different rates at different points in time. For example, Figure 1a shows the arrival times (denoted by black crosses) of tweets associated with an example rumour around Ferguson riots in 2014. Notice the existence of regions of both high and low density of arrival times over a one hour interval. We propose to address inter-arrival time prediction problem with log-Gaussian Cox process (LGCP), an inhomogeneous Poisson process (IPP) which models tweets to be generated by an underlying intensity function which varies across time. Moreover, it assumes a non-parametric form for the intensity function allowing the model complexity to depend on the data set. We also provide an approach to consider textual content of tweets to model inter-arrival times. We evaluate the models using Twitter rumours from the 2014 Ferguson unrest, and demonstrate that they provide good predictions for inter-arrival times, beating the baselines e.g. homogeneous Poisson Process, Gaussian Process regression and univariate Hawkes Process. Even though the central application is rumours, one could apply the proposed approaches to model the arrival times of tweets corresponding to other types of memes, e.g. discussions about politics.

This paper makes the following contributions: 1. Introduces log-Gaussian Cox process to predict tweet arrival times. 2. Demonstrates how incorporating text improves results of inter-arrival time prediction.

2 Related Work

Previous approaches to modeling inter-arrival times of tweets (Perera et al., 2010; Sakaki et al., 2010; Esteban et al., 2012; Doerr et al., 2013) were not complex enough to consider their time varying characteristics. Perera et al. (2010) modeled

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 250–255,
Lisbon, Portugal, 17-21 September 2015.



(a) rumour #39 (b) rumour #60

Figure 1: Intensity functions and corresponding predicted arrival times for different methods across example Ferguson rumours. Arrival times predicted by LGCP are denoted by red pluses, LGCPTXT by blue dots, and ground truth by black crosses. Light regions denote uncertainty of predictions.

inter-arrival times as independent and exponentially distributed with a constant rate parameter. A similar model is used by Sakaki et al. (2010) to monitor the tweets related to earthquakes. The renewal process model used by Esteban et al. (2012) assumes the inter-arrival times to be independent and identically distributed. Gonzalez et al. (2014) attempts to model arrival times of tweets using a Gaussian process but assumes the tweet arrivals to be independent every hour. These approaches do not take into account the varying characteristics of arrival times of tweets.

Point processes such as Poisson and Hawkess process have been used for spatio-temporal modeling of meme spread in social networks (Yang and Zha, 2013; Simma and Jordan, 2010). Hawkes processes (Yang and Zha, 2013) were also found to be useful for modeling the underlying network structure. These models capture relevant network information in the underlying intensity function. We use a log-Gaussian cox process which provides a Bayesian method to capture relevant information through the prior. It has been found to be useful e.g. for conflict mapping (Zammit-Mangion et al., 2012) and for frequency prediction in Twitter (Lukasik et al., 2015).

3 Data & Problem

In this section we describe the data and we formalize the problem of modeling tweet arrival times.

Data We consider the Ferguson rumour data set (Zubiaga et al., 2015), consisting of tweets on rumours around 2014 Ferguson unrest. It consists of conversational threads that have been manually labeled by annotators to correspond to rumours[1]. Since some rumours have few posts, we consider only those with at least 15 posts in the first hour as they express interesting behaviour (Lukasik et al., 2015). This results in 114 rumours consisting of a total of 4098 tweets.

Problem Definition Let us consider a time interval $[0, 2]$ measured in hours, a set of rumours $R = \{E_i\}_{i=1}^{n}$, where rumour E_i consists of a set of m_i posts $E_i = \{p_j^i\}_{j=1}^{m_i}$. Posts are tuples $p_j^i = (\mathbf{x}_j^i, t_j^i)$, where $\mathbf{x}_j^i$ is text (in our case a vector of Brown clusters counts, see section 5) and t_j^i is time of occurrence of post p_j^i, measured in time since the first post on rumour E_i.

