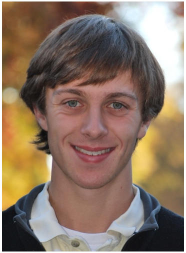
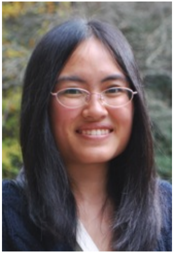
Sparsity and Scarcity in Discrete Event Data Analysis

Rebecca Willett, University of Wisconsin-Madison
Joint work with



Eric Hall

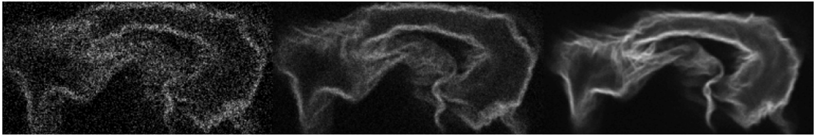


Xin Jiang

Paul Voyles

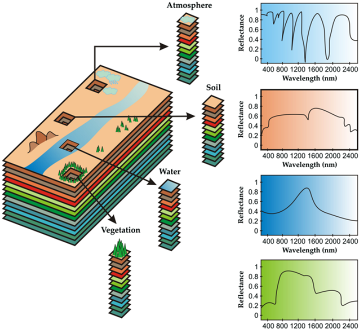
Garvesh Raskutti

Photon limited imaging



Fluorescence or electron microscopy

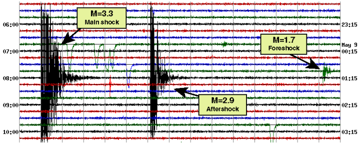
Spectral imaging

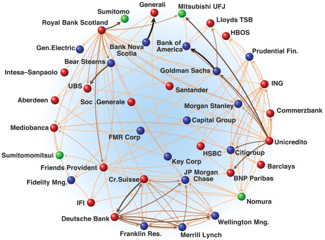


https://srogers.cartodb.com/viz/4a5eb582-23ed-11e4-bd6b-0e230854a1cb/embed_map

Discrete event data is prevalent

- Seismic shocks
- Financial transations
- Neurons firing
- Adverse drug events
- Crime

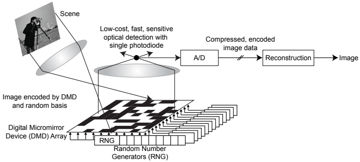


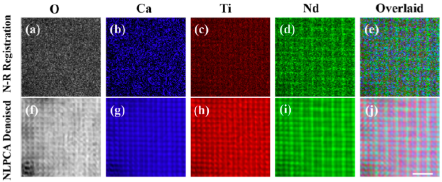


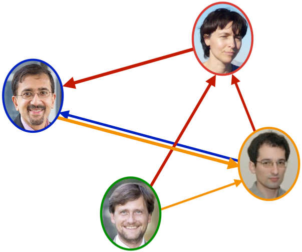
In all these settings, incorporating physical models into inference is essential

Today: three examples

- Electron microscopy in materials science
- Photon-limited compressive sensing
- Social and biological neural network inference



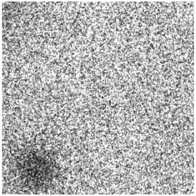




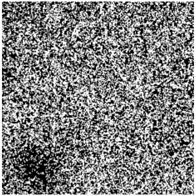
Example 1:
Electron microscopy in materials science

Case study: electron disperson spectroscopy imaging

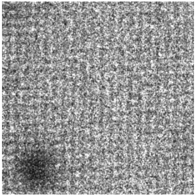
Calcium-doped neodymium titanate[1] (perovskite ceramic)



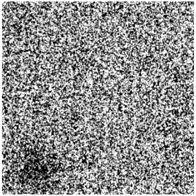
O Kα



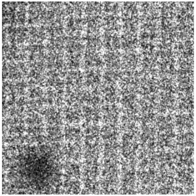
Ca Kα



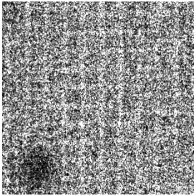
Ti Kα



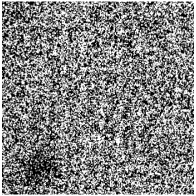
Ti Kβ



Nd Lα



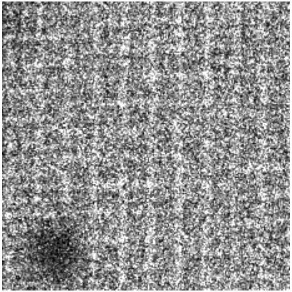
Nd Lβ



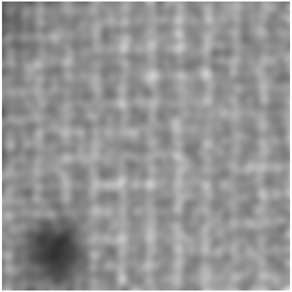
Nd Lβ4

[1] Raw data courtesy of Thomas Slater and Sarah Haigh at University of Manchester. Non-rigid alignment and averaging by Yankovich, Berkels, Dahmen, Binev, Sanchez, Bradley, Li, Szlufarska & Voyles (2014)

Naïve estimate



Nd Lα observations



Estimate via Gaussian smoothing

Can we do better?

Case study: EDS imaging

Long exposure times can damage samples

Short exposure times result in small numbers of detected photons per pixel. Statistical model:

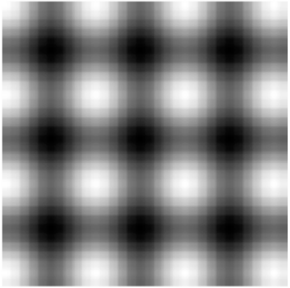
$$y \sim \mathrm{Poisson}(x^*),$$

where x^* is the spectral image and y is the noisy observation

Goal: estimate x^* from y using Poisson model for noise and structural models for x^*

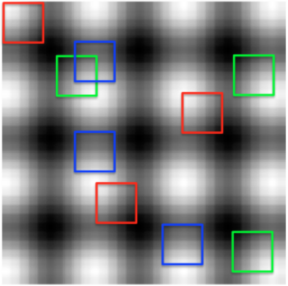
Spectral image model

Consider this phantom image:



Spectral image model

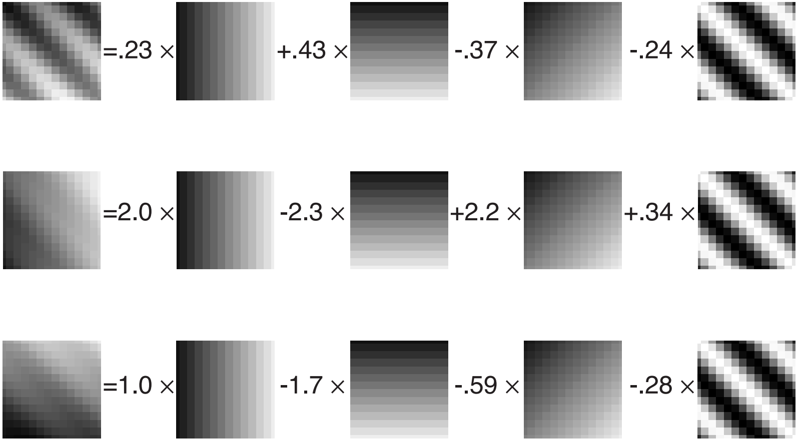
Consider this phantom image:



We want to exploit the redundancy in the image.

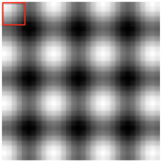
Key model idea: the (spectral) image patches lie in a union of subspaces

Patch subspaces

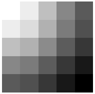


Each patch is a weighted sum of representative patches.

