Point Process Modelling of Rumour Dynamics in Social Media

Michal Lukasik[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, k.bontcheva}@shef.ac.uk
t.cohn@unimelb.edu.au

Abstract

Rumours on social media exhibit complex temporal patterns. This paper develops a model of rumour prevalence using a point process, namely a log-Gaussian Cox process, to infer an underlying continuous temporal probabilistic model of post frequencies. To generalize over different rumours, we present a multi-task learning method parametrized by the text in posts which allows data statistics to be shared between groups of similar rumours. Our experiments demonstrate that our model outperforms several strong baseline methods for rumour frequency prediction evaluated on tweets from the 2014 Ferguson riots.

1 Introduction

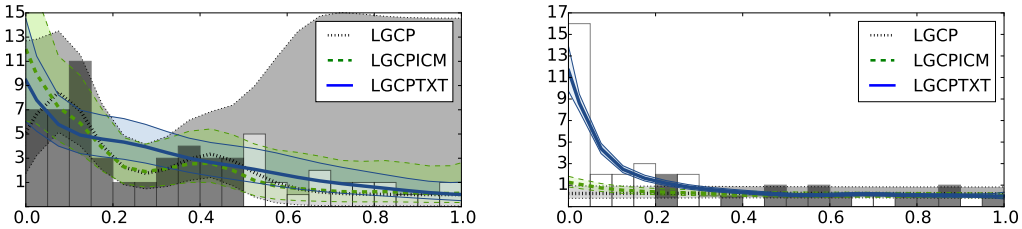
The ability to model rumour dynamics helps with identifying those, which, if not debunked early, will likely spread very fast. One such example is the false rumour of rioters breaking into McDonald's during the 2011 England riots. An effective early warning system of this kind is of interest to government bodies and news outlets, who struggle with monitoring and verifying social media posts during emergencies and social unrests. Another application of modelling rumour dynamics could be to predict the prevalence of a rumour throughout its lifespan, based on occasional spot checks by journalists.

The challenge comes from the observation that different rumours exhibit different trajectories. Figure 1 shows two example rumours from our dataset (see Section 3): online discussion of rumour #10 quickly drops away, whereas rumour #37 takes a lot longer to die out. Two characteristics can help determine if a rumour will continue to be discussed. One is the dynamics of post occurrences, e.g. if the frequency profile decays quickly, chances are it would not attract further attention. A second factor is text from the posts themselves, where phrases such as *not true*, *unconfirmed*, or *debunk* help users judge veracity and thus limit rumour spread (Zhao et al., 2015).

This paper considers the problem of modelling temporal frequency profiles of rumours by taking into account both the temporal and textual information. Since posts occur at continuous timestamps, and their density is typically a smooth function of time, we base our model on *point processes*, which have been shown to model well such data in epidemiology and conflict mapping (Brix and Diggle, 2001; Zammit-Mangion et al., 2012). This framework models count data in a continuous time through the underlying intensity of a Poisson distribution. The posterior distribution can then be used for several inference problems, e.g. to query the expected count of posts, or to find the probability of a count of posts occurring during an arbitrary time interval. We model frequency profiles using a log-Gaussian Cox process (Møller and Syversveen, 1998), a point process where the log-intensity of the Poisson distribution is modelled via a Gaussian Process (GP). GP is a non-parametric model which allows for powerful modelling of the underlying intensity function.

Modelling the frequency profile of a rumour based on posts is extremely challenging, since many rumours consist of only a small number of posts and exhibit complex patterns. To overcome this difficulty we propose *a multi-task learning approach*, where patterns are correlated across multiple rumours. In this way statistics over a larger training set are shared, enabling more reliable predictions for distant time periods, in which no posts from the target rumour have been observed. We demonstrate how text from observed posts can be used to weight influence across rumours. Using a set of Twitter rumours from the 2014 Ferguson unrest, we demonstrate that our models provide good

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 518–523,
Beijing, China, July 26-31, 2015.



(a) rumour #37 (b) rumour #10

Figure 1: Predicted frequency profiles for example rumours. Black bars denote training intervals, white bars denote test intervals. Dark-coloured lines correspond to mean predictions by the models, light shaded areas denote the 95% confidence interval, $\mu \pm 2\sigma$. This figure is best viewed in colour.

prediction of rumour popularity.

