Neural Spectral Marked Point Processes

Shixiang Zhu[†], Haoyun Wang[†], Zheng Dong[†], Xiuyuan Cheng[*], Yao Xie[†]

[†] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
[*] Department of Mathematics, Duke University

Introduction

Self- and mutually-exciting point processes are popular models for dependent discrete event data. To date, most existing models assume stationary kernels (including the classical Hawkes processes) and simple parametric models. Modern applications with complex event data require more general models that can incorporate contextual information of the events, called marks, besides the temporal and location information. Moreover, such applications often require non-stationarity to capture more complex spatio-temporal dependence. In this paper, we introduce a novel and general neural network-based non-stationary influence kernel with high expressiveness for handling complex discrete events data while providing theoretical performance guarantees. We demonstrate the superior performance of our proposed method compared with the state-of-the-art on synthetic and real data.

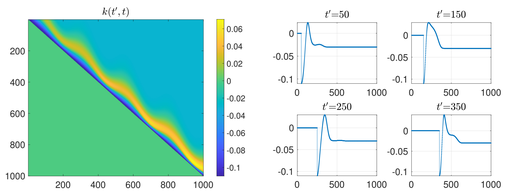


Figure: An example of non-stationary influence kernel $k(t', t)$ of event time t' and future time $t > t'$.

Marked Temporal Point Process

A data point in the discrete event sequence:

$$x = (t, m), \quad t \in [0, T), \quad m \in \mathcal{M},$$

Conditional intensity function:

$$\lambda(x)dx = \mathbb{E}\left(d\mathbb{N}(x)|\mathcal{H}_{t(x)}\right).$$

Hawkes process with kernel $k \in \mathcal{K} \subset C^0(\mathcal{X} \times \mathcal{X})$:

$$\lambda[k](x) = \mu + \int_{x' \in \mathcal{X}_{t(x)}} k(x', x) d\mathbb{N}(x'), \tag{1}$$

Log-likelihood given M sequences $\{x_{i,j}\}$, $i = 1, ..., N_j$, $j = 1, \ldots, M$

$$\ell[k] \coloneqq \frac{1}{M} \sum_{j=1}^{M} \left(\int_{\mathcal{X}} \log \lambda_j[k](x) d\mathbb{N}_j(x) - \int_{\mathcal{X}} \lambda_j[k](x) dx \right), \tag{2}$$

Contribution

- ✓ The kernel function is represented by a **spectral decomposition of the influence kernel**.
- ✓ The spectral decomposition of asymmetric influence kernel consists of a sum of the product of feature maps, which can be **parameterized by neural networks**.
- ✓ We establish **theoretical guarantees of the MLE for the true kernel function** based on functional variational analysis and finite-dimensional asymptotic analysis, which shed light on theoretical understanding of neural network-based kernel functions.

Neural Spectral Kernel

Influence kernel k is represented using finite-rank decomposition:

$$k(x', x) = \sum_{r=1}^{R} \nu_r \psi_r(x') \phi_r(x), \quad \nu_r \geq 0, \tag{3}$$

where

$$\psi_r : \mathcal{X} \to \mathbb{R}, \quad \phi_r : \mathcal{X} \to \mathbb{R}, \quad r = 1, \cdots, R,$$

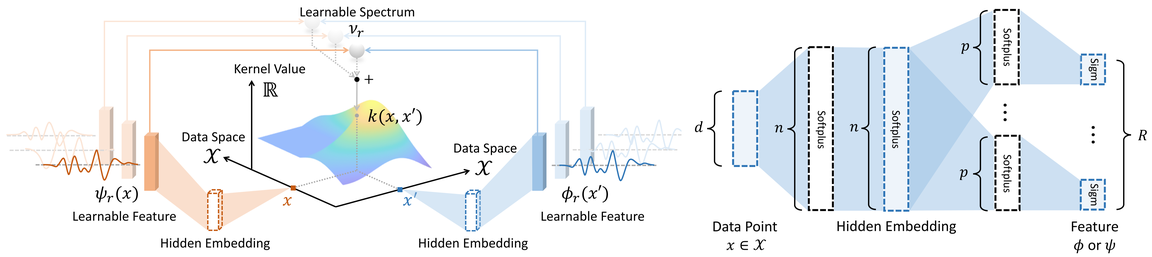
are two sets of *feature functions* in some smooth functional space $\mathcal{F} \subset C^0(\mathcal{X})$, and ν_r is the corresponding weight or *spectrum*.