Image model

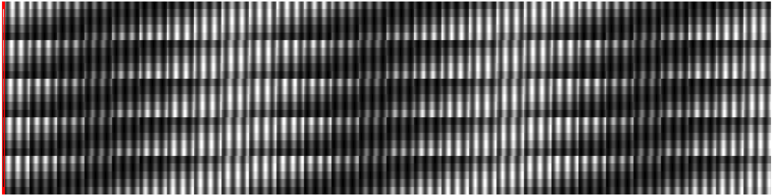


Image



Patch

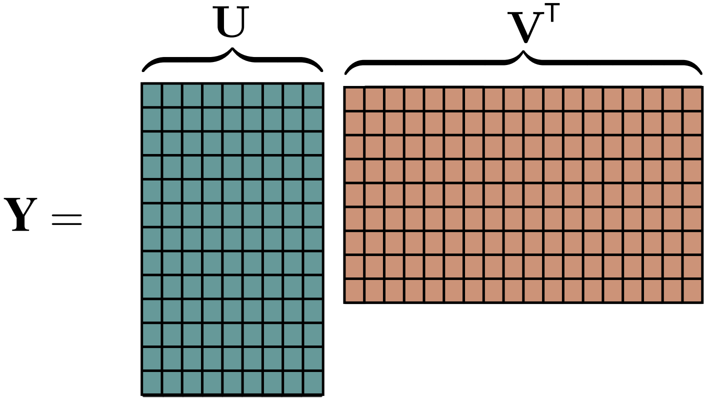
Vectorized Patch



Collection of patches

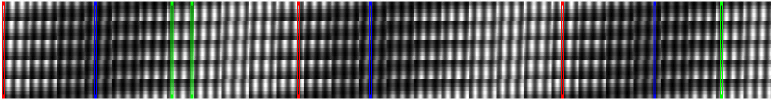
These ideas extend naturally to spectral images
(need to use 3-D patches)

Patch subspaces

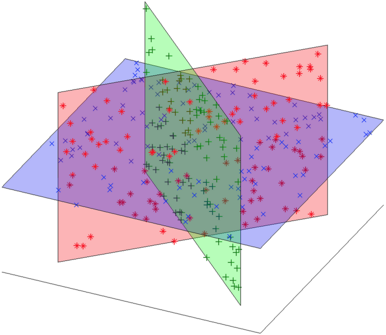


Each patch is a weighted sum of representative patches.

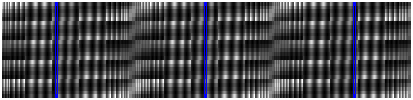
Image model



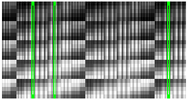
Collection of patches



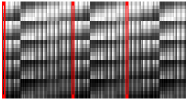
Union of subspaces



Cluster 1

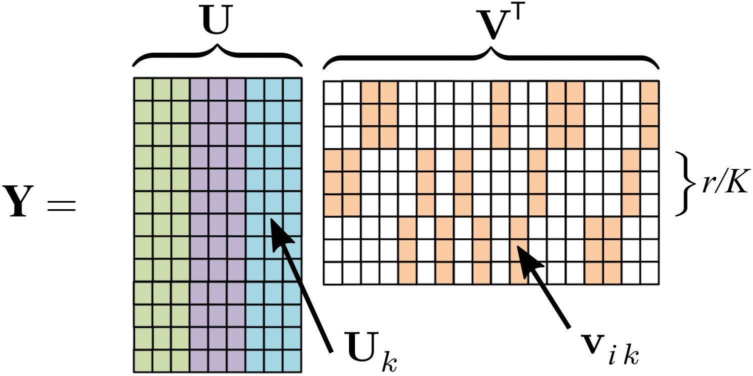


Cluster 2



Cluster 3

Mathematical model of union of subspaces



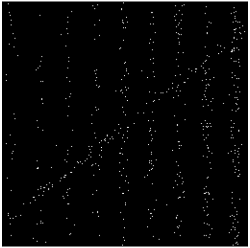
Matrix factorization associated with a union of subspaces model. The matrix U has $K = 3$ groups, each with $r/K = 3$ columns corresponding to three representative patches per group or nine total representative patches. U_k is the set of representative patches for the k^{th} group. $v_{i,k}$ is the set of weights for the i^{th} patch projected onto the k^{th} subspace. Note that each patch has nonzero weights for only *one* of the K subspaces.

Nonlocal PCA for photon-limited imaging[2]

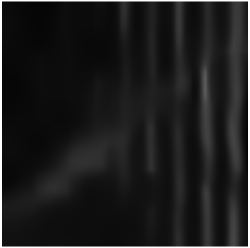
- Divide image into patches
- Cluster patches
  (using Poisson Bregman divergence to measure similarity of patches)
- Perform Poisson PCA on each cluster of patches to find low-dimensional patch subspace
  (by minimizing the negative Poisson log-likelihood with rank constraint)
- For each patch, estimate sparse PCA coefficients
  (by minimizing the negative Poisson log-likelihood + sparsity regularizer)

[2] Salmon, Deledalle, Harmany & Willett (2012)

Do not underestimate the power of the dark side



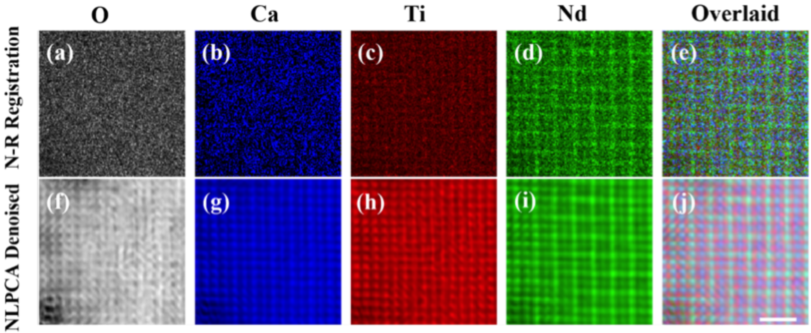
Original data



Anscombe + BM3D; PSNR = 18.99. Mäkitalo & Foi (2011)

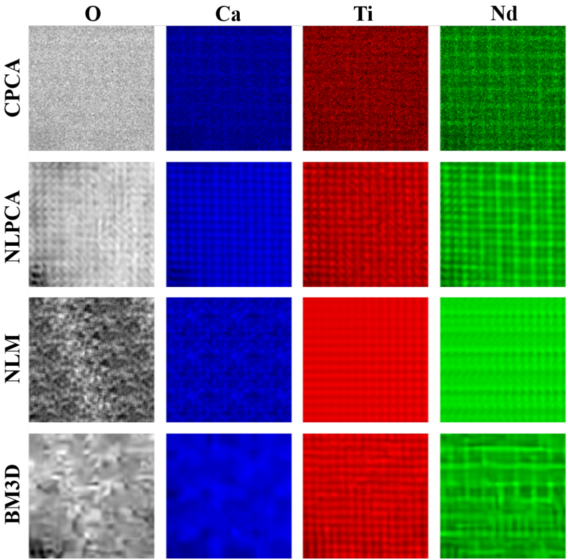
Poisson non-local PCA; PSNR = 23.27. Salmon, Deledalle, Harmany & Willett (2012)

EDS Imaging Experimental Results[3]



[3] Yankovich, Zhang, Oh, Slater, Azough, Freer, Haigh, Willett, and Voyles (2016)