We introduce the problem of predicting the exact time of posts in the future unobserved time interval, which is studied as *inter-arrival time prediction*. In our setting, we observe posts over a target rumour i for one hour and over reference rumours (other than i) for two hours. Thus, the training data set is $R^O = \{E_i^O\}_{i=1}^{n}$, where $E_i^O = \{p_j^i\}_1^{m_i^O}$ (m_i^O represents number of posts observed for i^{th} rumour). We query the model for a complete set of times $\{t_j^i\}_{m_i^O+1}^{m_i}$ of posts about rumour i in the future one hour time interval.

[1]For a fully automated approach, a system for early detection of rumours (Zhao et al., 2015) could be run first and our models then applied to the resulting rumours.

4 Model

The problem of modeling the inter-arrival times of tweets can be solved using Poisson processes (Perera et al., 2010; Sakaki et al., 2010). A homogeneous Poisson process (HPP) assumes the intensity to be constant (with respect to time and the rumour statistics). It is not adequate to model the inter-arrival times of tweets because it assumes constant rate of point arrival across time. Inhomogeneous Poisson process (IPP) (Lee et al., 1991) can model tweets occurring at a variable rate by considering the intensity to be a function of time, *i.e.* $\lambda(t)$. For example, in Figure 1a we show intensity functions learnt for two different IPP models. Notice how the generated arrival times vary according to the intensity function values.

Log-Gaussian Cox process We consider a log-Gaussian Cox process (LGCP) (Møller and Syversveen, 1998), a special case of IPP, where the intensity function is assumed to be stochastic. The intensity function $\lambda(t)$ is modeled using a latent function $f(t)$ sampled from a Gaussian process (Rasmussen and Williams, 2005). To ensure positivity of the intensity function, we consider $\lambda(t) = \exp(f(t))$. This provides a non-parametric Bayesian approach to model the intensity function, where the complexity of the model is learnt from the training data. Moreover, we can define the functional form of the intensity function through appropriate GP priors.

Modeling inter-arrival time Inhomogeneous Poisson process (unlike HPP) uses a time varying intensity function and hence, the distribution of inter-arrival times is not independent and identically distributed (Ross, 2010). In IPP, the number of tweets y occurring in an interval $[s, e]$ is Poisson distributed with rate $\int_s^e \lambda(t)dt$.

$$p(y|\lambda(t),[s,e]) = \text{Poisson}(y|\int_s^e \lambda(t)dt) = \frac{(\int_s^e \lambda(t)dt)^y \exp(-\int_s^e \lambda(t)dt)}{y!} \quad (1)$$

Assume that n^{th} tweet occurred at time $E_n = s$ and we are interested in the inter-arrival time T_n of the next tweet. The arrival time of next tweet E_{n+1} can be obtained as $E_{n+1} = E_n + T_n$. The cumulative distribution for T_n, which provides the probability that a tweet occurs by time $s + u$ can be obtained as[2]

$$\begin{aligned} p(T_n \le u) &= 1 - p(T_n > u|\lambda(t), E_n = s) \\ &= 1 - p(0 \text{ events in } [s, s+u]|\lambda(t)) \\ &= 1 - \exp(-\int_s^{s+u} \lambda(t)dt) \\ &= 1 - \exp(-\int_0^u \lambda(s+t)dt) \end{aligned} \quad (2)$$

The derivation is obtained by considering a Poisson probability for 0 counts with rate parameter given by $\int_s^{s+u} \lambda(t)dt$ and applying integration by substitution to obtain (2). The probability density function of the random variable T_n is obtained by taking the derivative of (2) with respect to u:

$$p(T_n = u) = \lambda(s+u)\exp(-\int_0^u \lambda(s+t)dt). \quad (3)$$

The computational difficulties arising from integration are dealt by assuming the intensity function to be constant in an interval and approximating the inter-arrival time density as (Møller and Syversveen, 1998; Vanhatalo et al., 2013)

