Online Prediction For High-dimensional Discrete Event Data

Chen Xu
Advised by Prof. Yao Xie

Georgia Tech ISyE
H. Milton Stewart School of
Industrial and Systems Engineering

INFORMS 2021 General Session, Adaptive Online Learning of High-dimensional Data, October 2021

Outline

Modeling high-dimensional spatio-temporal discrete data

- Approach #1: Interactive Categorical Point Process
- Approach #2: Marked spatio-temporal Hawkes Process and Conformal Prediction on dependent data (Ongoing Work)

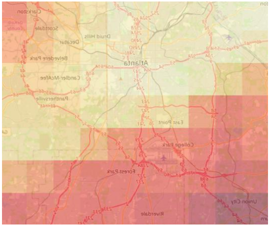
[1] *Solar Radiation Anomaly Events Modeling Using Spatial-Temporal Mutually Interactive Processes.* Minghe Zhang, Chen Xu, Andy Sun, Feng Qiu, Yao Xie. Submitted to Annals of Applied Statistics. 2020.

Problem setup and Goal

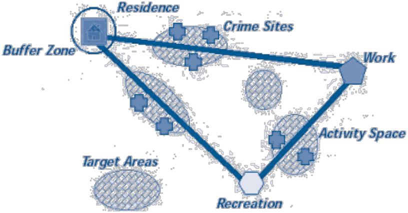
- Assume $\omega_0, \omega_1, \ldots$ are categorical random vectors in $\mathbb{R}^K$, where ω_{tk} indicates the event type at location k and time t.
- Examples includes the type of crime, magnitude of seismic activity, etc.
- Goal: Model the conditional distribution $\omega_t | \omega_{t-1}, \omega_{t-2}, \ldots$

Motivating applications

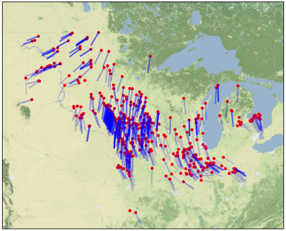
Solar Ramping event detection



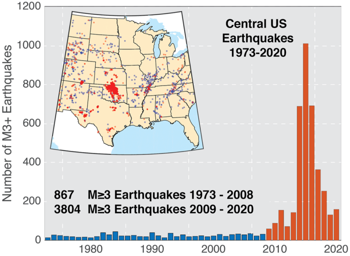
Crime Pattern Analysis



Wind power prediction



Seismic Activity



Literature

- Stochastic event modeling: Hawkes process model (Hawkes, 1971), self-correcting point process model (Isham and Westcott, 1979), event propagation (Wu et al., 2020), attention-based point processes (Zhu et al., 2020). . .
- Parametric and nonparametric spatio-temporal processes: non-parametric Hawkes process (Moller and Waagepetersen 2003), estimation in Bayesian framework (Python et al., 2016), sparse model estimation for point processes (Hansen et al., 2015). . .

Modeling

- Assume each $\omega_{tk} \in \{0, 1, \ldots, M\}$ for M non-zero categories. Denote $\boldsymbol{\omega}_t^{-d} = [\omega_{t-1}, \ldots, \omega_{t-d}]$.
- We model the conditional probability as

$$\mathbb{P}\left[\omega_{tk} = p \mid \boldsymbol{\omega}_t^{-d}\right] = \bar{\beta}_k(p) + \sum_{s=1}^{d} \sum_{\ell=1}^{K} \bar{\beta}_{k\ell}^{s}\left(p, \omega_{(t-s)\ell}\right).$$

Modeling

- Assume each $\omega_{tk} \in \{0, 1, \ldots, M\}$ for M non-zero categories. Denote $\boldsymbol{\omega}_t^{-d} = [\omega_{t-1}, \ldots, \omega_{t-d}]$.
- We model the conditional probability as

$$\mathbb{P}\left[\omega_{tk} = p \mid \boldsymbol{\omega}_t^{-d}\right] = \bar{\beta}_k(p) + \sum_{s=1}^{d} \sum_{\ell=1}^{K} \bar{\beta}_{k\ell}^{s}\left(p, \omega_{(t-s)\ell}\right).$$

- Compare with generalized linear models.
- Parameters to be estimated are $\bar{\beta}_k$ (birthrate) and $\bar{\beta}_{k\ell}^{s}$ (interaction), subject to probability constraints.