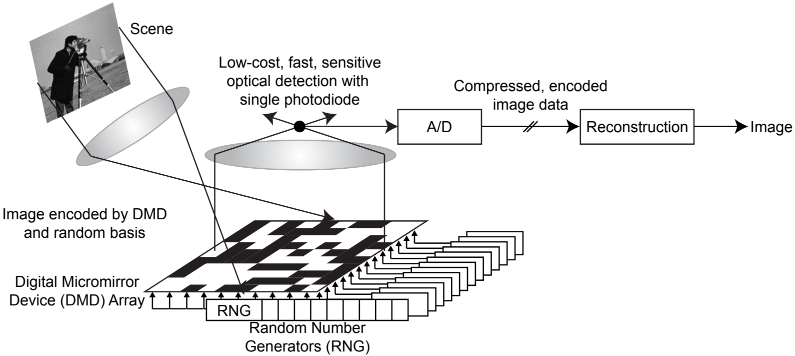
EDS Imaging Comparison[4]



[4] Yankovich, Zhang, Oh, Slater, Azough, Freer, Haigh, Willett, and Voyles (2016)

Example 2: Photon-limited compressive sensing

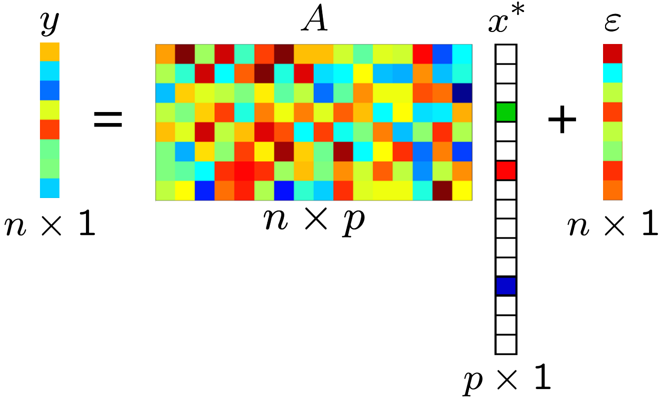
Compressive optical systems[5]



If we fix our total data acquisition time to T, then we have an explicit tradeoff between the number of projections, n, and the number of photons collected per projection, $O(T/n)$. As n increases, photon-limitations dominate errors.

[5] Duarte, Davenport, Laska, Sun, Takhar, Sarvotham, Baron, Wakin & Kelly, Baraniuk (2006)

The LASSO for sparse inverse problems



The LASSO estimator:

$$\min_x \frac{1}{2}\|y - Ax\|^2 + \gamma\|x\|_1$$

Sensing model

We observe

$$y \sim \text{Poisson}(TAx^*)$$

$$y_i \sim \text{Poisson}\left(T\sum_{j=1}^{p} A_{i,j}x_j^*\right), \qquad i \in \{1,\ldots,n\},$$

where

- $y \in \mathbb{Z}_+^n$
- $T \in \mathbb{R}_+$ is the total data acquisition time
- $A \in [0,1]^{n\times p}$ is a known sensing matrix
- $x^* \in \mathcal{X}$, where

$$\mathcal{X} = \left\{x \in \mathbb{R}_+^p \;:\; \|x\|_1 = 1,\; \|D^T x\|_0 \leq s+1\right\}$$

for an orthonormal basis D

This is not your ordinary CS problem

Sensing matrix A has several physical constraints

Think of $A_{i,j}$ as likelihood of photon from location j in x^* hitting detector at location i:

$$\begin{aligned} A_{i,j} &\in [0,1] & & \\ \mathbb{1}^\top A &\preceq \mathbb{1} & & \text{(columns sum to at most one)} \\ \|Ax\|_1 &\leq \|x\|_1 & & \forall x \end{aligned}$$

Typical CS sensing matrices do not satisfy these constraints!

Sensing matrix

Start with a sensing matrix $\widetilde{A} \in \frac{1}{\sqrt{n}}\{-1, 1\}^{n \times p}$ such that the product $\widetilde{A}D$ satisfies the RIP:

$$(1 - \delta_s)\|\theta\|_2^2 \leq \|\widetilde{A}D\theta\|_2^2 \leq (1 + \delta_s)\|\theta\|_2^2 \quad \forall \quad 2s - \text{sparse } \theta$$

Let

$$A \triangleq (\widetilde{A} + \frac{3}{\sqrt{n}}\mathbb{1}_{n \times p})/4\sqrt{n}.$$

Sensing matrix

Start with a sensing matrix $\widetilde{A} \in \frac{1}{\sqrt{n}}\{-1, 1\}^{n\times p}$ such that the product $\widetilde{A}D$ satisfies the RIP:

$$(1-\delta_s)\|\theta\|_2^2 \leq \|\widetilde{A}D\theta\|_2^2 \leq (1+\delta_s)\|\theta\|_2^2 \quad \forall \quad 2s - \text{sparse } \theta$$

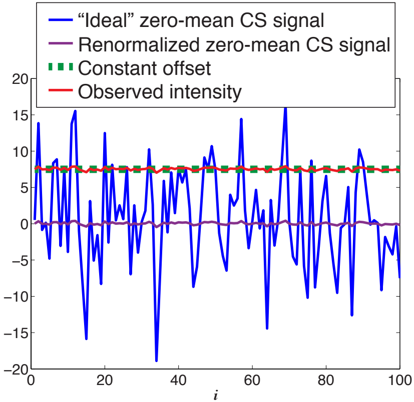
Let

$$A \triangleq (\widetilde{A} + \frac{3}{\sqrt{n}}\mathbb{1}_{n\times p})/4\sqrt{n}.$$

We observe

$$y \sim \text{Poisson}(TAx^*)$$

$$\sim \text{Poisson}\left(\frac{T\widetilde{A}x^*}{4\sqrt{n}} + \underbrace{\frac{3T}{4n}\mathbb{1}_{n\times 1}}_{\text{determines variance}}\right)$$



Rates for high-intensity settings (large T) [6]

Theorem:

$$\inf_{\widehat{x}} \sup_{x^* \in \mathcal{X}} \mathbb{E}[\|\widehat{x} - x^*\|_2^2] \asymp \frac{s \log p}{T}$$

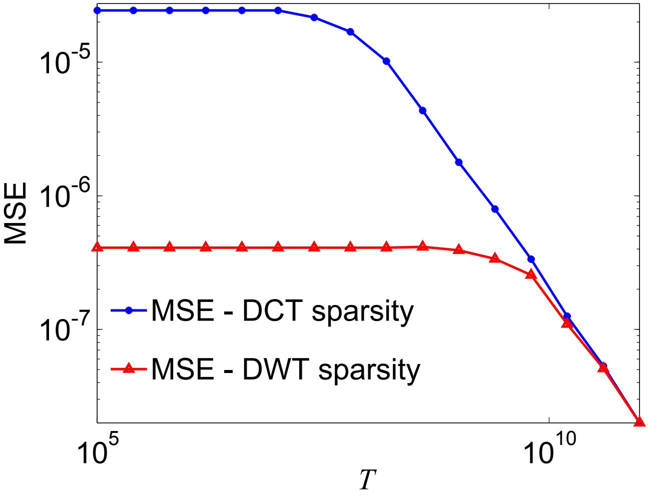
where

$$\mathcal{X} = \left\{ x \in \mathbb{R}_+^p \ : \ \|x\|_1 = 1, \ \|D^T x\|_0 \leq s + 1 \right\}$$

- The data acquisition time T, which reflects the signal-to-noise ratio, controls the error decay
- Once the number of measurements, n, is sufficiently large to ensure a RIP-like sensing matrix, it does not impact errors

[6] Jiang, Raskutti & Willett (2014)

MSE vs. T: An elbow in the rates



So far we have only considered high-intensity (large T) settings. What happens in low intensities?

