Visualization of Predictive Distributions for Discrete Spatial-Temporal Log Cox Processes approximated with MCMC

David Rohde*, Jonathon Corcoran**, Gentry White***, Ruth Huang**

* Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
** School of Geography, University of Queensland, Brisbane Australia
***Institute for Social Science Research, University of Queensland, Brisbane Australia

Abstract. An important aspect of decision support systems involves applying sophisticated and flexible statistical models to real datasets and communicating these results to decision makers in interpretable ways. An important class of problem is the modelling of incidence such as fire, disease etc. Models of incidence known as point processes or Cox processes are particularly challenging as they are 'doubly stochastic' i.e. obtaining the probability mass function of incidents requires two integrals to be evaluated. Existing approaches to the problem either use simple models that obtain predictions using plug-in point estimates and do not distinguish between Cox processes and density estimation but do use sophisticated 3D visualization for interpretation. Alternatively other work employs sophisticated non-parametric Bayesian Cox process models, but do not use visualization to render interpretable complex spatial temporal forecasts. The contribution here is to fill this gap by inferring predictive distributions of Gaussian-log Cox processes and rendering them using state of the art 3D visualization techniques. This requires performing inference on an approximation of the model on a discretized grid of large scale and adapting an existing spatial-diurnal kernel to the log Gaussian Cox process context.

Keywords: non-parametric Bayesian Inference, Markov chain Monte Carlo, Visualization, Spatial-Temporal Cox Processes, Geographical Information Systems

1 Introduction

There are considerable applications for applying pattern recognition techniques to exploring spatial-temporal datasets of incidents. Applications include modelling the occurrence of disease outbreaks, modelling the observations of new species, or the temporal occurrence of incidents such as coal mine disasters. Another pertinent example is the occurrence in space and in the hour of day (diurnal) of urban fires in Australia which we use as a case study in this paper. The goal of such a spatial-diurnal analysis can be used to evaluate the consequences

of different operational decisions of the fire service and to direct attention to preventative action. While a formal model may be employed for the occurrence of fire incidents, the available decisions and their utility under different hypothetical outcomes is usually not formalized. As such it can be preferable to use advanced three dimensional visualizations of the output using 3D rendering techniques such as the iso-surface, cut planes and volume rendering.

Statistical models of incidence are complex entities as they are doubly stochastic models, i.e. identifying the probability mass function of the count of incidence in a region involves computing two integrals instead of the usual one. These statistical models are known as point processes or Cox processes [5].

Existing research on forecasting point processes follows two main paths. The first path is to use simple statistical approaches based on kernel smoothing applied to a point process in essentially the same way as applied to problems of probability density estimation [7]. If a new spatial-diurnal kernel that is Gaussian in space and has a 24 hour period in time is adopted spatial-diurnal surface of the Cox process can be computed and visualized using techniques such as the 3D isosurface [3]. There are several shortcomings in adopting such a simple model. As predictions are obtained using a plug-in point estimate of the underlying intensity of the Cox process model uncertainty is ignored. The non-Bayesian method also does not provide principled means of bandwidth selection or for avoiding edge effects. Finally and perhaps most compelling kernel smoothing is just a simple function of the data which means that the computed surface is most reasonably seen as a summary of history rather than a forecast. A more sophisticated modelling approach may intelligently identified patterns in the data that are likely to continue into the future. Using a smoothed summary of history for forecasting can be particularly problematic in temporal models and simplifications such as only using the hour of day (i.e. diurnal time) component must be employed. However despite the statistical shortcomings of such an approach it is delivering something useful to decision makers in the form of comprehensible outputs related to operational concerns.

A second path uses sophisticated Bayesian models to obtain predictions, but usually on relatively simple problems and without sophisticated 3D visualization in order to make the implications of these complex mathematical objects apparent to the decision maker.

An important class of Bayesian non-parametric models of Cox processes is the Gaussian Cox Processes which models the intensity function as a non-negative function of a Gaussian process, usually an exponential function resulting in a log Gaussian Cox Process.

