High-Dimensional Autoregressive Point Processes

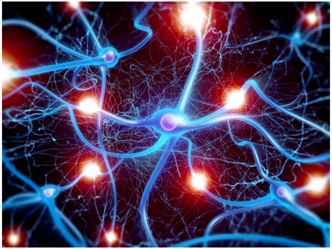
Ben Mark, joint work with Rebecca Willett and Garvesh Raskutti

UW-Madison

Point Processes

Interested in studying cascading series of events in networks. Examples include:

- Biological neural networks: neuron firings can inhibit or stimulate other neurons (Smith & Brown (2004))
- Social networks: users share their friends' content (Zhou et al. (2013))
- Crime: violence from one gang can lead to retaliatory violence from another gang (Bertozzi et al. (2011))



- Goal: estimate network structure from event data
- Network is possibly large relative to number of events we observe, but we assume it is sparse.

Related Work

Multi-Variate Poisson Autoregressive Model[1]:

$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1})$$

$$\log(\lambda_{t+1}) = \nu + A^* X_t$$

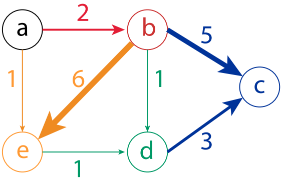
- $\nu =$ background rate
- $X_{t,m} =$ number of events from node m during time period t
- $A^* =$ influence matrix to be estimated

[1] cf., Hall et al. (2016)

Related Work

Multi-Variate Poisson Autoregressive Model:

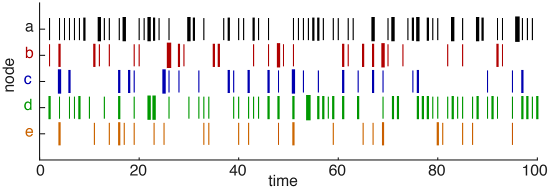
$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1}) \text{ where } \log(\lambda_{t+1}) = \nu + A^* X_t$$



Toy inhibitory network

$$A^* = -\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 3 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 6 & 0 & 0 & 0 \end{bmatrix}$$

Corresponding A^* (weighted adjacency matrix)



Observed events

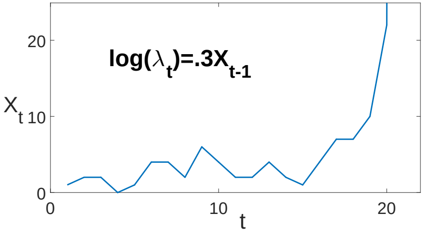
Related Work

Multi-Variate Poisson Autoregressive Model:

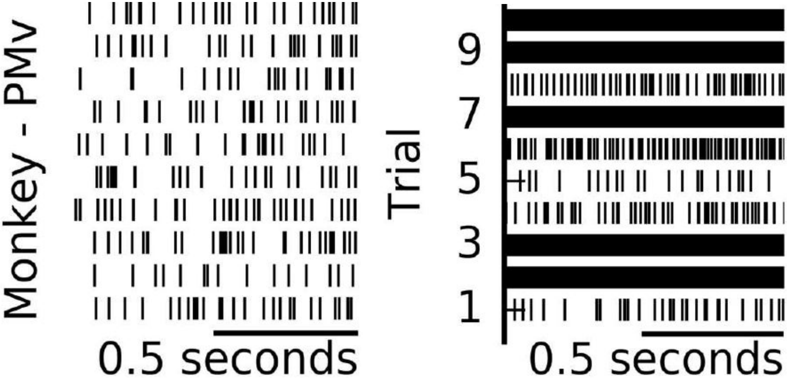
$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1})$$

$$\log(\lambda_{t+1}) = \nu + A^* X_t$$

- Two key limitations:
  - Model only considers first order effects
  - Due to log link function, process can be highly unstable with positive A^*. Hall et al. give sample complexity bounds assuming $A^* \leq 0$



Log-linear point processes make bad generative models



Spike train data from monkey cortex. Each row represents a single trial.

Simulated data generated by model learned from spike train data.

Figure 1: Figures from Gerhard et al. (2016).