[2] *Convex Parameter Recovery for Interacting Marked Processes.* Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, Yao Xie. IEEE Journal on Selected Areas in Information Theory. 2020

Parameter Estimation with Guarantee

- Estimation via constrained (1) Maximum Likelihood Estimation or (2) Least-square Estimation.

  Both are convex programming and (2) is equivalent to solving a special case of variational inequality (Details see [1, 2]).

Parameter Estimation with Guarantee

- Estimation via constrained (1) Maximum Likelihood Estimation or (2) Least-square Estimation.

  Both are convex programming and (2) is equivalent to solving a special case of variational inequality (Details see [1, 2]).

- Guarantee: Suppose we have N observations of ω_t and let κ be total number of parameters. With probability at least $1-\epsilon$,

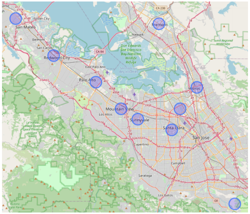
$$\text{For (1), } \left\|\hat{\boldsymbol{\beta}}^{\mathrm{LS}}-\boldsymbol{\beta}\right\|_p \leq C_{1,p}\frac{\ln(2\kappa/\epsilon)}{N}.$$

$$\text{For (2), } \left\|\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}-\boldsymbol{\beta}\right\|_p \leq C_{2,p}\sqrt{\frac{\ln(2\kappa/\epsilon)}{N}}.$$

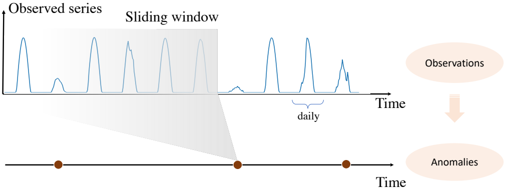
- Proof Ingredients: Martingale-difference, Concentration inequality, Properties of variational inequality (Details see [1, 2]).

Experiment: Solar ramping event prediction

- Ramping events are anomalies in the time-series, due to extremely high/low observations.
- Some raw radiation values come from $K = 10$ sparsely distributed sensors over two years (Fig (a)).
- We define ramping events via comparing with lower/upper quantiles of past values (Fig (b)).
- We also fit models by season for better estimation (Details in [1]).



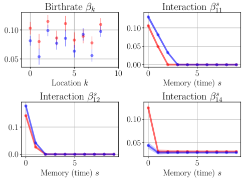
(a)



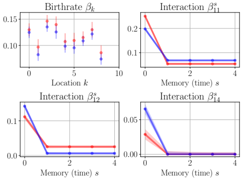
(b)

Parameter estimation results

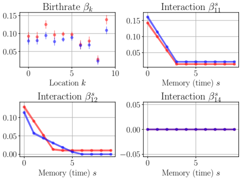
- Birthrate and Interactions (Temporal view)



(a) Atlanta



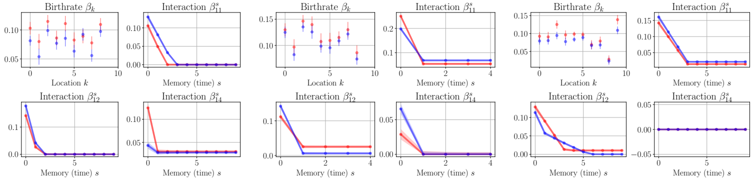
(b) Los Angeles



(c) California

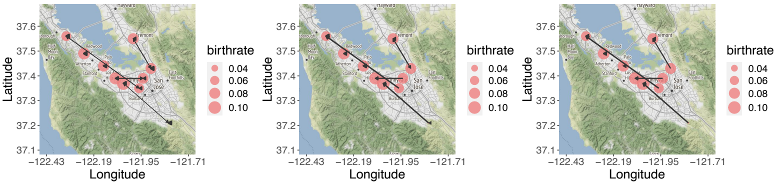
Parameter estimation results

- Birthrate and Interactions (Temporal view)