$$p(T_n = u) = \lambda(s+u)\exp(-u\lambda(s+\frac{u}{2})). \quad (4)$$

We associate a distinct intensity function $\lambda_i(t) = \exp(f_i(t))$ with each rumour E_i as they have varying temporal profiles. The latent function f_i is modelled to come from a zero mean Gaussian process (GP) (Rasmussen and Williams, 2005) prior with covariance defined by a squared exponential (SE) kernel over time, $k_{time}(t,t') = a\exp(-(t-t')^2/l)$. We consider the likelihood of posts E_i^{O} over the entire training period to be product of Poisson distribution (1) over equal length sub-intervals with the rate in a sub-interval $[s, e]$ approximated as $(e-s)\exp(f_i(\frac{1}{2}(s+e)))$. The likelihood of posts in the rumour data is obtained by taking the product of the likelihoods over individual rumours.

The distribution of the posterior $p(f_i|E_i^{\text{O}})$ is intractable and a Laplace approximation (Rasmussen and Williams, 2005) is used to obtain the posterior. The predictive distribution $f_i(t_*^i)$ at time t_*^i is obtained using the approximated posterior. The intensity function value at the point t_*^i is then obtained as

$$\lambda_i(t_*^i|E_i^{\text{O}}) = \int \exp\left(f_i(t_*^i)\right) p\left(f_i(t_*^i)|E_i^{\text{O}}\right) df_i(t_*^i).$$

[2] We suppress the conditioning variables for brevity.

Algorithm 1 Importance sampling for predicting the next arrival time

1: **Input:** Intensity function $\lambda(t)$, previous arrival time s, proposal distribution $q(t) = \exp(t; 2)$, number of samples N
2: **for** $i = 1$ **to** N **do**
3: Sample $u_i \sim q(t)$.
4: Obtain weights $w_i = \frac{p(u_i)}{q(u_i)}$, where $p(t)$ is given by (4).
5: **end for**
6: Predict expected inter-arrival time as $\bar{u} = \sum_{i=1}^{N} u_i \frac{w_i}{\sum_{j=1}^{N} w_j}$
7: Predict the next arrival time as $\bar{t} = s + \bar{u}$.
8: **Return:** $\bar{t}$

Importance sampling We are interested in predicting the next arrival time of a tweet given the time at which the previous tweet was posted. This is achieved by sampling the inter-arrival time of occurrence of the next tweet using equation (4). We use the importance sampling scheme (Gelman et al., 2003) where an exponential distribution is used as the proposal density. We set the rate parameter of this exponential distribution to 2 which generates points with a mean value around 0.5. Assuming the previous tweet occurred at time s, we obtain the arrival time of next tweet as outlined in Algorithm 1. We run this algorithm sequentially, i.e. the time $\bar{t}$ returned from Algorithm 1 becomes starting time s in the next iteration. We stop at the end of the interval of interest, for which a user wants to find times of post occurrences.

Incorporating text We consider adding the kernel over text from posts to the previously introduced kernel over time. We join text from the observed posts together, so a different component is added to kernel values across different rumours. The full kernel then takes form $k_{\text{TXT}}((t,i),(t',i')) = k_{time}(t,t') + k_{text}\big(\sum_{p_j^i \in E_i^O} \mathbf{x}_j^i, \sum_{p_j^{i'} \in E_{i'}^O} \mathbf{x}_j^{i'}\big)$. We compare text via linear kernel with additive underlying base similarity, expressed by $k_{text}(\mathbf{x}, \mathbf{x}') = b + c\mathbf{x}^T\mathbf{x}'$.

Optimization All model parameters (a, l, b, c) are obtained by maximizing the marginal likelihood $p(E_i^O) = \int p(E_i^O|f_i)p(f_i)df_i$ over all rumour data sets.