Low-intensity settings (small T) [7]

Let $\overline{x^*} \equiv \mathbb{1}_{p\times 1}/\sqrt{p}$ be the average of x^*. Then

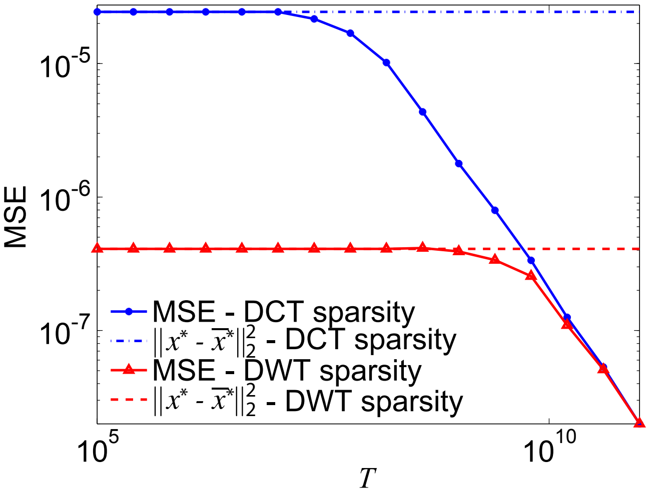
$$\mathbb{E}[\|\widehat{x} - x^*\|_2^2] \asymp \|x^* - \overline{x^*}\|_2^2$$

Rates depend on how much x^* deviates from its mean ("residual energy"), subject to the constraint that $\|x^*\|_1 = 1$ for $x^* \in \mathcal{X}$.

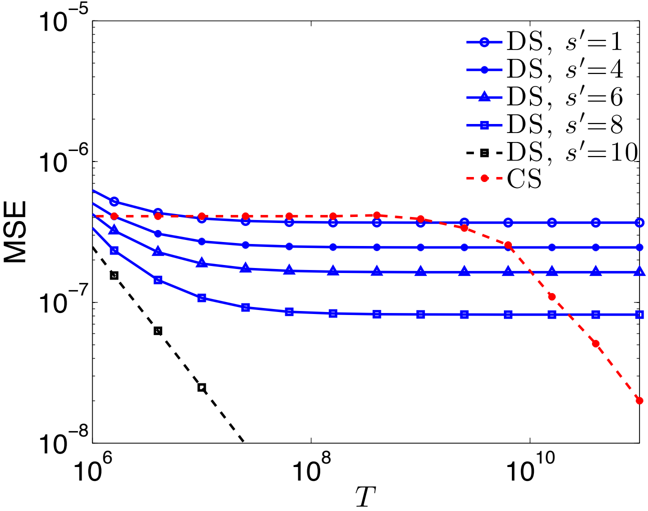
For different sparsifying bases D, this residual energy falls in different ranges, giving different rates.

[7] Jiang, Raskutti & Willett (2014)

MSE vs. T: An elbow in the rates



CS can be suboptimal at low intensities



$s' =$ number of non-zero coarse-scale wavelet coefficients, $s = 10$ is total number of non-zero wavelet coefficients.

Ramifications

CS conventional wisdom (for Gaussian noise settings) tells us rates are

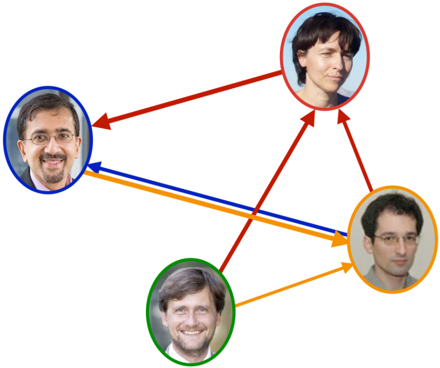
- Independent of sparsifying basis
- Not much worse than if we collected non-compressive measurements

In Poisson noise settings, because of the interaction between physical constraints and sparsity assumptions

- Rates are highly dependent on sparsifying basis
- Depending on the sparsity assumptions, we can do far better using non-compressive measurements

Example 3: Social and biological neural network inference

Cascading chains of interactions



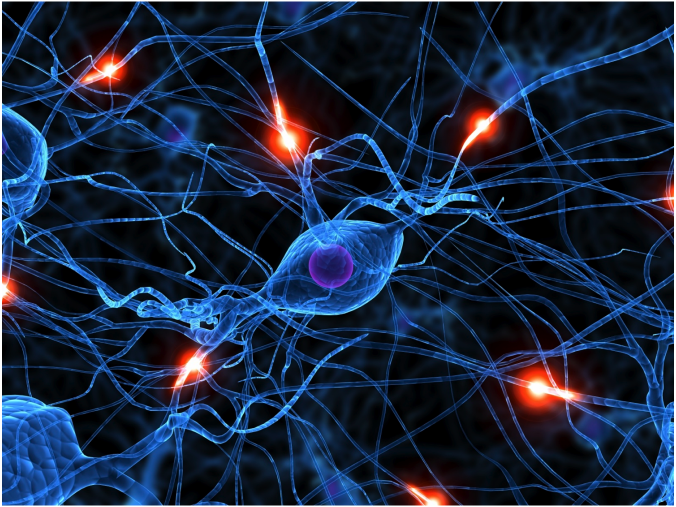
- Internet memes quickly propagate[a]
- Gang violence begets retaliations[b]
- Nation-state conflicts are accompanied by proxy wars[c]

---

[a] K. Zhou, H. Zha, and L. Song, 2013
[b] A. Stomakhin, M. B. Short, and A. Bertozzi, 2011
[c] C. Blundell, K. A. Heller, and J. M. Beck, 2012

Can we infer the underlying network of influences from observations of individual events?

Functional neural network connectivity[8]

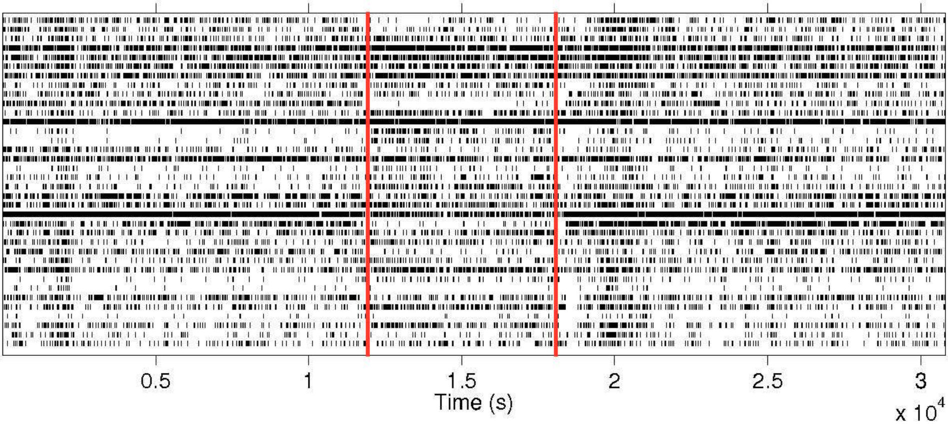


We record neurons firing in response to different stimuli.

Can we estimate the functional network?