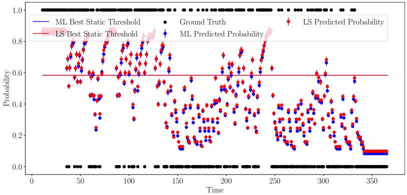
(a) Atlanta **(b)** Los Angeles **(c)** California

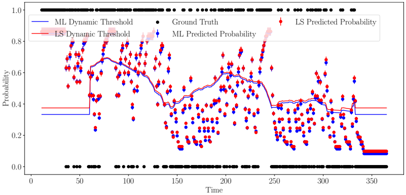
- Interactions (Spatio-temporal view, $s \in \{1, 5, 10\}$).



Prediction performance

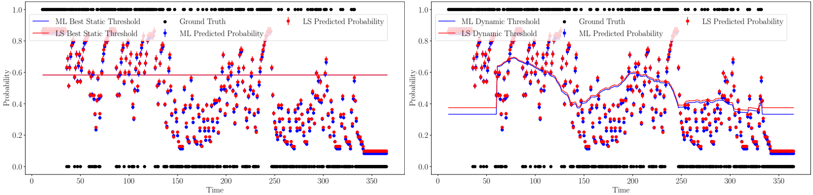
- Conditional probabilities when $M = 2$ (e.g. up and down ramping events).





Prediction performance

- Conditional probabilities when $M = 2$ (e.g. up and down ramping events).



- Precision, Recall, and F_1 under static and dynamic thresholds (Details see [1]).

		Least Square			Maximum Likelihood		
Location	τ	Precision	Recall	F_1	Precision	Recall	F_1
Atlanta	Static	0.79	0.95	0.86	**0.97**	**0.98**	**0.97**
	Dynamic	0.82	0.95	0.88	0.96	0.96	0.96
Los Angeles	Static	0.91	0.78	0.84	0.91	0.81	0.86
	Dynamic	**0.91**	0.83	0.87	0.85	**0.91**	**0.88**
Palo Alto	Static	**0.95**	0.60	0.73	0.94	0.52	0.67
	Dynamic	0.78	**0.89**	**0.83**	0.76	0.88	0.82

Extension to continuous processes

- The model can be naturally extended to model continuous random vectors (below is ongoing work).
- In particular, we can assume $\omega_t \sim N(\mu_t, \Theta_t)$, where

$$\mu_t = \sum_{s=1}^{d} A_s \omega_{t-s}, \quad \Theta_t = \sum_{s=1}^{d} \Gamma_s \circ \left(\omega_{t-s}\omega_{t-s}^T\right).$$

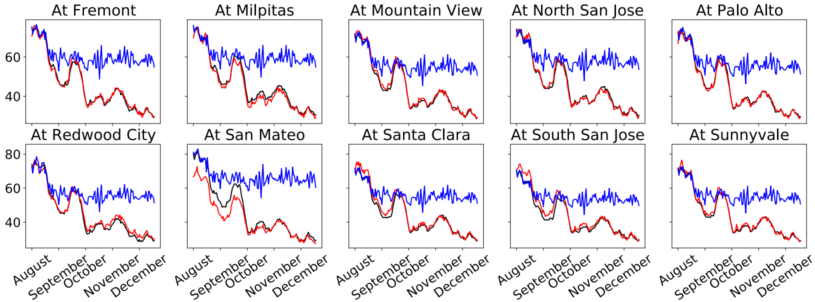
- Similar parameter estimation and guarantee hold.