5 Experiments

Data preprocessing In our experiments, we consider the first two hours of each rumour lifespan. The posts from the first hour of a target rumour is considered as observed (training data) and we predict the arrival times of tweets in the second hour. We consider observations over equal sized time intervals of length six minutes in the rumour lifespan for learning the intensity function. The text in the tweets is represented by using Brown cluster ids associated with the words. This is obtained using 1000 clusters acquired on a large scale Twitter corpus (Owoputi et al., 2013).

Evaluation metrics Let the arrival times predicted by a model be $(\hat{t}_1, \ldots, \hat{t}_M)$ and let the actual arrival times be $(t_1, \ldots, t_N)$. We introduce two metrics based on root mean squared error (RMSE) for evaluating predicted inter-arrival times. First is aligned root mean squared error (ARMSE), where we align the initial $K = \min(M, N)$ arrival times and calculate the RMSE between such two subsequences. The second is called penalized root mean squared error (PRMSE). In this metric we penalize approaches which predict a different number of inter-arrival times than the actual number. The PRMSE metric is defined as the square root of the following expression.

$$\frac{1}{K}\sum_{i=1}^{K}(\hat{t}_i - t_i)^2 + \mathbb{I}[M > N] \sum_{i=N+1}^{M} (T - \hat{t}_i)^2 + \mathbb{I}[M < N] \sum_{i=M+1}^{N} (T - t_i)^2 \quad (5)$$

The second and third term in (5) respectively penalize for the excessive or insufficient number of points predicted by the model.

Baselines We consider a homogeneous Poisson process (HPP) (Perera et al., 2010) as a baseline which results in exponentially distributed inter-arrival times with rate λ. The rate parameter is set to the maximum likelihood estimate, the reciprocal of the mean of the inter-arrival times in the training data. The second baseline is a GP with a linear kernel (GPLIN), where the inter-arrival time is modeled as a function of time of occurrence of last tweet. This model tends to predict small inter-arrival times yielding a huge number of points. We limit the number of predicted points

method	ARMSE	PRMSE
GPLIN	20.60±22.01⋆	1279.78±903.90⋆
HPP	21.85±22.82⋆	431.4±96.5⋆
HP	15.94±18.20	363.70±59.01⋆
LGCP	13.31±14.28	261.26±92.97⋆
LGCP Pooled	19.18±20.36⋆	183.25±102.20⋆
LGCPTXT	15.52±18.79	154.05±115.70

Table 1: ARMSE and PRMSE between the true event times and the predicted event times expressed in minutes (lower is better) over the 114 Ferguson rumours, showing mean ± std. dev. Key ⋆ denotes significantly worse than LGCPTXT method according to one-sided Wilcoxon signed rank test ($p < 0.05$). In case of ARMSE, LGCP is not significantly better than LGCP TXT according to Wilcoxon test.

to 1000 (above the maximum count yielded by any rumour from our dataset), thus reducing the error from this method.

We also compare against Hawkes Process (HP) (Yang and Zha, 2013), a self exciting point process where an occurrence of a tweet increases the probability of tweets arriving soon afterwards. We consider a univariate Hawkes process where the intensity function is modeled as $\lambda_i(t) = \mu + \sum_{t^i_j < t} k_{time}(t^i_j, t)$. The kernel parameters and μ are learnt by maximizing the likelihood. We apply the importance sampling algorithm discussed in Algorithm 1 for generating arrival times for Hawkess process. We consider this baseline only in the single-task setting, where reference rumours are not considered.

LGCP settings In the case of LGCP, the model parameters of the intensity function associated with a rumour are learnt from the observed inter-arrival times from that rumour alone. LGCP Pooled and LGCPTXT consider a different setting where this is learnt additionally using the inter-arrival times of all other rumours observed over the entire two hour life-span.