A general expression for the count C of incidents in a region of space and time R under a log Gaussian process is given by

$$C \sim \text{Poisson}\left(\iiint_R e^{\phi(x,y,t)} \mathrm{d}x\mathrm{d}y\mathrm{d}t\right)$$

where $\phi(x, y, t)$ is a Gaussian Process. A further restriction is that the distribution of counts for non-overlapping regions R and R' should be independent.

Fully Bayesian approaches to this problem are hampered by the presence of doubly intractable integrals [12]. That is the evaluation of the likelihood and as such the posterior itself contains an intractable integral, this makes applying Markov chain Monte Calro (MCMC) algorithms difficult. Two possible solutions have been proposed, the first makes use of the fact that MCMC algorithms have been found to handle doubly intractable integrals for the special cases where the model can be sampled from [11] [13]. While sampling from a log-Gaussian Cox process is not possible sampling from a modified model the sigmoidal Gaussian Cox Process is possible resulting in the first MCMC algorithm to handle Gaussian Cox Processes [1], i.e. the first MCMC algorithm with the stationary distribution of the posterior of a Gaussian Cox process.

Although this approach appears to offer many advantages and allows MCMC inference to be performed on a very appealing model there are some drawbacks. The first is computational cost, this algorithm performs a matrix inversion on a matrix that includes all observed data and additionally latent data incorporated into the model, this is likely to be a considerable burden. Secondly the main advantage of this approach is to avoid imposing the model on a discrete grid as in [9] however one of our goals is to visualize the model using 3D graphical tools such as Mayavi [14] which itself uses a discrete scalar field. Thirdly there are comforting results that given a sufficiently fine grid the discrete model will converge to a continuous model [16]. Finally the discrete model has a much longer history being applied to real problems including to problems of fire incidence [10].

There exist Bayesian non-parametric models of Cox processes which do not rely on the Gaussian process for the intensity function, but rather use a mixture model for the intensity function such as [8] which uses a Dirichlet process mixture of Beta distributions. This has the advantage of allowing a relatively standard MCMC sampler, but the disadvantage of losing the intepretability of the Gaussian process which has been extensively used in both spatial statistics under the name of Krigging [4] and in machine learning [15]. Indeed one of our goals here is to incorporate the spatial diurnal kernel developed in [3] to the log Gaussian Cox process context, it is very unclear if a Cox process with periodicity could be formulated in a mixture model framework.

2 Discretised Log Gaussian Cox Process Model

An outline of the model is developed first in space only and then in space and diurnal time.

2.1 Spatial Model

The discretized spatial only model takes the following form

$$C_{i,j} \sim \text{Poisson}(e^{\phi_{i,j}}) \tag{1}$$

where ϕ is given a matrix variate normal distribution, or alternatively $\gamma = e^{\phi}$ is given a matrix variate log normal distribution. The purpose of the prior is to impose a correlation between elements of C that are close space. For our application the desirable dimension of C and consequently ϕ is $I \times J$ which in the current context is 240×240 or 57600 parameters for ϕ. While this is large it remains manageable, on the other hand the covariance matrix for the normal distribution over ϕ is this size squared which is not easily manageable. Fortunately both for conceptual and computational reasons a matrix variate normal distribution can be employed instead. While the matrix variate distribution can be interpreted as a special case of the multivariate normal distribution in that

$$\text{vec}(\phi) \sim \mathcal{N}_{I \times J}(\text{vec}(\mu), \Omega \otimes \Gamma). \tag{2}$$

here $\text{vec}(\phi)$ denotes that the $I \times J$ array or matrix of ϕ is converted to a single vector $\text{vec}(\phi) = [\phi_{1,1}, ..., \phi_{1,J}, ..., \phi_{I,1}, ..., \phi_{I,J}]^T$, Ω is an $I \times I$ matrix and Γ is a $J \times J$ matrix and $\otimes$ is the Kronecker tensor product, note that for many applications computing $\Omega \otimes \Gamma$ would require allocating very large amounts of memory for a highly redundant array.

