Interpretable Deep Generative Spatio-Temporal Point Processes

Shixiang Zhu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
shixiang.zhu@gatech.edu

Shuang Li
Department of Statistics
Harvard University
Cambridge, MA 02138
shuangli@fas.harvard.edu

Zhigang Peng
School of Earth and Atmospheric Sciences
Georgia Institute of Technology
Atlanta, GA 30332
zpeng@gatech.edu

Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
yao.xie@isye.gatech.edu

Abstract

We present a novel Neural Embedding Spatio-Temporal (NEST) point process model for spatio-temporal discrete event data and develop an efficient imitation learning (a type of reinforcement learning) based approach for model fitting. Despite the rapid development of one-dimensional temporal point processes for discrete event data, the study of spatial-temporal aspects of such data is relatively scarce. Our model captures complex spatio-temporal dependence between discrete events by carefully design a mixture of heterogeneous Gaussian diffusion kernels, whose parameters are parameterized by neural networks. This is the key that our model can capture intricate spatial dependence patterns and yet still lead to interpretable results as we examine maps of Gaussian diffusion kernel parameters. Furthermore, the likelihood function under our model enjoys tractable expression due to Gaussian kernel parameterization. Experiments based on real data show our method's good performance relative to the state-of-the-art and the good interpretability of NEST's result.

1 Introduction

Spatio-temporal event data has become ubiquitous, emerging from various applications. Studying generative models for discrete events data has become a hot area in machine learning and statistics: it reveals of pattern in the data, helps us to understand the data dynamic and information diffusion, as well as serves as an important step to enable subsequent machine learning tasks. Point process models (see [17] for an overview) have become a standard choice for generative models of discrete event data. In particular, the self and mutual exciting processes, also known as the Hawkes processes, are popular since they can capture past events' influence on future events over time, space, and networks.

AI for Earth Sciences Workshop at NeurIPS 2020.

Despite the rapid development of one-dimensional temporal point processes models for discrete event data, the study focusing on *spatial-temporal* aspects of such data is relatively scarce. The original works of [15, 16] develop the so-called ETAS model, which is still widely used, suggesting an exponential decaying diffusion kernel function. This model captures the seismic activities' mechanism and is convenient to fit, as the kernel function is homogeneous at all locations with the same oval shape. However, these classical models for spatio-temporal event data (usually statistical models in nature) tend to make strong parametric assumptions on the conditional intensity.

However, in specific scenarios, the simplifying spatio-temporal model based on ETAS may lack flexibility. It does not capture the anisotropic spatial influence and cannot capture the complex spatial dependence structure (see a motivating example in Appendix A). On the other hand, when developing spatio-temporal models, we typically want to generate some statistical interpretations (e.g., temporal correlation, spatial correlation), which may not be easily derived from a complete neural network model. Thus, generative model based on specifying conditional intensity of point process models is a popular approach. For example, recent works [2, 12, 11, 18, 20, 19, 24] has achieved many successes in modeling temporal event data (some with marks) which are correlated in time. It remains an open question on extending this type of approach to include the spatio-temporal point processes. One challenge is how to address the computational challenge associated with evaluating the log-likelihood function. This can be intractable for the general model without a carefully crafted structure since it requires the integration of the conditional intensity function in a continuous spatial and time-space.

In this paper, we present a novel point-process based model for spatio-temporal discrete event data. Our proposed NEST model tackles flexible representation for complex spatial dependence, interpretability, and computational efficiency, through meticulously designed neural networks with embedding capturing spatial information. We generalize the idea of using a Gaussian diffusion kernel to model spatial correlation by introducing the more flexible heterogeneous mixture of Gaussian diffusion kernels with shifts, rotations, and non-isotropic shapes. Such a model can still be efficiently represented using a handful of parameters (compared with a full neural network model such as convolutional neural networks (CNN) over space). The Gaussian diffusion kernels are parameterized by neural networks, which allows the kernels to vary continuously over locations. This is the key that our model can capture intricate spatial dependence patterns and yet still lead to interpretable results as we examine maps of Gaussian diffusion kernel parameters. As shown in Figure 6 in Appendix A, our model is able to represent arbitrary diffusion shape at different locations in contrast to ETAS developed by [14, 15, 16]. Moreover, the likelihood function under our model enjoys tractable expression due to Gaussian kernel parameterization. Experiments based on real data show our method's good performance relative to the state-of-the-art and NEST results' interpretability.