[8] Smith & Brown, 2003; Pillow, Shlens, Paninski, Sher, Litke, Chichilnisky, & Simoncelli. 2008

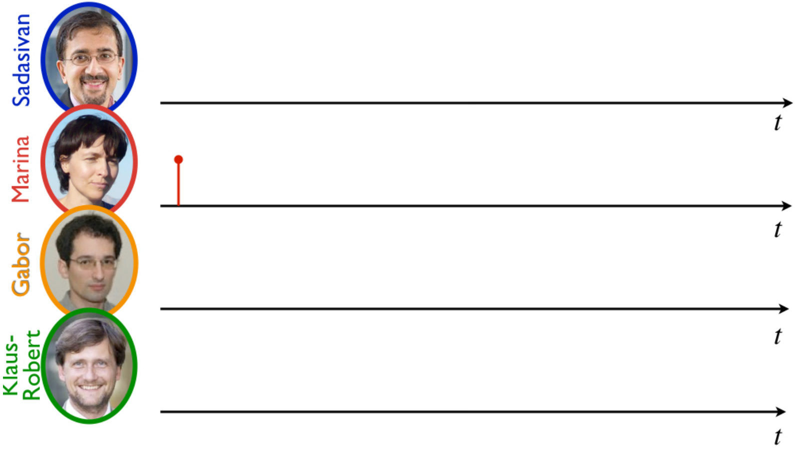
Functional neural network connectivity



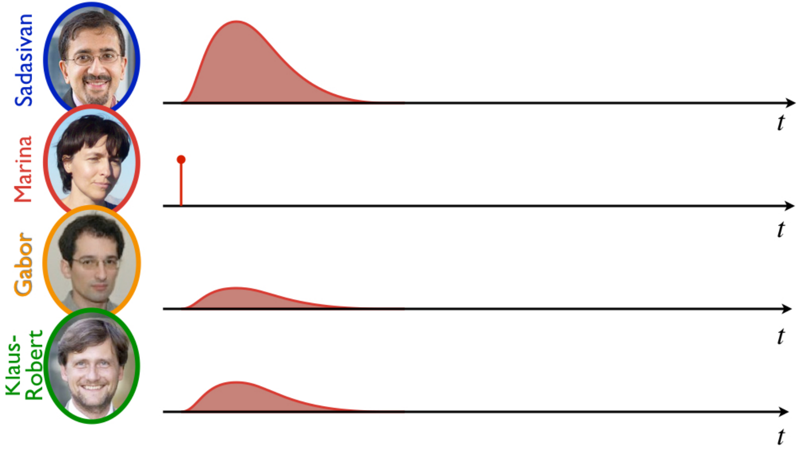
Raster plot of $M = 36$ spike trains. Each row corresponds to a spike train, with small, black, vertical lines indicating the time of individual spikes. The vertical red lines indicate the start and end of a maze exploration period.[9]