It is instructive to consider the covariance between $\phi_{i,j}$ and $\phi_{i',j'}$ which is $\Omega_{i,i'}\Gamma_{j,j'}$, in the context of a Gaussian process a covariance function would specify a covariance between two points (i, j) and (i', j'), in practice there are a number of popular forms that result in close points having high covariance such as exponential and the more general Matern family [4] [6] . If an exponential covariance function is adopted then $\text{Cov}(\phi_{i,j}, \phi_{i',j'}) = \exp(-(\frac{i-i'}{I\sigma})^2 - \frac{j-j'}{J\sigma})^2)$, it is easily seen that this can be obtained by letting $\Omega_{i,i'} = \exp(-(\frac{i-i'}{I\sigma})^2)$ and $\Gamma_{j,j'} = \exp(-(\frac{j-j'}{J\sigma})^2)$.

The main advantage of employing a matrix normal rather than multivariate normal prior is that by factorizing the covariance matrix $\Sigma = \Omega \otimes \Gamma$ it is possible to evaluate the prior and therefore the posterior using reasonable amounts of memory, by using the following expression for the matrix variate normal distribution.

$$P(\phi|\mu, \Omega, \Sigma) = \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\Omega^{-1}(\phi - \mu)^T\Sigma^{-1}(\phi - \mu)\right]\right)}{(2\pi)^{\frac{1}{2}IJ}|\Omega|^{I/2}|\Sigma|^{J/2}} \tag{3}$$

Alternatively $\lambda = e^{\phi}$ can be given a matrix variate log normal distribution

$$P(\lambda|\mu, \Omega, \Sigma) = f_\lambda(\lambda) = \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\Omega^{-1}(\log(\lambda) - \mu)^T\Sigma^{-1}(\log(\lambda) - \mu)\right]\right)}{(2\pi)^{\frac{1}{2}IJ}|\Omega|^{I/2}|\Sigma|^{J/2}\prod_{i=1..I}\prod_{j=1..J}\lambda_{i,j}} \tag{4}$$

By using this expression we are able to formulate a complete expression for the posterior

$$P(\lambda|C) = f_\lambda(\lambda) \prod_{i=1..I} \prod_{j=1..J} \text{Poiss}(C_{i,j}|\lambda_{i,j}) \tag{5}$$

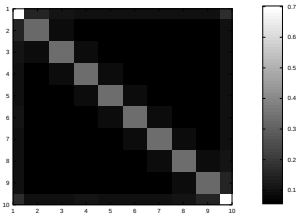

Fig. 1. Approximation of $E[\lambda|C] = E[e^\phi|C]$ depicted graphically, the approximation is computed using multiplicative random walk Metropolis algorithm.

Using this expression it is possible to use MCMC algorithms to generate samples from $P(e^\phi|C)$. A simple random walk Metropolis algorithm can be used, or because e^ϕ is non-negative a multiplicative Metropolis hastings algorithm can also be used with multivariate log normal proposals with low variance and expectation of 1. The advantage of this procedure is that the proposals like the target have non-negative support. This method was applied to the model where $I = J = 10$ and where C is equal to the identity matrix and $\text{Cov}(\phi_{i,j}, \phi_{i',j'}) = \exp(-(i-i')^2 - (j-j')^2)$. A multiplicative Metropolis proposal was used with an expectation of 1 and a variance of 0.25 the chain was run for 2 million iterations, with thinning so that only 1 in 100 samples were retained this resulted in a fraction of 0.08 samples being accepted (rather than the optimal 0.25). A visual display of $E[e^\phi|C]$ is shown in Figure 1. This output shows some pleasing features in that the highest values are where counts have been observed on the diagonal, and it appears that the covariance is operating correctly as cells adjacent to the diagonal also have an increased value. Very similar results are obtained by applying a standard random walk Metropolis algorithm to ϕ. This distribution can be obtained by multiplying the posterior by the Jacobian, and in our experience resulted in a slightly more efficient algorithm.

$$P(\phi|C) = f_\phi(\phi) = f_\lambda(e^\phi) \left|J(e^\phi, \phi)\right| = f_\lambda(e^\phi) \prod_{i=1..I} \prod_{j=1..J} e^{\phi_{i,j}} \tag{6}$$

where $J(\cdot)$ is the Jacobian.