2 Proposed model

To capture the complex and heterogenous spatial dependence in discrete events, we present a novel *continuous-time and continuous-space* point process model (see Appendix B for a basic introduction in point processes), called the Neural Embedding Spatio-Temporal (NEST) model. The NEST uses the flexible neural network structure to represent the conditional intensity's spatial heterogeneity while retaining interpretability as a semi-parametric statistical model.

Spatially heterogeneous Gaussian diffusion kernel We start by specifying the conditional probability of the point process model, as it will uniquely specify the joint distribution of a sequence of events. First, to obtain a similar interpretation as the ETAS model [15], we start from a similar parametric form for the conditional intensity function

$$\lambda^*(t, s) = \lambda_0 + \sum_{j: t_j < t} \nu(t, t_j, s, s_j), \tag{1}$$

where $\lambda_0 > 0$ is a constant background rate, ν is the kernel function that captures the influence of the past events $\mathcal{H}_t$. The form of the kernel function ν determines the profile of the spatio-temporal dependence of events.

We assume the kernel function takes the form of a standard Gaussian diffusion kernel over space and decays exponential over time. To enhance the spatial expressiveness, we adopt a mixture of generalized Gaussian diffusion kernels, which is location dependent. Thus, it can capture more complicated spatio-nonhomogeneous structure. Given all past events $\mathcal{H}_t$, we define

$$\nu(t, t', s, s') = \sum_{k=1}^{K} \phi_{s'}^{(k)} \cdot g(t, t', s, s' | \Sigma_{s'}^{(k)}, \mu_{s'}^{(k)}), \qquad \forall t' < t, s \in \mathcal{S}, \tag{2}$$

where $\{\mu_{s'}^{(k)}, \Sigma_{s'}^{(k)}\}$ are the mean and covariance matrix parameters (which we will specify later), K is the hyper-parameter that defines the number of components of the Gaussian mixture; $\phi_{s'}^{(k)} : \mathcal{S} \to \mathbb{R}$ (form specified later) is the weight for the k-th Gaussian component that satisfies $\sum_{k=1}^{K} \phi_{s'}^{(k)} = 1$, $\forall s' \in \mathcal{S}$. In the following discussions, we omit the superscript k for the notational simplicity.

Now each Gaussian diffusion kernel is defined as

$$g(t, t', s, s' | \Sigma_{s'}, \mu_{s'}) = \frac{Ce^{-\beta(t-t')}}{2\pi\sqrt{|\Sigma_{s'}|}(t-t')} \cdot \exp\left\{ -\frac{(s - s' - \mu_{s'})^T \Sigma_{s'}^{-1} (s - s' - \mu_{s'})}{2(t-t')} \right\}, \tag{3}$$

where $\beta > 0$ controls the temporal decay rate, $C > 0$ is constant to control the magnitude, $\mu_s = [\mu_x(s), \mu_y(s)]^T$, and Σ_s denote the mean and covariance parameters of the diffusion kernel (which may vary over time t and "source" locations $s \in \mathcal{S}$); $|\cdot|$ denotes the determinant of a covariance matrix; Σ_s is defined as a positive semi-definite matrix

$$\Sigma_s = \begin{pmatrix} \sigma_x^2(s) & \rho(s)\sigma_x(s)\sigma_y(s) \\ \rho(s)\sigma_x(s)\sigma_y(s) & \sigma_y^2(s) \end{pmatrix}.$$

The parameters μ_s and Σ_s control the shift, rotation and shape of each Gaussian component. As shown in Figure 1, parameters $\sigma_x(s), \sigma_y(s), \rho(s)$ may vary according to the location s and jointly control the spatial structure of diffusion at s. The $\mu_x(s), \mu_y(s)$ define the offset of the center of the diffusion from the location s. Note that $\mu_x : \mathcal{S} \to \mathbb{R}$, $\mu_y : \mathcal{S} \to \mathbb{R}$, $\sigma_x : \mathcal{S} \to \mathbb{R}^+$, $\sigma_y : \mathcal{S} \to \mathbb{R}^+$, $\rho : \mathcal{S} \to (-1, 1)$ are the non-linear mappings that project location s to the parameters. To capture intricate spatial dependence, we represent such non-linear mappings from location to the parameters of Gaussian components (defined by (3)) using neural networks. An illustration for our neural network architecture has been shown in Figure 2, and the detailed description has been elaborated in Appendix C.