[9] http://seis.bris.ac.uk/~mb0184/projects/dtsonn/

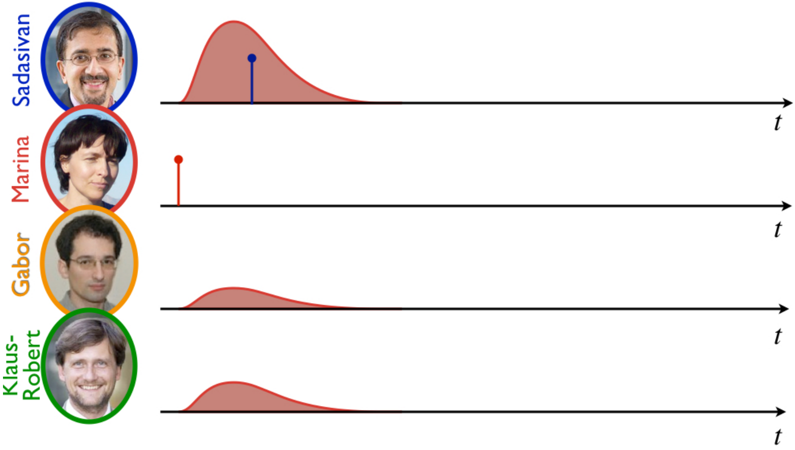
Autoregressive point processes



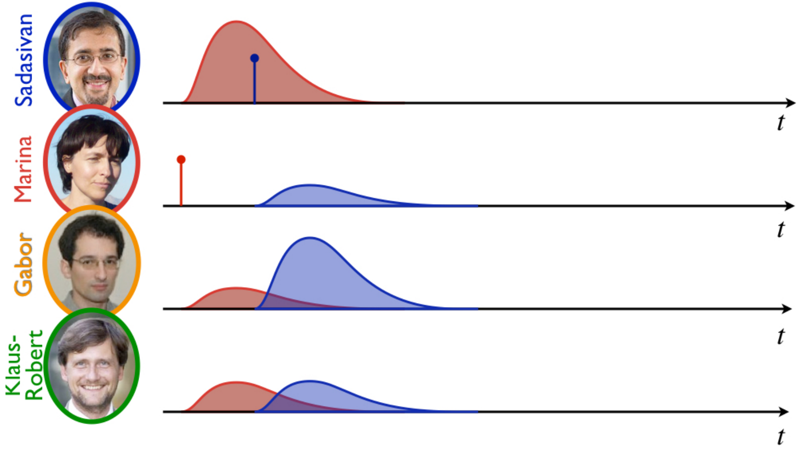
Autoregressive point processes



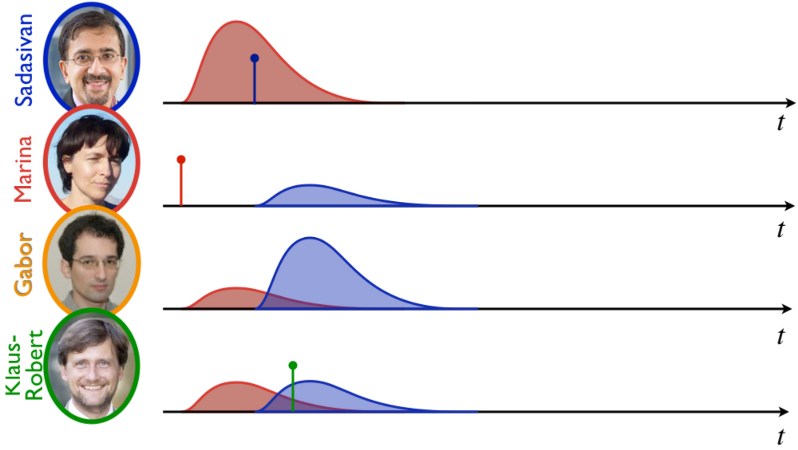
Autoregressive point processes



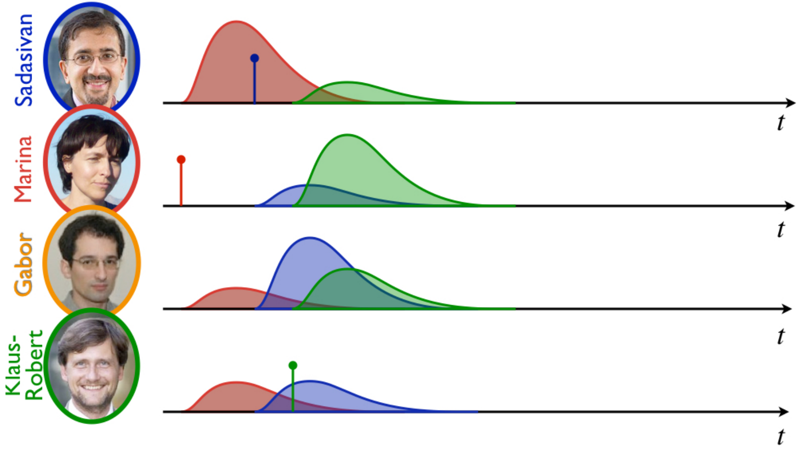
Autoregressive point processes



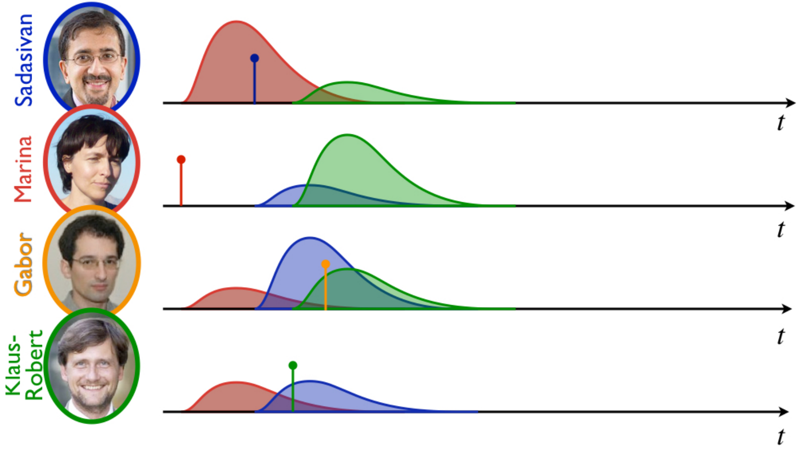
Autoregressive point processes



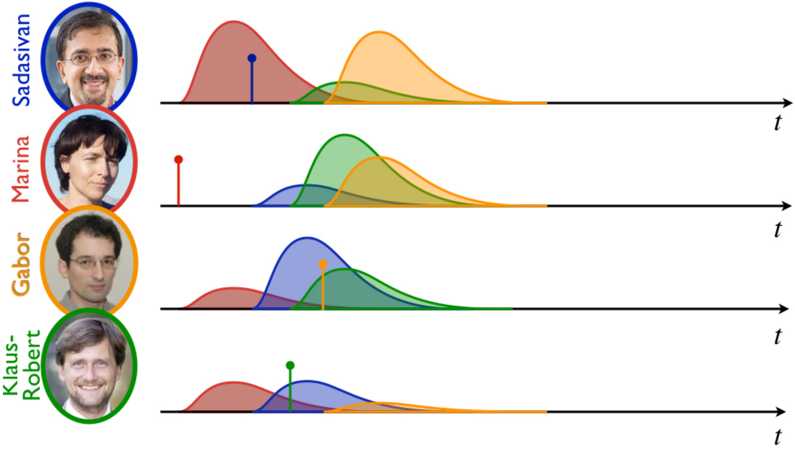
Autoregressive point processes



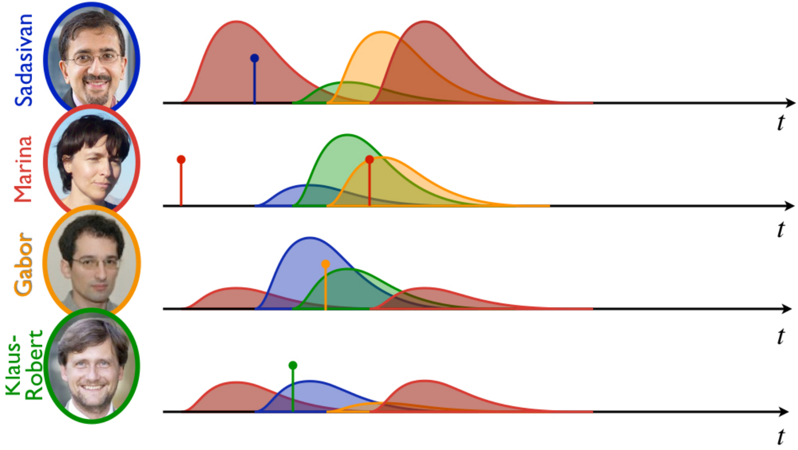
Autoregressive point processes



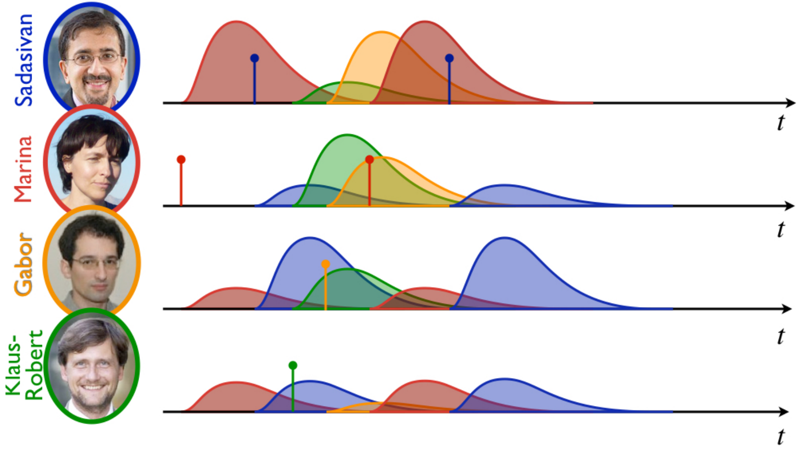
Autoregressive point processes



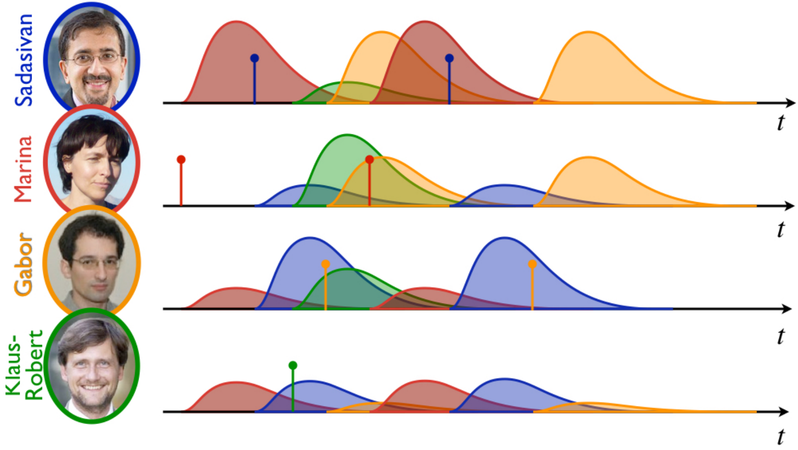
Autoregressive point processes



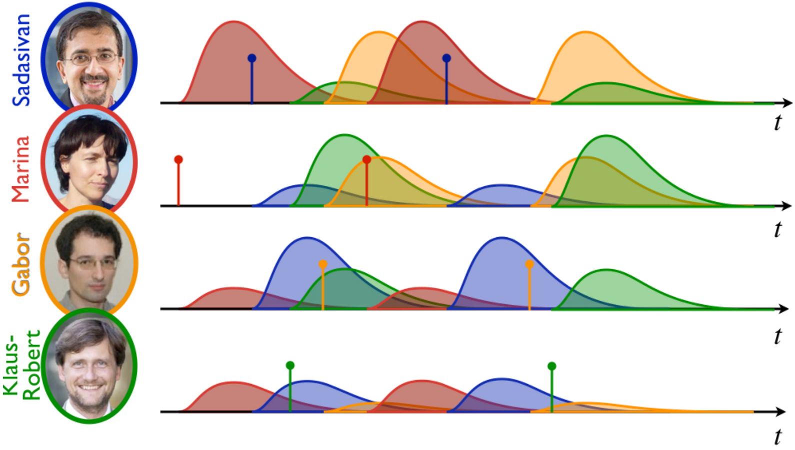
Autoregressive point processes



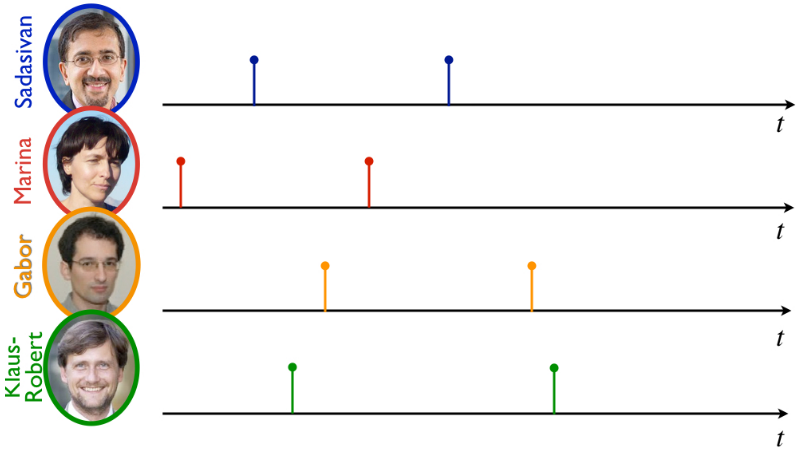
Autoregressive point processes

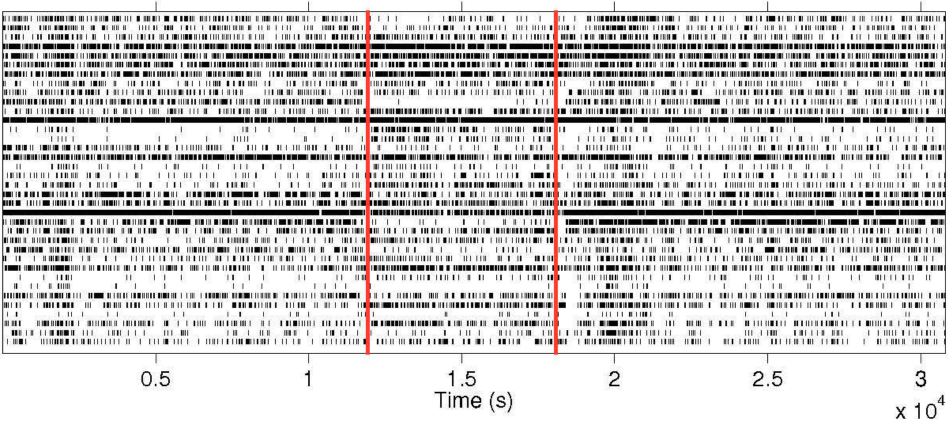


Autoregressive point processes



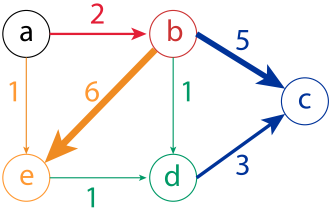
Autoregressive point processes



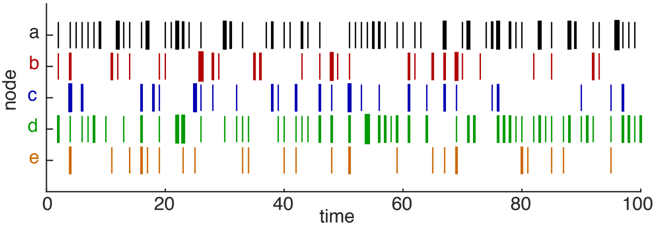

0.5
1
1.5
2
2.5
3
Time (s)
x 10^4

Log-linear Poisson autoregressive process

$$x_{t+1} \sim \text{Poisson}\left(\exp\{\nu - A^* x_t\}\right), \qquad t = 1, \ldots, T$$

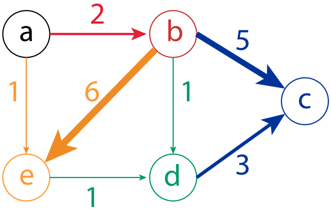


$$A^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 3 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 6 & 0 & 0 & 0 \end{bmatrix}$$



Log-linear Poisson autoregressive process

$$x_{t+1} \sim \text{Poisson}\left(\exp\{\nu - A^* x_t\}\right), \qquad t = 1, \ldots, T$$



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

How should we estimate A^*?

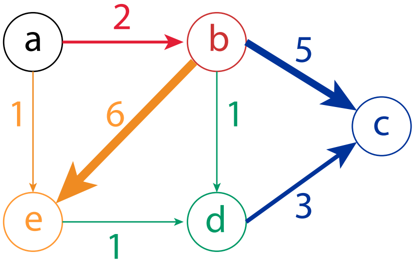
- How much sensing time is required for a desired level of accuracy?
- How do network properties influence achievable accuracy?

Sparsity

- We assume A^* is non-negative and bounded:

$$A^* \in [0, A_{\max}]^{M \times M}$$

- s is the number of non-zero elements in A^* (e.g. number of network edges)