Kernel Identifiability

Assumption: (A1) The kernel function family $\overline{\mathcal{K}} \subset C^0(\mathcal{X} \times \mathcal{X})$, and kernel functions in $\overline{\mathcal{K}}$ is uniformly bounded; (A2) There exist c_1, c_2 positive constants, such that for any $k \in \overline{\mathcal{K}}$, $c_1 \leq \lambda_j[k](x) \leq c_2$, $\forall x \in \mathcal{X}$ and $\forall j$.

Kernel identifiability using maximum likelihood: Under Assumption, the true kernel function k^* is locally identifiable in that k^* is a local minimum solution of maximum likelihood (2) in expectation.

Kernel Architecture and Network Structure



(a) Kernel architecture

(b) Network structure

Figure: (a) The architecture of the proposed non-stationary neural spectral kernel. (b) The structure of the multi-branch neural network.

Optimal fit using feature function basis representation

- Assume that the feature functions of the kernel can be well-approximated by a linear combination of basis functions: $b_i(x) : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, S$:

$$\psi_r(x) = \sum_{i=1}^{S} \alpha_{ri} b_i(x), \ \phi_r(x) = \sum_{i=1}^{S} \beta_{ri} b_i(x), \ r = 1, \cdots, R.$$

The kernel function in (3) can be written as $k_A(x', x) = b(x')^T A b(x)$, where $A_{pq} = \sum_{r=1}^{R} \nu_r \alpha_{rp} \beta_{rq}$.

- Let $\widetilde{A} \in \mathcal{A}$ be the one which maximizes the expected log-likelihood function. i.e.

$$\widetilde{A} = \arg\max_{A \in \mathcal{A}} \mathbb{E}\left(\ell_A\right). \tag{4}$$

Under Assumption, let the ℓ_2-norm of a kernel be

$$\|k\|_2^2 = \int_{\mathcal{X}} \int_{\mathcal{X}_{t(x)}} k(x', x)^2 dx' dx. \tag{5}$$

Then we have

$$\|k^* - k_{\widetilde{A}}\|_2^2 \leq \frac{c_2^5 |\mathcal{M}| T + c_2^4}{c_1^4} \exp(2(c_2 - c_1)|\mathcal{M}|T) D(k^*, \mathcal{K}_{\text{finite}})^2,$$

where $D(k^*, \mathcal{K}_{\text{finite}})$ is the ℓ_2-distance between the true kernel and the set $\mathcal{K}_{\text{finite}}$,

$$D(k^*, \mathcal{K}_{\text{finite}}) = \min_{k \in \mathcal{K}_{\text{finite}}} \|k^* - k\|_2.$$

Asymptotic normality of low-rank MLE

Let the singular value decomposition of $\widetilde{A}$ be $\widetilde{A} = U \Lambda V^T$ and $\widetilde{I}$ be the expected Hessian matrix of the log-likelihood at $\widetilde{A}$. Let $\widetilde{J}$ be the covariance matrix of a single trajectory's score function at $\widetilde{A}$, and $\widetilde{G} \in \mathbb{R}^{S \times S}$ be the expected score at $\widetilde{A}$, Let $F = (\mathbb{I}_S \otimes U, V \otimes \mathbb{I}_S) \in \mathbb{R}^{S^2 \times 2SR}$ where $\otimes$ is the Kronecker product, $\mathbb{I}_S$ is the identity matrix of size S, $\widetilde{C} = (\widetilde{A}^\dagger \otimes \widetilde{G}) Q_{S,S} + ((\widetilde{A}^\dagger \otimes \widetilde{G}) Q_{S,S})^T$, where $\dagger$ represents pseudo-inverse and $Q_{a,b} \in \mathbb{R}^{ab \times ab}$ is the permutation matrix such that $\text{vec}(P^T) = Q_{a,b} \text{vec}(P)$ for any a-by-b matrix P.