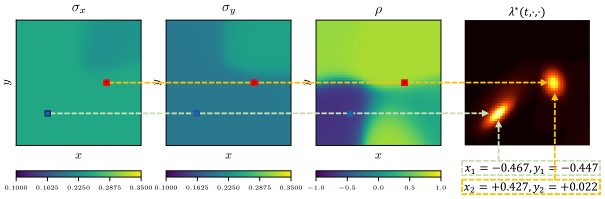

Figure 1: An example of kernel used in the NEST model: σ_x, σ_y, ρ defines a Gaussian component in the heterogeneous Gaussian diffusion kernel. The right hand side is the conditional intensity at time t, where two points occurred at location (x_1, y_1) and (x_2, y_2) have triggered the two diffusions (the bright spots) with different shapes.

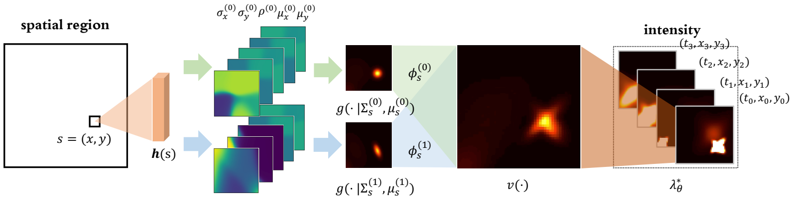
spatial region
$s = (x, y)$
$\boldsymbol{h}(s)$
$\sigma_x^{(0)} \sigma_y^{(0)} \rho^{(0)} \mu_x^{(0)} \mu_y^{(0)}$
$\sigma_x^{(1)} \sigma_y^{(1)} \rho^{(1)} \mu_x^{(1)} \mu_y^{(1)}$
$\phi_s^{(0)}$
$\phi_s^{(1)}$
$g(\cdot | \Sigma_s^{(0)}, \mu_s^{(0)})$
$g(\cdot | \Sigma_s^{(1)}, \mu_s^{(1)})$
$v(\cdot)$
intensity
(t_3, x_3, y_3)
(t_2, x_2, y_2)
(t_1, x_1, y_1)
(t_0, x_0, y_0)
λ_θ^*

Figure 2: An illustration for NEST's neural network architecture based on a mixture of heterogeneous Gaussian diffusion kernel. Note that each Gaussian kernel is specified by neural networks, which can be viewed as summarizing the latent embedding information from data.

Comparison with ETAS model In the standard ETAS model, the kernel function is defined as a *single component whose parameters do not vary over space and time*: the kernel function (2) is simplified to $\nu(t, t', s, s') = g(t, t', s, s' | \Sigma, 0)$, where the spatial and temporal parameters are invariant $\Sigma \equiv \text{diag}\{\sigma_x^2, \sigma_y^2\}$ and $\mu_s \equiv 0$. Compared with the standard Gaussian diffusion kernel used in ETAS, here we introduce additional parameters ρ, μ_x, μ_y that allows the diffusion to shift, rotate, or stretch in the space. An example of the comparison of the spatio-temporal kernels between ETAS and NEST models is presented in Figure 6, Appendix A.

3 Computationally efficient learning

The model parameters can be estimated via maximum likelihood estimate (MLE) since we have the explicit form of the conditional intensity function. Given a sequence of events $\boldsymbol{a} = \{a_0, a_1, \dots, a_n\}$ occurred on $(0, T] \times \mathcal{S}$ with length n, where $a_i = (t_i, s_i)$, the log-likelihood is given by

$$\ell(\theta) = \left(\sum_{i=1}^{n} \lambda^*_\theta(t_i, s_i) \right) - \int_0^T \int_{\mathcal{S}} \lambda^*_\theta(\tau, r) dr d\tau. \tag{4}$$

A crucial step to tackle the computational challenge is to evaluate the integral in (4). Here, we can obtain a closed-form expression for the likelihood function, using the following proposition. This can reduce the integral to an analytical form, which can be evaluated directly without numerical integral.