Extension to continuous processes

- The model can be naturally extended to model continuous random vectors (below is ongoing work).
- In particular, we can assume $\omega_t \sim N(\mu_t, \Theta_t)$, where

$$\mu_t = \sum_{s=1}^{d} A_s \omega_{t-s}, \quad \Theta_t = \sum_{s=1}^{d} \Gamma_s \circ \left(\omega_{t-s}\omega_{t-s}^T\right).$$

- Similar parameter estimation and guarantee hold.
- Prediction result on earlier *raw* solar data



Outline

Modeling high-dimensional spatio-temporal discrete data

- Approach #1: Interactive Categorical Point Process
- Approach #2: Marked Spatio-temporal Hawkes Process and Conformal Prediction on dependent data (Ongoing Work)

Motivation

- More complex marks/features are available (e.g. weather variables for predicting natural hazard).
- Events happen in continuous time (e.g. influence changes over time and could not be captured by fixed parameters).
- For prediction purposes, classification methods are more widely used than point processes, whose uncertainty analyses remain largely unexplored.

Problem setup

- Assume each observed datum

$$x_i = (t_i, u_i, m_i), t_i \in [0, T], u_i \in \mathbb{R}^2, m_i \in \mathbb{R}^d$$

is a tuple consisting of time, location, and marks.
In particular, $m_i = (z_i, m_i')$ contains both static marks z_i and dynamic marks m_i'.

Problem setup

- Assume each observed datum

$$x_i = (t_i, u_i, m_i), t_i \in [0, T], u_i \in \mathbb{R}^2, m_i \in \mathbb{R}^d$$

  is a tuple consisting of time, location, and marks.
  In particular, $m_i = (z_i, m'_i)$ contains both static marks z_i and dynamic marks m'_i.

- We have two tasks:
  1. Model the probability of a new event occurring after time T.
     Solution: Marked spatio-temporal Hawkes process models
  2. Suppose each event is of type $y_i \in \{0, \ldots, M\}$. Predict the type of new event that occurs.
     Solution: Machine learning classifiers, which are calibrated to provide uncertainty sets for true event types.

Method–Marked Spatio-temporal Hawkes process

- The point process intends to model the conditional intensity

$$\lambda\left(t, u \mid \mathcal{H}_t\right) := \lim_{\Delta t, \Delta u \to 0} \frac{\mathbb{E}\left[N\left([t, t+\Delta t) \times B(u, \Delta u) \mid \mathcal{H}_t\right]\right.}{\Delta t \times B(u, \Delta u)}.$$

- In particular, we assume the intensity has the form

$$\lambda\left(t, u \mid \mathcal{H}_t\right) = f_{u,z} + \sum_{j: t_j < t} f_u\left(u, u_j\right) \cdot f_t\left(t, t_j\right) \cdot f_{m'}\left(m'_j\right), \qquad (1)$$

Method–Marked Spatio-temporal Hawkes process

- The point process intends to model the conditional intensity

$$\lambda\left(t, u \mid \mathcal{H}_t\right) := \lim_{\Delta t, \Delta u \to 0} \frac{\mathbb{E}\left[N\left([t, t+\Delta t) \times B(u, \Delta u) \mid \mathcal{H}_t\right]\right.}{\Delta t \times B(u, \Delta u)}.$$

- In particular, we assume the intensity has the form

$$\lambda\left(t, u \mid \mathcal{H}_t\right) = f_{u,z} + \sum_{j: t_j < t} f_u\left(u, u_j\right) \cdot f_t\left(t, t_j\right) \cdot f_{m'}\left(m_j'\right), \quad (1)$$

  - $f_{u,z}$ is the baseline rate, assuming location is discretized. It may include influence from static marks.
  - $f_u\left(u, u_j\right)$ captures interactions, which can be bi-variate kernels and/or graph neural networks.
  - $f_t\left(t, t_j\right)$ denotes temporal influence from past events.
  - $f_{m'}\left(m_j'\right)$ measures contribution from dynamic marks,
- Please see [3] in the references for a survey of Hawkes process models.