If $F^T(\widetilde{I} + \widetilde{C})F$ shares the same null-space with F, then the low-rank estimator $\widehat{A}_{\text{MLE}}$, solved from the constrained maximum likelihood problem, when $M \to \infty$, satisfies

$$\sqrt{M}(\text{vec}(\widehat{A}_{\text{MLE}}) - \text{vec}(\widetilde{A})) \to \mathcal{N}(0, F(F^T(\widetilde{I} + \widetilde{C})F)^\dagger F^T \widetilde{J} F (F^T(\widetilde{I} + \widetilde{C})F)^\dagger F),$$

Real Data Experiments

Table: Predictive log-likelihood on real data.

ℓ	Earthquake (2D)	Robbery (1D)	#Parameters	Training/Testing time[2]
NSMPP	**-56.50**	**-74.47**	171,555	0.766 / 0.84
RMTPP	-218.39	-132.55	274,168	0.245 / 7.29
Neural Hawkes	-189.39	-96.10	282,755	0.204 / 6.09
Hawkes	NA	-197.84	2	0.021 / <0.01

Numerical Experiments

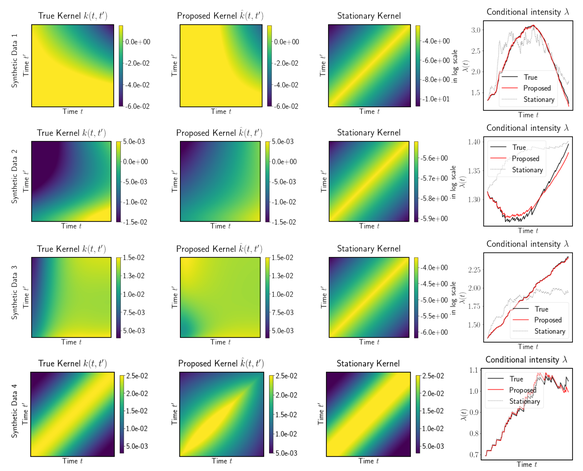


Figure: Recovered kernels using three one-dimensional synthetic data sets.

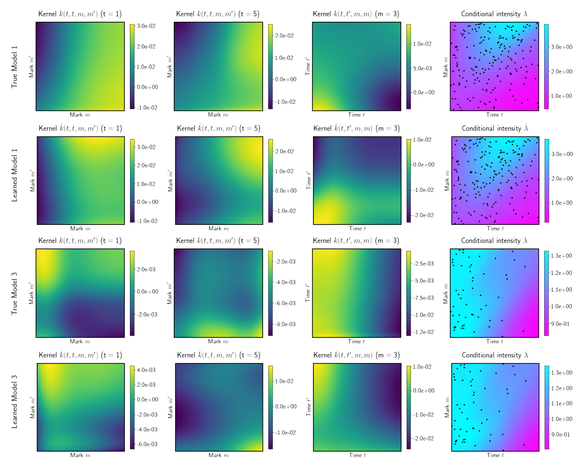


Figure: Recovered kernels using three two-dimensional synthetic data sets.

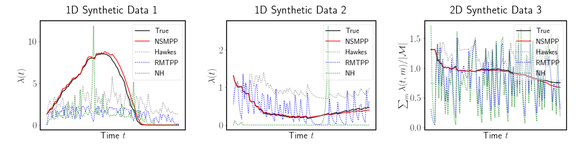


Figure: Predicted conditional intensity using our method and other baseline approaches for synthetic data sets.

Reference

Zhu, Shixiang, Haoyun Wang, Zheng Dong, Xiuyuan Cheng, and Yao Xie. "Neural Spectral Marked Point Processes." arXiv:2106.10773 (2021).