Proposition 1 (Integral of conditional intensity function)**.** *Given ordered event times* $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = T$*, for* $i = 0, \dots, n$*,*

$$\int_{t_i}^{t_{i+1}} \int_{\mathcal{S}} \lambda^*_\theta(\tau, r) dr d\tau = \lambda_0 (t_{i+1} - t_i)|\mathcal{S}| + (1 - \epsilon) \frac{C}{\beta} \sum_{j: t_j < t_i} C_j \left(e^{-\beta(t_i - t_j)} - e^{-\beta(t_{i+1} - t_j)} \right),$$

where $C_j = \sum_{k=1}^{K} \phi^{(k)}_{s_j} (\sigma^{(k)}_x(s_j) \sigma^{(k)}_y(s_j) / |\Sigma^{(k)}_{s_j}|^{1/2})$*, and the constant*

$$\epsilon = \max_{j: t_j < t_{i+1}} \frac{\int_{t_i}^{t_{i+1}} \int_{\mathcal{S}} g(\tau, t_j, r, s_j) dr d\tau}{\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^2} g(\tau, t_j, r, s_j) dr d\tau}.$$

Since spatially, the kernel g is a Gaussian concentrated around s, when $\mathcal{S}$ is chosen sufficiently large, and most events s_i locates in the relatively interior of $\mathcal{S}$, we can ignore the marginal effect and ϵ can become a number much smaller than 1. Due to the decreased activity in the region's edges, the boundary effect is usually negligible [16]. Define $t_0 = 0$ and $t_{n+1} = T$. Since $\int_0^T \int_{\mathcal{S}} \lambda^*_\theta(\tau, r) dr d\tau = \sum_{i=0}^{n+1} \int_{t_i}^{t_{i+1}} \int_{\mathcal{S}} \lambda^*_\theta(\tau, r) dr d\tau$, using Proposition 1 we can write down the integral in the log-likelihood function in closed-form expression. Finally, the optimal parameters trained the maximum-likelihood is thus obtained by $\hat{\theta} = \text{argmax}_\theta \log \ell(\theta)$. Due to the non-convex nature of this problem, we solve the problem by stochastic gradient descent. In addition to MLE, we also introduce a more flexible, imitation learning framework for model fitting as described in Appendix D. The major benefit of this approach is that it does not rely on the likelihood function model and thus is more robust to model misspecification.

4 Numerical results

We describe our experimental setting and the real data we have used in Appendix E. More simulation studies can be found in Appendix F. We first quantitatively compare our NEST+MLE and ETAS by evaluating Mean Squared Error (MSE) of one-step-ahead prediction. The results has been summarized in Figure 3, which shows that both our methods NEST+IL and NEST+MLE outperform the state-of-the-art (ETAS) regarding two metrics. In addition, we also show that our method has a competitive performance even without considering spatial information in contrast to RLPP.

Figure 3: MSE for five methods on two real data sets.

Data set	Random	ETAS	NEST+IL	NEST+MLE	RLPP
Robbery (space-time)	.6323	.1425	**.0503**	.0649	N/A
Seismic (space-time)	.2645	.0221	**.0119**	.0153	N/A
Robbery (time only)	.4783	.0857	.0104	**.0094**	.0183
Seismic (time only)	.1266	.0173	**.0045**	.0150	.0122

Interpretable conditional intensity To interpret the spatial dependence learned using our model, we plot the conditional intensity as a heatmap over the space at a specific time frame. For the state-of-the-art, as shown in Figure 4, we can see that the ETAS (the third column) captured the general pattern of the conditional intensity over the space, where regions with more events tend to have higher intensity. Comparing with the result shown in first two columns of Figure 4, our NEST is able to capture complex spatial pattern at different locations and the shape of the captured diffusion also differ from application to application. For 911 calls-for-service data, shown in the first row of Figure 4, the spatial influence of some robbery events diffuse to the surrounding streets and the community blocks unevenly. For seismic data, shown in the second row of Figure 4, the spatial influence of some events majorly diffuses along the earthquake fault lines.