This paper makes the following contributions: 1. Introduces the problem of modelling rumour frequency profiles, and presents a method based on a log-Gaussian Cox process; 2. Incorporates multi-task learning to generalize across disparate rumours; and 3. Demonstrates how incorporating text into multi-task learning improves results.

2 Related Work

There have been several descriptive studies of rumours in social media, e.g. Procter et al. (2013) analyzed rumours in tweets about the 2011 London riots and showed that they follow similar lifecycles. Friggeri et al. (2014) showed how Facebook constitutes a rich source of rumours and conversation threads on the topic. However, none of these studies tried to model rumour dynamics.

The problem of modelling the temporal nature of social media explicitly has received little attention. The work most closely related modelled hash tag frequency time-series in Twitter using GP (Preotiuc-Pietro and Cohn, 2013). It made several simplifications, including discretising time and treating the problem of modelling counts as regression, which are both inappropriate. In contrast we take a more principled approach, using a point process. We use the proposed GP-based method as a baseline to demonstrate the benefit of using our approaches.

The log-Gaussian Cox process has been applied for disease and conflict mapping, e.g. Zammit-Mangion et al. (2012) developed a spatio-temporal model of conflict events in Afghanistan. In contrast here we deal with temporal text data, and model several correlated outputs rather than their single output. Related also is the extensive work done in spatio-temporal modelling of meme spread. One example is application of Hawkes processes (Yang and Zha, 2013), a probabilistic framework for modelling self-excitatory phenomena. However, these models were mainly used for network modelling rather than revealing complex temporal patterns, which may emerge only implicitly, and are more limited in the kinds of temporal patterns that may be represented.

3 Data & Problem

In this section we describe the data and we formalize the problem of modelling rumour popularity.

Data We use the Ferguson rumour data set (Zubiaga et al., 2015), consisting of tweets collected in August and September 2014 during the Ferguson unrest. It contains both source tweets and the conversational threads around these (where available). All source tweets are categorized as rumour vs non-rumour, other tweets from the same thread are assigned automatically as belonging to the same event as the source tweet. Since some rumours have few posts, we consider only those with at least 15 posts in the first hour as rumours of particular interest. This results in 114 rumours consisting of a total of 4098 tweets.

Problem Definition Let us consider a time interval $[0, l]$ of length l=2 hours, a set of n 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 bag of words text representation) and t_j^i is a timestamp describing post p_j^i, measured in time elapsed since the first post on rumour E_i.

Posts occur at different timestamps, yielding varying density of posts over time, which we are interested in estimating. To evaluate the predicted density for a given rumour E_i we leave out posts from a set of intervals $T_{te} = \{[s_k^i, e_k^i]\}_{k=1}^{K_i}$ (where s_k^i and e_k^i are respectively start and end points of

interval k for rumour i) and estimate performance at predicting counts in them by the trained model.

The problem is considered in supervised settings, where posts on this rumour outside of these intervals form the training set $E_i^{\mathrm{O}} = \{p_j^i : t_j^i \notin \bigcup_{k=1}^{K_i} [s_k^i, e_k^i]\}$. Let the number of elements in E_i^{O} be m_i^{O}. We also consider a domain adaptation setting, where additionally posts from other rumours are observed $R_i^O = R \setminus E_i$.

Two instantiations of this problem formulation are considered. The first is *interpolation*, where the test intervals are not ordered in any particular way. This corresponds to a situation, e.g., when a journalist analyses a rumour during short spot checks, but wants to know the prevalence of the rumour at other times, thus limiting the need for constant attention. The second formulation is that of *extrapolation*, where all observed posts occur before the test intervals. This corresponds to a scenario where the user seeks to predict the future profile of the rumour, e.g., to identify rumours that will attract further attention or wither away.

Although our focus here is on rumours, our model is more widely applicable. For example, one could use it to predict whether an advertisement campaign would be successful or how a political campaign would proceed.