2.2 Spatial-Diurnal Model

The discretized spatial-diurnal model has the following form

$$C_{i,j,k} \sim \text{Poisson}(e^{\phi_{i,j,k}}) \tag{7}$$

where ϕ is given a 'tensor' variate normal distribution, or alternatively $\gamma = e^{\phi}$ is given a 'tensor' variate log normal distribution. The purpose of the prior is to impose a correlation between elements of C that are close space. For our application the desirable dimension of C and consequently ϕ is $I \times J \times K$ which in the current context is $240 \times 240 \times 24$ or approximately 1.3 million parameters for ϕ. In this model like the previous model, i, j and k index cells with different spatial co-ordinates and j index the diurnal or hourly of day component of time. The covariance between two points is given by $\text{Cov}(\phi_{i,j,k}, \phi_{i',j',k'}) = \exp(-(\frac{i-i'}{I\sigma})^2 - (\frac{j-j'}{J\sigma})^2 - \frac{c^2}{\sigma^2}\Delta(k,k'))$, where $\Delta(k,k') = -\frac{1}{2}\cos(\frac{2\pi(k-k')}{24K}) + \frac{1}{2}$, and the parameter c is a constant which converts a distance in time of 12 hours to a distance in meters, note that $\Delta(k,k'+12K) = 1$, so that $\text{Cov}(\phi_{i,j,k}, \phi_{i,j,k+12K}) = \exp(-\frac{c^2}{\sigma^2})$, an equivalent covariance can be achieved purely at a distance in this case in the i direction as $\text{Cov}(\phi_{i,j,k}, \phi_{i+cI,j,k}) = \exp(-\frac{c^2}{\sigma^2})$, so that

$$\text{vec}(\phi) \sim \mathcal{N}_{I\times J\times K}(\text{vec}(\mu), \Omega \otimes \Gamma \otimes \Psi). \tag{8}$$

where, $\Omega_{i,i'} = \exp(-(\frac{i-i'}{I\sigma})^2)$ and $\Gamma_{j,j'} = \exp(-(\frac{j-j'}{J\sigma})^2)$, and $\Psi_{k,k'} = \exp(-\frac{1}{2}\cos(\frac{2\pi(k-k')}{24K}) + \frac{1}{2})$.

While a tensor variate normal distribution can be expressed using tensor products, it can also be expressed using linear algebra and the Kronecker tensor product available in most scientific programming languages such as GNU Octave [2] as used here. This involves reshaping ϕ and μ which have $I \times J \times K$ to ϕ' and μ' to dimensions $I \times JK$. The distribution over ϕ' can then be expressed as

$$P(\phi'|\mu', \Omega, \Sigma, \Psi) = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\text{tr}\left[\Omega^{-1}(\phi'-\mu')^T(\Sigma^{-1}\otimes\Psi^{-1})(\phi'-\mu')\right]\right)}{(2\pi)^{\frac{1}{2}IJK}|\Omega|^{I/2}|\Sigma\otimes\Psi|^{JK/2}}. \tag{9}$$

The full posterior is then $P(C'|\phi')P(\phi'|\mu', \Omega, \Sigma, \Psi)$, where C' is also reshaped from a tensor to a matrix with dimensions $I \times JK$. The advantage of this solution is that a compact representation of the posterior is available which can be written in standard scientific programming languages (in this case GNU Octave). The disadvantage is that $\Sigma \otimes \Psi$ will require a large amount of memory to be allocated and this is in principle avoidable.

3 Results and Discussion

A log Gaussian Cox Process was run on a discrete $I \times J \times K$ grid where $I = J = 240$ and $K = 24$ with the following covariance function. The dataset this was

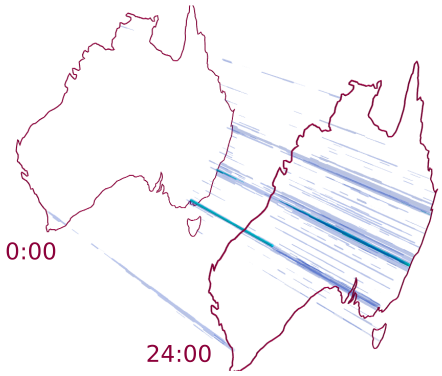

Fig. 2. Iso surfaces of $E[\lambda|C] = E[e^{\phi}|C]$ for a spatial-diurnal Cox process on a discretized $240 \times 240 \times 24$ grid, with the expectation computed with MCMC methods.