Related Work

Multi-Variate Poisson Autoregressive Model:

$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1})$$

$$\log(\lambda_{t+1}) = \nu + A^* X_t$$

- Practitioners are interested in log-linear point process models [2], but unrealistic as generative model [3]
- Need stability to facilitate analysis, and understand space of networks we can infer.
- No infinite rates in practice, real systems have dampening effects [4]

[2] cf., Laub (2015); Mensi et al. (2011); Weber & Pillow (2016)

[3] cf., Gerhard et al. (2016)

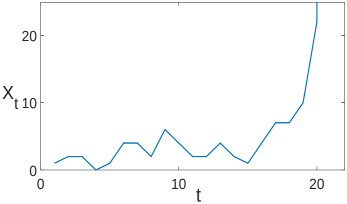
[4] cf., Ertekin et al. (2015)

Saturation effects ensure stability

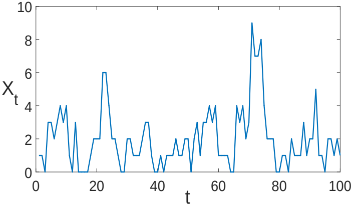
Clipped PAR Model

$\log(\lambda_{t+1}) = \nu + A^* \min(X_t, K)$ for some constant K

- Clipped PAR model is stable with stimulatory effects.
- How does clipping function effect our ability to learn?



Unclipped PAR for $\log(\lambda_t) = .3X_{t-1}$.
$X_{20} = 23$ while $X_{22} \approx 10^{98}$



Clipped PAR for $\log(\lambda_t) = .3\min(X_{t-1}, 6)$.

Saturation effects

- What is the space of networks we can reconstruct with clipping? How many observations do we need?
- Should depend on amount of clipping and structure of network. Connections can't be so stimulatory that we're constantly clipping.

ARMA(1,1) model

In a variety of applications want to consider longer term memory. Consider:

$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1})$$

$$\log(\lambda_{t+1}) = \nu_t + A^*(\sum_{i=0}^{t} \alpha^i X_{t-i}) \tag{1}$$

Similar form to PAR, but is equivalent to an ARMA(1,1) model:

$$\log(\lambda_{t+1}) = \nu + A^* X_t + \alpha \log(\lambda_t) \tag{2}$$

Clipped ARMA(1,1) Model

Notation

Let $g(\mathcal{X}_t) = \sum_{i=0}^{t} \alpha^i \min(X_{t-i}, K)$ for some constant K

Clipped ARMA(1,1) Model:

$$X_{t+1} \sim \text{Poisson}(\lambda_{t+1})$$

$$\log(\lambda_{t+1}) = \nu_t + A^* g(\mathcal{X}_t)$$

- Guarantees stability and incorporates long range memory but clipping creates challenges in deriving performance guarantees.
- When $\alpha = 0$ and $K = \infty$ this reduces to PAR model

Related Work

- Discrete-time point process models in low-dimensional setting (e.g. INGARCH model) [5]
- Continuous time models (e.g. Hawkes process) [6]
- Application driven works incorporating saturation effects [7]

[5] cf., Heinen (2003); Ferland et al. (2006)

[6] Hansen et al. (2012); Etesami et al. (2016)

[7] cf., Ertekin et al. (2015)

Regularized MLE

Estimate A^* using regularized maximum likelihood estimation:

$$\widehat{A} = \arg\min_{A} \quad \underbrace{-L(A|\mathcal{X}_T)}_{\text{negative log-likelihood}} \quad + \underbrace{\lambda||A||_1}_{\text{regularizer}}$$

- Convex optimization problem
- Incorporates sparsity assumption
- Decomposable in rows of A:

$$\widehat{a_m} = \arg\min_{a} -L(a_m|\mathcal{X}_T) + \lambda||a||_1$$

Statistical Learning Bounds

Two key ingredients needed for sample complexity bounds:

1. Deviation Bound: Let $\epsilon_{t,m} = X_{t+1,m} - \exp(\nu_m + \langle a_m, g(\mathcal{X}_t)\rangle)$. Need to find λ such that

$$\max_m \frac{1}{T} || \sum_{t=1}^{T} g(\mathcal{X}_t) \epsilon_{t,m} ||_\infty \leq \lambda$$

Statistical Learning Bounds

Two key ingredients needed for sample complexity bounds:

1. Deviation Bound: Let $\epsilon_{t,m} = X_{t+1,m} - \exp(\nu_m + \langle a_m, g(\mathcal{X}_t))$. Need to find λ such that

$$\max_m \frac{1}{T} || \sum_{t=1}^{T} g(\mathcal{X}_t) \epsilon_{t,m} ||_\infty \leq \lambda$$

2. Restricted Eigenvalue: The smallest eigenvalue of $\mathbb{E}[g(\mathcal{X}_t) g(\mathcal{X}_t)^T | \mathcal{X}_{t-1}]$ is lower bounded by $\omega > 0$. Strong dependencies make ω smaller.

Definition

$$\epsilon_{t,m} = X_{t,m} - \mathbb{E}[X_{t,m}|\mathcal{X}_{t-1}] = X_{t,m} - \exp(\nu_m + \langle a_m, g(\mathcal{X}_{t-1})\rangle)$$

- Commonly studied settings where noise is iid and subgaussian do not apply. Instead use martingale concentration inequalities to bound.