- ρ is the maximum number of non-zero elements in any row A^* (e.g. maximum in-degree of any node)



$$M = 5,\ s = 7,\ \rho = 2$$

Autoregressive challenges

Let $y_m = \begin{bmatrix} X_{1,m} & X_{2,m} & \cdots & X_{T,m} \end{bmatrix}^\top$ be the time series of observations at the m^{th} node and let

$$\boldsymbol{X}_{\setminus m} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,m-1} & X_{1,m+1} & \cdots & X_{1,M} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,m-1} & X_{2,m+1} & \cdots & X_{2,M} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{T,1} & X_{T,2} & \cdots & X_{T,m-1} & X_{T,m+1} & \cdots & X_{T,M} \end{bmatrix}$$

be the observations at all other nodes, corresponding to the potential influences on node m.

Consider estimating each row a_m^* via the LASSO:

$$\hat{a}_m = \arg\min_a \| \underbrace{y_m}_{\text{node } m \text{ obs}} - \underbrace{\boldsymbol{X}_{\setminus m}}_{\text{node } m \text{ inputs}} a\|_2^2 + \lambda \|a\|_1$$

Regularized maximum likelihood estimator

Let a_m be the m^{th} row of A and

$$\hat{A} := \arg\min_{A} \underbrace{\sum_{t=0}^{T-1} \sum_{m=1}^{M} \exp(\nu_m - \langle a_m, x_t \rangle) - \langle a_m, x_t \rangle x_{t+1,m}}_{\text{negative log-likelihood}} + \underbrace{\lambda \|A\|_{1,1}}_{\text{regularizer}}$$

or, row-wise

$$\hat{a}_m := \arg\min_{a} \underbrace{\sum_{t=0}^{T-1} \sum_{m=1}^{M} \exp(\nu_m - \langle a, x_t \rangle) - \langle a, x_t \rangle x_{t+1,m}}_{\text{negative log-likelihood}} + \underbrace{\lambda \|a\|_{1}}_{\text{regularizer}}$$

Main result (sample complexity bound)

If

$$T \gtrsim \rho^2 s \log M$$

and

$$\lambda \approx \frac{\log^2(MT)}{\sqrt{T}},$$

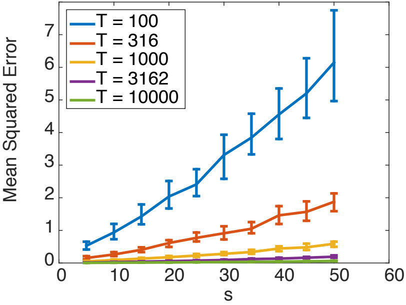
then with probability at least $1 - 1/M$

$$\|\hat{A} - A^*\|_F^2 \leq O\left(\frac{e^{\rho} s \log^6(MT)}{T}\right).$$

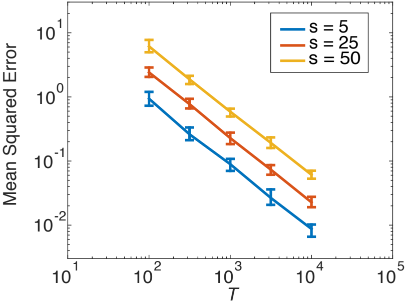
Theory does not depend on all observations coming from stationary distribution!