applied to was the occurrence in time and hour of day of malicious hoax calls within metropolitan Australia.

$$\begin{aligned}\mathrm{Cov}(\phi_{i,j,k}, \phi_{i',j',k'}) = \\ \exp\{-(x_{\max} - x_{\min})((i-i')/(I\sigma))^2 - (y_{\max} - y_{\min})((j-j')/(J\sigma))^2 \\ + \frac{c}{\sigma^2}(\frac{1}{2}\cos(2\pi(k-k')/(24K)) - \frac{1}{2})\}\end{aligned} \tag{10}$$

where $x_{\max} - x_{\min} = 4000$ km and $y_{\max} - y_{\min} = 3600$ km and $\sigma = 5$.

The initial value of the Markov chain was obtained by setting $\phi^{(0)}_{i,j,k} = \log(C_{i,j,k} + 1)$, the MCMC algorithm then proceeds using the random walk Metropolis algorithm for 5000 iterations with a burn in of 1000, and with thinning such that only every tenth sample is used.

A 3D iso surface rendered with the Mayavi library is shown in Fig 2. Incidence of malicious hoax calls are visible in time and hour of day, with large numbers of incidents in metropolitan regions such as Sydney and Melbourne and with evident trends of higher and lower incidence through different parts of the day.

This surface is qualitatively similar to the spatial diurnal iso surfaces of Cox processes used in previous studies such as [3]. As such the visual output of this model qualitatively has the same appealing features that are useful in operational context. The work here improves the state of the art by applying the spatial diurnal Cox process a more complex model that previously considered and demonstrating 3D visualization techniques on this. Several aspects of this problem deserve deeper consideration including statistical modelling issues including inference for the bandwidth and more complex models for time i.e. that incorporate more than the diurnal component. Computational issues are also of interest, it seems likely that more powerful MCMC algorithms may be required if inference is applied to a sufficiently fine grid as the approach described for setting the initial value of the chain will become less effective on a smaller grid.

In doing this work we sometimes encountered numerical problems where the covariance matricies where either singular or were not positive definite to working precision, more work is needed in order to work out when this occurs and how to avoid this. Finally the fact that a full posterior distribution is available should enrich visualization possibilities. We are considering these issues in future work.

References

1. R. P. Adams, I. Murray, and D. J. C. MacKay. Tractable nonparametric Bayesian inference in poisson processes with gaussian process intensities. In *In Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, Montreal, Quebec, 2009.
2. John W. Eatonand David Bateman and Soren Hauberg. *GNU Octave Manual Version 3*. Network Theory Limited, 2008.
3. C. Brunsdon, J. Corcoran, and G. Higgs. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, 31(1):52–75, 2007.
4. J. Chiles and P. Delfiner. *Geostatistics*. Wiley, New York, 1999.
5. D. R. Cox. Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society. Series B*, 17(2):129–164, 1955.
6. N. A. C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1991.
7. P. J. Diggle. A kernel method for smoothing point process data. *Applied Statistics*, 34:138–47, 1985.
8. A. Kottas and B Sansó. Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137:31513163, 2009.
9. J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451482, 1998.
10. J. Moller and F. B. Vej. Structured spatio-temporal shot-noise cox point process models, with a view to modelling forest fires. *Scandinavian Journal of Statistics*, 37(1), 2010.
11. Jesper Moller, Anthony N. Pettitt, Robert W. Reeves, and Kasper K. Berthelsen. An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, June 2006.
12. Iain Murray and Zoubin Ghahramani. Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 392–399, Arlington, Virginia, 2004. AUAI Press.
13. Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.
14. P. Ramachandran and G. Varoquaux. Mayavi: 3D Visualization of Scientific Data. *Computing in Science & Engineering*, 13(2):40–51, 2011.
15. C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.
16. R Waagepetersen. Convergence of posteriors for discretized log gaussian cox processes. *Statistics & Probability Letters*, 66:229–235, 2003.