Results Table 1 reports the results of predicting arrival times of tweets in the second hour of the rumour lifecycle. In terms of ARMSE, LGCP is the best method, performing better than LGCP-TXT (though not statistically significantly) and outperforming other approaches. However, this metric does not penalize for the wrong number of predicted arrival times. Figure 1b depicts an example rumour, where LGCP greatly overestimates the number of points in the interval of interest. Here, the three points from the ground truth (denoted by black crosses) and the initial three points predicted by the LGCP model (denoted by red pluses), happen to lie very close, yielding a low ARMSE error. However, LGCP predicts a large number of arrivals in this interval making it a bad model compared to LGCPTXT which predicts only four points (denoted by blue dots). ARMSE fails to capture this and hence we use PRMSE. Note that Hawkes Process is performing worse than the LGCP approach.

According to PRMSE, LGCPTXT is the most successful method, significantly outperforming all other according to Wilcoxon signed rank test. Figure 1a depicts the behavior of LGCP and LGCP-TXT on rumour 39 with a larger number of points from the ground truth. Here, LGCPTXT predicts relatively less number of arrivals than LGCP. The performance of Hawkes Process is again worse than the LGCP approach. The self excitory nature of Hawkes process may not be appropriate for this dataset and setting, where in the second hour the number of points tends to decrease as time passes.

We also note, that GPLIN performs very poorly according to PRMSE. This is because the inter-arrival times predicted by GPLIN for several rumours become smaller as time grows resulting in a large number of arrival times.

6 Conclusions

This paper introduced the log-Gaussian Cox processes for the problem of predicting the inter-arrival times of tweets. We showed how text from posts helps to achieve significant improvements. Evaluation on a set of rumours from Ferguson riots showed efficacy of our methods comparing to baselines. The proposed approaches are generalizable to problems other than rumours, e.g. disaster management and advertisement campaigns.

Acknowledgments

Work partially supported by the European Union under grant agreement No. 611233 PHEME.

References

Christian Doerr, Norbert Blenn, and Piet Van Mieghem. 2013. Lognormal infection times of online information spread. *PLOS ONE*, 8.

J. Esteban, A. Ortega, S. McPherson, and M. Sathiamoorthy. 2012. Analysis of Twitter Traffic based on Renewal Densities. *ArXiv e-prints*.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Chapman and Hall/CRC.

Roberto Gonzalez, Alfonso Muñoz, José Alberto Hernández, and Ruben Cuevas. 2014. On the tweet arrival process at twitter: analysis and applications. *Trans. Emerging Telecommunications Technologies*, 25(2):273–282.

S. H. Lee, M. M. Crawford, and J. R. Wilson. 1991. Modeling and simulation of a nonhomogeneous poisson process having cyclic behavior. *Communications in Statistics Simulation, 20(2):777–809*.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Point process modelling of rumour dynamics in social media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, pages 518–523.

Jesper Møller and Anne Randi Syversveen. 1998. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, pages 451–482.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of NAACL*.

Rohan DW Perera, Sruthy Anand, KP Subbalakshmi, and R Chandramouli. 2010. Twitter analytics: Architecture, tools and analysis. In *Military Communications Conference, 2010-MILCOM 2010*, pages 2186–2191.

Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Sheldon M. Ross. 2010. *Introduction to Probability Models, Tenth Edition*. Academic Press, Inc., Orlando, FL, USA.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860.

Aleksandr Simma and Michael I. Jordan. 2010. Modeling events with cascades of poisson processes. In *UAI*, pages 546–555.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. 2013. Gpstuff: Bayesian modeling with gaussian processes. *J. Mach. Learn. Res.*, 14(1):1175–1179.

Shuang-Hong Yang and Hongyuan Zha. 2013. Mixture of mutually exciting processes for viral diffusion. In *ICML (2)*, volume 28 of *JMLR Proceedings*, pages 1–9.

Andrew Zammit-Mangion, Michael Dewar, Visakan Kadirkamanathan, and Guido Sanguinetti. 2012. Point process modelling of the afghan war diary. In *Proceedings of the National Academy of Sciences*, Vol. 109, No. 31, pages 12414–12419.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Early detection of rumors in social media from enquiry posts. In *International World Wide Web Conference Committee (IW3C2)*.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. In *AAAI Workshop on AI for Cities*.