Estimation–Marked Spatio-temporal Hawkes process

- Let Θ denote parameters. The full log-likelihood can be derived as

$$\ell(\Theta) = \sum_{i=1}^{n} \log\left(\lambda\left(u_i, t_i\right)\right) - \int_0^T \int_U \lambda(u,t)\mathrm{d}u\mathrm{d}t, \qquad (2)$$

  whereby parameters are solved via maximum likelihood estimation.

- In general, (2) is non-convex in Θ with high computational cost. However, our work parametrizes $\lambda(t,k|\mathcal{H}_t)$ so that $\ell(\Theta)$ is convex in all but one parameter. The estimates can thus be found via simple iterative procedures.

- Once parameters are estimated, they are substituted in (1) for prediction.

Method–Conformal Prediction for Dependent Data

- Typical classification setting, where $X_i \in \mathbb{R}^{1+2+d}$ (time, location, feature dimension) is used to predict the type Y_i.
- Predictor $\hat{f}$ can be generic machine learning models.
- In particular, we want to produce uncertainty sets C_t^α for future observations, so that marginal coverage holds:

$$\mathbb{P}(Y_t \in C_t^\alpha) \geq 1 - \alpha. \tag{3}$$

Importantly, we want C_t^α to be distribution-free.

Method–Conformal Prediction for Dependent Data

- Typical classification setting, where $X_i \in \mathbb{R}^{1+2+d}$ (time, location, feature dimension) is used to predict the type Y_i.
- Predictor $\hat{f}$ can be generic machine learning models.
- In particular, we want to produce uncertainty sets C_t^α for future observations, so that marginal coverage holds:

  $$\mathbb{P}(Y_t \in C_t^\alpha) \geq 1 - \alpha. \tag{3}$$

  Importantly, we want C_t^α to be distribution-free.
- One solution is conformal prediction [4], which includes in C_t^α all possible types that *conform* to past observed types.
  Limitation: Data must be exchangeable.
- Remedy: Inspired by our recent work [5], we design methods that achieve (3) approximately.

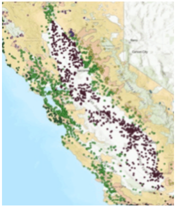
  Specifically, we carefully design the conformity metrics and efficiently train leave-one-out ensemble predictors to maximize power.

Literature

- Hawkes Process: Origin (Hawkes, 1971), Cluster process (Daley and Vere-Jones, 2003, Section 6.3; Gonzalez et al., 2016), Spatial point processes (Diggle, 2014), Applications in finance (Bauwens and Hautsch, 2009; Bacry, Mastromatteo and Muzy, 2015) and in neuron activity (Johnson, 1996)...
- Conformal Prediction: Regression (Papadopoulos et al., 2007; Barber et al., 2019; Kim, Xu, and Barber, 2020) and Classification (Angelopoulos et al., 2020; Romano et al., 2020).

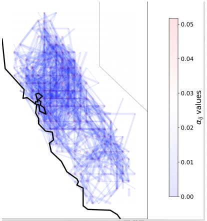
Experiment–Marked Spatio-temporal Hawkes process

- We model the occurrence of California wildfire, which is known to have strong dependencies.
  Figure on the right shows the distribution of fire events, with different clustering patterns.