4 Model

We consider a log-Gaussian Cox process (LGCP) (Møller and Syversveen, 1998), a generalization of inhomogeneous Poisson process. In LGCP the intensity function is assumed to be a stochastic process which varies over time. In fact, the intensity function $\lambda(t)$ is modelled using a latent function $f(t)$ sampled from a Gaussian process (Rasmussen and Williams, 2005), such that $\lambda(t) = \exp(f(t))$ (exponent ensures positivity). This provides a non-parametric approach to model the intensity function. The intensity function can be automatically learned from the data set and its complexity depends on the data points.

We model the occurrence of posts in a rumour E_i to follow log-Gaussian Cox process (LGCP) with intensity $\lambda_i(t)$, where $\lambda_i(t) = \exp(f_i(t))$. We associate a distinct intensity function with each rumour as they have varying temporal profiles. LGCP models the likelihood that a single tweet occurs at time t in the interval $[s, t]$ for a rumour E_i given the latent function $f_i(t)$ as

$$p(y = 1|f_i) = \exp(f_i(t)) \exp(-\int_s^t \exp(f_i(u))du).$$

Then, the likelihood of posts E_i^{O} in time interval T given a latent function f_i can be obtained as

$$p(E_i^{\mathrm{O}}|f_i) = \exp\left(-\int_{T-T_{te}} \exp\left(f_i(u)\right) du + \sum_{j=1}^{m_i^{\mathrm{O}}} f_i(t_j^i)\right) \quad (1)$$

The likelihood of posts in the rumour data is obtained by taking the product of the likelihoods over individual rumours. The likelihood (1) is commonly approximated by considering sub-regions of T and assuming constant intensities in sub-regions of T (Møller and Syversveen, 1998; Vanhatalo et al., 2013) to overcome computational difficulties arising due to integration. Following this, we approximate the likelihood as $p(E_i^{\mathrm{O}}|f_i) = \prod_{s=1}^{S} \mathrm{Poisson}(y_s \mid l_s \exp\left(f_i(\dot{t}_s)\right))$. Here, time is divided into S intervals indexed by s, $\dot{t}_s$ is the centre of the s^{th} interval, l_s is the length of the s^{th} interval and y_s is number of tweets posted during this interval.

The latent function f is modelled via a Gaussian process (GP) (Rasmussen and Williams, 2005): $f(t) \sim \mathcal{GP}(m(t), k(t, t'))$, where m is the mean function (equal 0) and k is the kernel specifying how outputs covary as a function of the inputs. We use a Radial Basis Function (RBF) kernel, $k(t, t') = a \exp(-(t - t')^2/l)$, where lengthscale l controls the extent to which nearby points influence one another and a controls the scale of the function.

The distribution of the posterior $p(f_i(t)|E_i^{\mathrm{O}})$ at an arbitrary timestamp t is calculated based on the specified prior and the Poisson likelihood. It is intractable and approximation techniques are required. There exist various methods to deal with calculating the posterior; here we use the Laplace approximation, where the posterior is approximated by a Gaussian distribution based on the first 2 moments. For more details about the model and inference we refer the reader to (Rasmussen and Williams, 2005). The predictive distribution over time t_* is obtained using the approximated posterior. This predictive distribution is then used to obtain the intensity function value at the point t_*:

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

The predictive distribution over counts at a particular time interval of length w with a mid-point t_* for rumour E_i is Poisson distributed with rate $w\lambda_i(t_*|E_i^{\mathrm{O}})$.

Multi-task learning and incorporating text In order to exploit similarities across rumours we propose a multi-task approach where each rumour represents a task. We consider two approaches.

First, we employ a multiple output GP based on the Intrinsic Coregionalization Model (ICM) (Álvarez et al., 2012). It is a method which has been successfully applied to a range of NLP tasks (Beck et al., 2014; Cohn and Specia, 2013). ICM parametrizes the kernel by a matrix representing similarities between pairs of tasks. We expect it to find correlations between rumours exhibiting similar temporal patterns. The kernel takes the form

$$k_{\text{ICM}}((t,i),(t',i')) = k_{time}(t,t')B_{i,i'},$$

where B is a square coregionalization matrix (rank 1, $B = \boldsymbol{\kappa} I + \boldsymbol{v}\boldsymbol{v}^T$), i and i' denote the tasks of the two inputs, k_{time} is a kernel for comparing inputs t and t' (here RBF) and $\boldsymbol{\kappa}$ is a vector of values modulating the extent of each task independence.