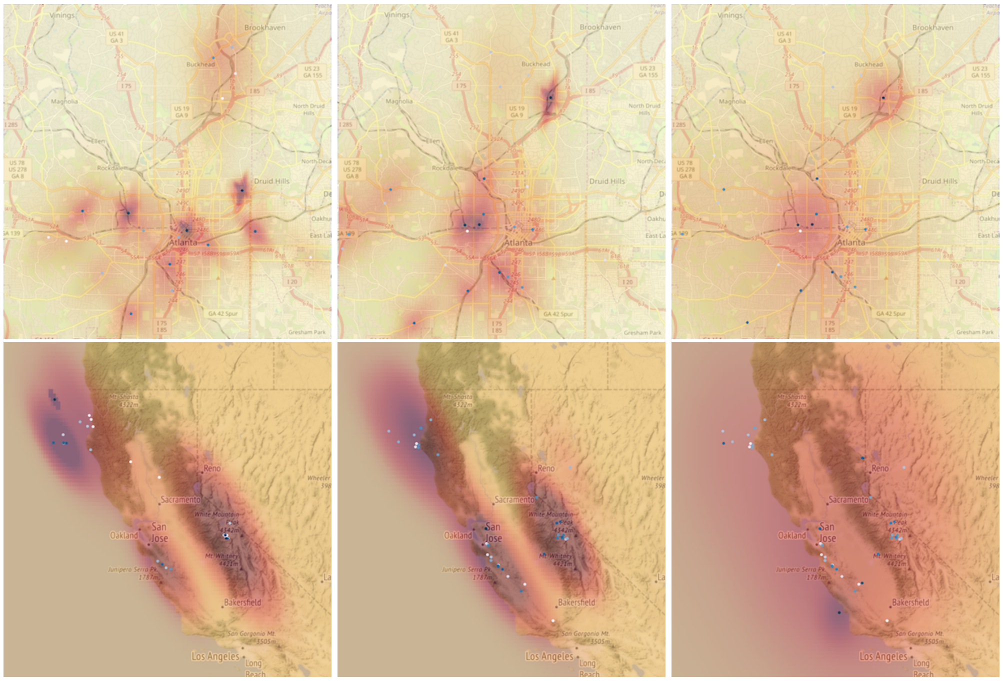

Figure 4: Snapshots of the conditional intensity for two real data sequences (crime events in Atlanta and seismic events): First and second row show snapshots of the conditional intensity for a series of robberies in Atlanta and a series of earthquakes in North of California, respectively. First two columns are generated by `NEST+MLE` ($K = 5$) and the third column is generated by `ETAS`. The color depth indicate the value of intensity. The region in darker red has higher risk to have next event happened again.

References

[1] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. II*. Probability and its Applications (New York). Springer, New York, second edition, 2008. General theory and structure.

[2] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1555–1564, New York, NY, USA, 2016. ACM.

[3] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pages 258–267, Arlington, Virginia, United States, 2015. AUAI Press.

[4] Eric W. Fox, Martin B. Short, Frederic P. Schoenberg, Kathryn D. Coronges, and Andrea L. Bertozzi. Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, 111(514):564–584, 2016.

[5] Edith Gabriel, Barry Rowlingson, and Peter Diggle. stpp: An r package for plotting, simulating and analyzing spatio-temporal point patterns. *Journal of Statistical Software*, 53:1–29, 04 2013.

[6] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

[7] ALAN G. HAWKES. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 04 1971.

[8] Beomjoon Kim and Joelle Pineau. Maximum mean discrepancy imitation learning. In *Robotics: Science and Systems*, 2013.

[9] Northern California Earthquake Data Center. UC Berkeley Seismological Laboratory. NCEDC, 2014.

[10] Erik Lewis, George Mohler, P Jeffrey Brantingham, and Andrea L Bertozzi. Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264, Jul 2012.

[11] Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning temporal point processes via reinforcement learning. In *Advances in Neural Information Processing Systems 31*, pages 10781–10791. Curran Associates, Inc., 2018.

[12] Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6757–6767, USA, 2017. Curran Associates Inc.

[13] F. Musmeci and D. Vere-Jones. A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, 44(1):1–11, Mar 1992.

[14] Yosihiko Ogata. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, January 1981.

[15] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[16] Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

[17] Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.*, 33(3):299–318, 08 2018.

[18] Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodriguez. Deep reinforcement learning of marked temporal point processes. In *Advances in Neural Information Processing Systems 31*, 2018.

[19] Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Le Song, and Hongyuan Zha. Wasserstein learning of deep generative point process models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3250–3259. Curran Associates Inc., 2017.

[20] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M. Chu. Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1597–1603. AAAI Press, 2017.

[21] Shixiang Zhu and Yao Xie. Crime event embedding with unsupervised feature selection, 2018.

[22] Shixiang Zhu and Yao Xie. Crime incidents embedding using restricted boltzmann machines. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2376–2380, 2018.

[23] Shixiang Zhu and Yao Xie. Spatial-temporal-textual point processes with applications in crime linkage detection, 2019.

[24] Shixiang Zhu, Henry Shaowu Yuchi, and Yao Xie. Adversarial anomaly detection for marked spatio-temporal streaming data, 2019.

[25] Joseph R. Zipkin, Frederic P. Schoenberg, Kathryn Coronges, and Andrea L. Bertozzi. Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics*, 27(3):502–529, 2016.