Deviation Bound

$\max_m ||\frac{1}{T}\sum_{t=1}^{T} \epsilon_{t,m} g(\mathcal{X}_{t-1})||_\infty \leq \frac{C\log^2(MT)}{\sqrt{T}}$ whp

Restricted Eigenvalue Condition

ω is a lower bound on eigenvalues of $\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^T|\mathcal{X}_{t-1}]$

- Show this equivalent to lower bound on

$$\text{Var}\ (\min(X_{t,m}, K)|\mathcal{X}_{t-1})$$

- Two worst case scenarios: $\lambda_{t,m} = \exp(\nu_m + \langle a_m, g(\mathcal{X}_{t-1})\rangle)$ is very small, or very large

Restricted Eigenvalue Condition

ω is a lower bound on eigenvalues of $\mathbb{E}[g(\mathcal{X}_t)g(\mathcal{X}_t)^T|\mathcal{X}_{t-1}]$

- If $\lambda_{t,m}$ small, variance can be bounded in terms of $||A^*||_\infty, K, \alpha$ (but independent of M, T)
- If $\lambda_{t,m}$ large, variance can be bounded in terms of a constant κ which is related to the fraction of observations that are clipped.

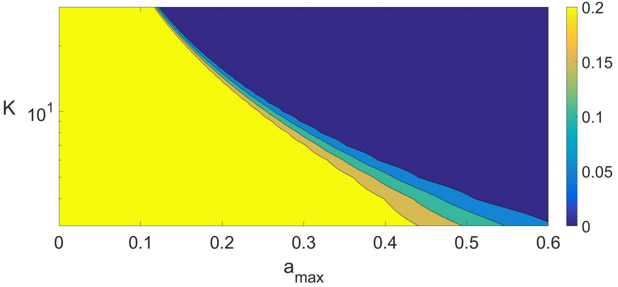


Figure 2: Values of κ for varying $||A^*||_\infty$ and K

Performance Guarantees

Theorem 1:

Suppose data is generated according to the clipped ARMA(1,1) model. Then:

$$||\widehat{A} - A^*||_F^2 \leq C \frac{R_{\text{max}}^2}{R_{\text{min}}^2 \min(\frac{1}{2} R_{\text{min}}, \kappa)^2} \frac{||A^*||_0 \log^4(MT)}{T}$$

whp for T sufficiently large.

Notation

- M: size of network
- T: number of time periods
- $R_{\text{min}}, R_{\text{max}}, \kappa$: Independent of M and T.

Key takeaway: bound scales well in $||A||_0 \ll T \ll M^2$ setting.

Simulations

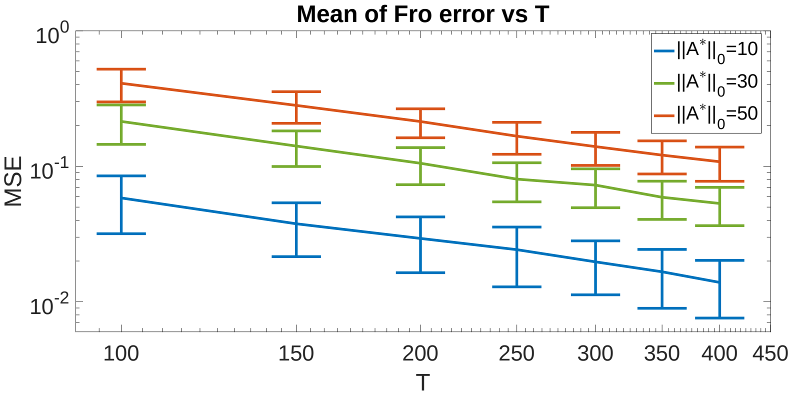


Figure 3: MSE vs. T. Median of 100 trials is shown, with error bars denoting 25th-75th percentiles. M=50, $\alpha = .25$.

Experiments

Can we identify geographic patters in criminal activity? [8]

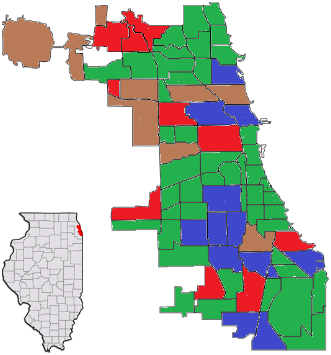


Figure 4: Spectral clustering of community areas in Chicago based on network learned from crime data. Half day time discretization period used with $\alpha = .2$. Log-likelihood of events on test set larger than for constant Poisson process.

[8] cf., Moher et al. (2014); Adams & Linderman (2014)

Conclusions

- The clipped ARMA(1,1) model incorporates saturation effects common in real-world systems.
- Performance guarantees applicable in high-dimensional and sparse setting.
- Lays groundwork for extensions to general autoregressive models or to different regularizers.

Thank You!