In a second approach, we parametrize the inter-task similarity measures by incorporating text of the posts. The full multi-task kernel takes form

$$k_{\text{TXT}}((t,i),(t',i')) = k_{time}(t,t') \times k_{text}\Bigg(\sum_{p_j^i \in E_i^O} \mathbf{x}_j^i, \sum_{p_j^{i'} \in E_{i'}^O} \mathbf{x}_j^{i'}\Bigg).$$

We compare text vectors using cosine similarity, $k_{text}(\mathbf{x},\mathbf{y}) = b + c\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$, where the hyperparameters $b > 0$ and $c > 0$ modulate between text similarity and a global constant similarity. We also consider combining both multi-task kernels, yielding $k_{\text{ICM+TXT}} = k_{\text{ICM}} + k_{\text{TXT}}$.

Optimization All hyperparameters are optimized by maximizing the marginal likelihood of the data $L(E_i^O|\theta)$, where $\theta = (a, l, \boldsymbol{\kappa}, \boldsymbol{v}, b, c)$ or a subset thereof, depending on the choice of kernel.

5 Experimental Setup

Evaluation metric We use mean squared error (MSE) to measure the difference between true counts and predicted counts in the test intervals. Since probabilistic models (GP, LGCP) return distributions over possible outputs, we also evaluate them via the log-likelihood (LL) of the true counts under the returned distributions (respectively Gaussian and Poisson distribution).

Baselines We use the following baselines. The first is the Homogenous Poisson Process (HPP) trained on the training set of the rumour. We select its intensity λ using maximum likelihood estimate, which equals to the mean frequency of posts in the training intervals. The second baseline is Gaussian Process (GP) used for predicting hashtag frequencies in Twitter by Preotiuc-Pietro and Cohn (2013). Authors considered various kernels in their experiments, most notably periodic kernels. In our case it is not apparent that rumours exhibit periodic characteristics, as can be seen in Figure 1. We restrict our focus to RBF kernel and leave inspection of other kernels such as periodic ones for both GP and LGCP models for future. The third baseline is to always predict 0 posts in all intervals. The fourth baseline is tailored for the interpolation setting, and uses simple interpolation by averaging over the frequencies of the closest left and right intervals, or the frequency of the closest interval for test intervals on a boundary.

Data preprocessing In our experiments, we consider the first two hours of each rumour lifespan, which we split into 20 evenly spaced intervals. This way, our dataset consists in total of 2280 intervals. We iterate over rumours using a form of folded cross-validation, where in each iteration we exclude some (but not all) time intervals for a single target rumour. The excluded time intervals form the test set: either by selecting half at random (interpolation); or by taking only the second half for testing (extrapolation). To ameliorate the problems of data sparsity, we replace words with their Brown cluster ids, using 1000 clusters acquired on a large scale Twitter corpus (Owoputi et al., 2013).

The mean function for the underlying GP in LGCP methods is assumed to be 0, which results in intensity function to be around 1 in the absence of nearby observations. This prevents our method from predicting 0 counts in these regions. We add 1 to the counts in the intervals to deal with this problem as a preprocessing step. The original counts can be obtained by decrementing 1 from the predicted counts. Instead, one could use a GP with a non-zero mean function and learn the mean function, a more elegant way of approaching this problem, which we leave for future work.

6 Experiments

The left columns of Table 1 report the results for the extrapolation experiments, showing the mean and variance of results across the 114 rumours. According to log likelihood evaluation metric, GP is the worst from the probabilistic ap-

	Extrapolation		Interpolation	
	MSE	LL	MSE	LL
HPP	7.14±10.1⋆	-23.5±10.1⋆	7.66±7.55⋆	-25.8±11.0⋆
GP	4.58±11.0⋆	-34.6±8.78⋆	6.13±6.57⋆	-90.1±198 ⋆
Interpolate	4.90±13.1⋆	-	5.29±6.06⋆	-
0	2.76±7.81⋆	-	7.65±11.0⋆	-
LGCP	3.44±9.99⋆	-15.8±11.6†⋆	6.01±6.29⋆	-21.0±8.77†⋆
LGCP ICM	2.46±7.82†⋆	-14.8±11.2†⋆	8.59±19.9⋆	-20.7±9.87†⋆
LGCP TXT	2.32±7.06†	-14.7±9.12†	3.66±5.67†	-16.9±5.91†
LGCP ICM+TXT	2.31±7.80†	-14.6±10.8†	3.92±5.20†	-16.8±5.34†