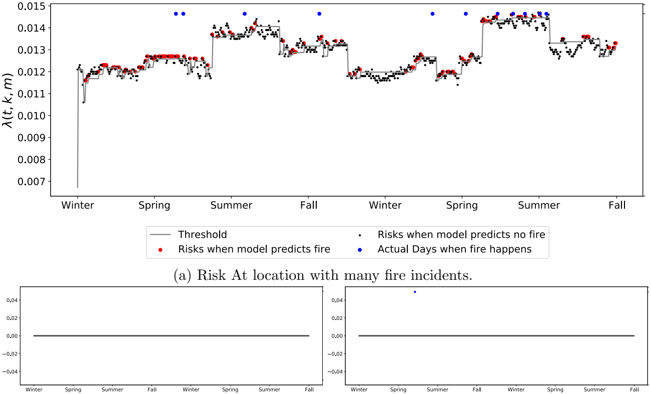
- Estimated parameters

	Top 3 Largest Estimates			Bottom 3 Smallest Estimates		
baseline estimates	0.027	0.019	0.017	0.001	0.001	0.001
grid location	127	41	152	121	36	86
static mark estimates	0.605	0.424	0.093	0.002	0.002	0.001
feature name	PG&E	Fire Tier1	Fire Tier2	PHYS=Exotic Tree-Shrub	PHYS=Hardwood	PHYS=Grassland
dynamic mark estimates	0.374	0.358	0.345	0.202	0.107	0.038
feature name	Summer	Temperature	Relative Humidity	LFP	Spring	Winter



Experiment–Marked Spatio-temporal Hawkes process (cont.)

- Prediction with dynamic hedging thresholds, which take into account the accuracy of past predictions and historical patterns.



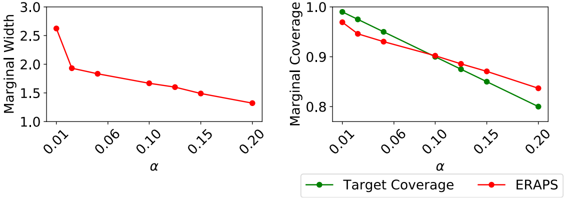
(a) Risk At location with many fire incidents.

(b) Risk At location with no fire incidents.

(c) Risk At location with very few fire incidents.

Experiment–Conformal Prediction for Dependent Data

- Our method is named ERAPS. The current results are under random forest classifiers.



Conditional Coverage

α	[0,.25Acres)	[.25Acres,10Acres)	[10Acres,100Acres)	[100Acres,1000Acres)	[1000Acres,É)
0.01	1.00	1.00	0.43	0.06	0.00
0.02	1.00	0.94	0.04	0.00	0.00
0.05	1.00	0.85	0.04	0.00	0.00
0.10	1.00	0.70	0.00	0.00	0.00
0.12	0.99	0.63	0.00	0.00	0.00
0.15	0.99	0.56	0.02	0.00	0.00
0.20	0.98	0.43	0.00	0.00	0.00

Conditional Set Size

α	[0,.25Acres)	[.25Acres,10Acres)	[10Acres,100Acres)	[100Acres,1000Acres)	[1000Acres,É)
0.01	2.65	2.53	2.53	2.56	2.67
0.02	1.93	1.95	1.84	1.94	2.00
0.05	1.83	1.86	1.78	2.00	1.83
0.10	1.66	1.70	1.69	1.88	1.83
0.12	1.58	1.63	1.71	1.94	1.83
0.15	1.46	1.56	1.61	2.00	1.67
0.20	1.29	1.42	1.47	1.56	1.67

Summary

Approach #1

- Model high-dimensional categorical random vectors with a flexible categorical point process model.
- The problem is convex, with provable performance guarantee on estimated parameters.
- Can be generalized to Gaussian processes and beyond.

Approach #2

- Model the conditional intensity of correlated observations with a flexible marked spatio-temporal Hawkes process.
- Design the process to yield convex likelihood with efficient solving procedures.
- Provide uncertainty quantification for machine learning classifiers using recent advances in conformal prediction.

References

[1] *Solar Radiation Anomaly Events Modeling Using Spatial-Temporal Mutually Interactive Processes.* Minghe Zhang, Chen Xu, Andy Sun, Feng Qiu, Yao Xie. Submitted to Annals of Applied Statistics. 2020

[2] *Convex Parameter Recovery for Interacting Marked Processes.* Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, Yao Xie. IEEE Journal on Selected Areas in Information Theory. 2020

[3] *A Review of Self-Exciting Spatio-Temporal Point Processes and Their Application.* Alex Reinhart. Statistical Science. 2018

[4] *A Tutorial on Conformal Prediction.* Glenn Shafer, Vladimir Vovk. Journal of Machine Learning Research. 2008

[5] *Conformal Prediction Interval for Dynamic Time-series.* Chen Xu, Yao Xie. Thirty-eighth International Conference on Machine Learning. 2021