Table 1: MSE between the true counts and the predicted counts (lower is better) and predictive log likelihood of the true counts from probabilistic models (higher is better) for test intervals over the 114 Ferguson rumours for extrapolation (left) and interpolation (right) settings, showing mean ± std. dev. Baselines are shown above the line, with LGCP models below. Key: † denotes significantly better than the best baseline; ⋆ denotes significantly worse than LGCP TXT, according to one-sided Wilcoxon signed rank test $p < 0.05$.

proaches. This is due to GP modelling a distribution with continuous support, which is inappropriate for modelling discrete counts. Changing the model from a GP to a better fitting to the modelling temporal count data LGCP gives a big improvement, even when a point estimate of the prediction is considered (MSE). The 0 baseline is very strong, since many rumours have comparatively little discussion in the second hour of their lifespan relative to the first hour. Incorporating information about other rumours helps outperform this method. ICM, TXT and ICM+TXT multi-task learning approaches achieve the best scores and significantly outperform all baselines. TXT turns out to be a good approach to multi-task learning and outperforms ICM. In Figure 1a we show an example rumour frequency profile for the extrapolation setting. TXT makes a lower error than LGCP and LGCPICM, both of which underestimate the counts in the second hour.

Next, we move to the interpolation setting. Unsurprisingly, Interpolate is the strongest baseline, and outperforms the raw LGCP method. Again, HPP and GP are outperformed by LGCP in terms of both MSE and LL. Considering the output distributions (LL) the difference in performance between the Poisson Process based approaches and GP is especially big, demonstrating how well the principled models handle uncertainty in the predictive distributions. As for the multi-task methods, we notice that text is particularly useful, with TXT achieving the highest MSE score out of all considered models. ICM turns out to be not very helpful in this setting. For example, ICM (just as LGCP) does not learn there should be a peak at the beginning of a rumour frequency profile depicted in Figure 1b. TXT manages to make a significantly smaller error by predicting a large posting frequency there. We also found, that for a few rumours ICM made a big error by predicting a high frequency at the start of a rumour lifespan when there was no such peak. We hypothesize ICM performs poorly because it is hard to learn correct correlations between frequency profiles when training intervals do not form continuous segments of significant sizes. ICM manages to learn correlations more properly in extrapolation setting, where the first hour is fully observed.

7 Conclusions

This paper introduced the problem of modelling frequency profiles of rumours in social media. We demonstrated that joint modelling of collective data over multiple rumours using multi-task learning resulted in more accurate models that are able to recognise and predict commonly occurring temporal patterns. We showed how text data from social media posts added important information about similarities between different rumours. Our method is generalizable to problems other than modelling rumour popularity, such as predicting success of advertisement campaigns.

8 Acknowledgments

We would like to thank Srijith P. K. for helpful comments. This work was funded by the PHEME FP7 project (grant No. 611233) and partially supported by the Australian Research Council.

References

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. 2012. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266.

The GPy authors. 2012–2015. GPy: A Gaussian process framework in Python. `http://github.com/SheffieldML/GPy`.

Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task Gaussian processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1798–1803.

Anders Brix and Peter J. Diggle. 2001. Spatiotemporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society Series B*, 63(4):823–841.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 32–42.

Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *International AAAI Conference on Weblogs and Social Media*.

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 *Proceedings of NAACL*, pages 380–390.

Daniel Preotiuc-Pietro and Trevor Cohn. 2013. A temporal model of text periodicities using Gaussian processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 977–988.

Rob Procter, Jeremy Crump, Susanne Karstedt, Alex Voss, and Marta Cantijoch. 2013. Reading the riots: What were the police doing on twitter? *Policing and society*, 23(4):413–436.

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

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. *Proceedings of the National Academy of Sciences of the United States of America*, 109(31):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*.