For constant ρ, error grows linearly in s, but only polylogarithmically in M, showing the benefit of sparsity

Up to log factors, error decreases like $1/T$, which will dictate how much data needs to be collected for a desired accuracy



MSE vs s



MSE vs. T

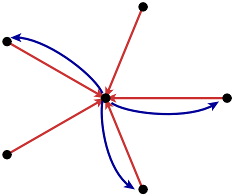
In both plots the median value of 100 trials is shown, with error bars denoting the $25^{\mathrm{th}} - 75^{\mathrm{th}}$ percentile. $M = 20$, 100 trials, $\lambda = 0.1/\sqrt{T}$.

The role of ρ

Our bounds scale like e^{ρ}. **Is this tight?**

Recall that ρ is the maximum in-degree of any node in the network.
If ρ is large, many nodes can simultaneously inhibit a single node.

Example: star network



$$A = \begin{bmatrix} 0 & \mathbf{1} & \mathbf{1} & \dots & \mathbf{1} \\ \mathbf{1} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & 0 & 0 & \dots & 0 \end{bmatrix}$$

Here

$$\mathbb{E}[x_{t+1}] = \exp(\nu - Ax_t) \approx [0, \nu_2, \nu_3, ..., \nu_M]^\top$$

rare events on center node $\Rightarrow$ cannot infer inhibitions
$\Rightarrow$ large errors

Proof elements

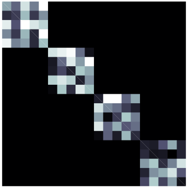
Lasso analysis

1. Bound $\|X^\top \epsilon\|_\infty$; require λ to exceed this bound
2. Show $\|X(a^* - \hat{a})\|_2^2 \lesssim \lambda\sqrt{s}\|a^* - \hat{a}\|_2$
3. Show/assume $\|X(a^* - \hat{a})\|_2^2 \geq \kappa\|a^* - \hat{a}\|_2^2$
4. Algebra: $\|a^* - \hat{a}\|_2 \lesssim \frac{\lambda\sqrt{s}}{\kappa}$

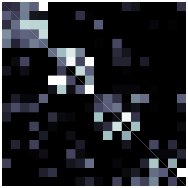
Sparse PAR analysis

1. $\left\|\frac{1}{T}\sum_{t=1}^{T-1} x_{t-1}\epsilon_{t,m}\right\|_\infty \leq \frac{C\log^3(MT)}{\sqrt{T}} \leq \lambda$
2. Show $\|a_m^* - \hat{a}_m\|_T^2 \triangleq \frac{1}{T}\sum_t \langle a_m^* - \hat{a}_m, x_t\rangle^2 \leq \lambda\sqrt{\rho_m}\|a_m^* - \hat{a}_m\|_2$
3. Lower bound $\omega =$ minimum eigenvalue of $\mathbb{E}[x_t x_t^\top | x_{t-1}]$ to show $\|a_m^* - \hat{a}_m\|_2^2 \leq \max\{\|a_m^* - \hat{a}_m\|_2^2, \|a_m^* - \hat{a}_m\|_T^2\} \leq \frac{\rho_m\lambda^2}{\omega^2}$

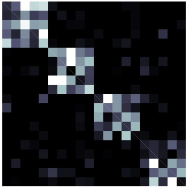
Experimental results



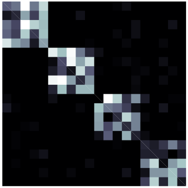
Ground Truth A Matrix



Estimate for $T = 100$



Estimate for $T = 316$

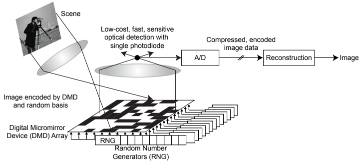


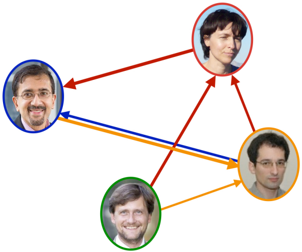
Estimate for $T = 1000$

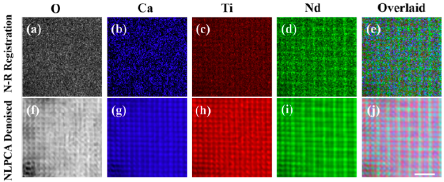
Even for a relatively low amount of data we have picked out most of the support but with several spurious artifacts. As the amount of data increases, fewer of the erroneous elements are estimated.

Conclusions

- Principled mechanisms for analyzing discrete event data arising in real physical systems
- Results provide new insights into how different sensor or network characteristics influence sample complexities and recovery guarantees
- Interesting open questions remain!



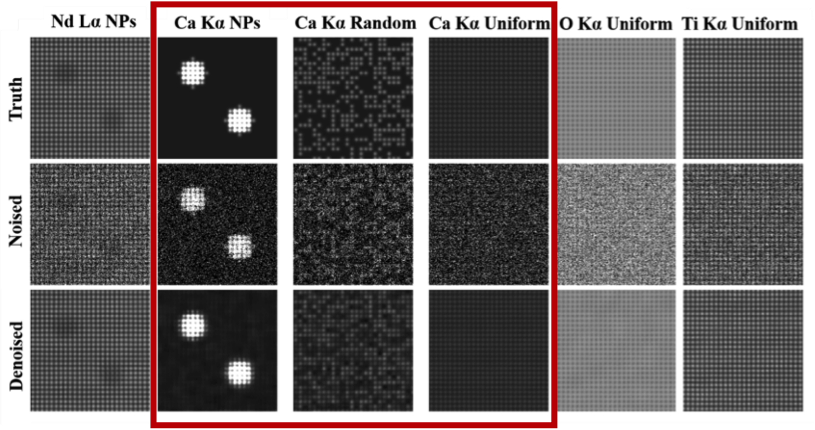




Thank you.

Backup slides

EDS Imaging Phantom Experiment[10]



10 Yankovich, Zhang, Oh, Slater, Azough, Freer, Haigh, Willett, and Voyles (2016)

Main result (sample complexity bound)

Let $\delta \in (0,1)$. There exist constants $c_1, c_2 > 0$ independent of M, δ, ρ and s such that if

$$T > c_1 \max\left\{\rho^2[s \log M + \log(1/\delta)], 1/(M\delta)\right\}$$

and

$$\lambda = \frac{c_2 \log^2(MT)}{\sqrt{T}},$$

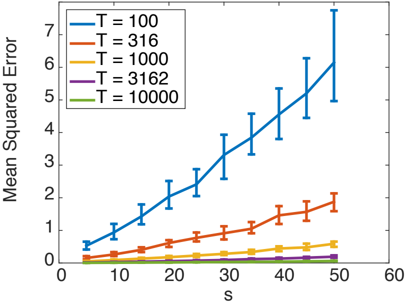
then with probability at least $1 - \delta$

$$\|\hat{A} - A^*\|_F^2 \leq O\left(\frac{e^\rho s \log^6(MT)}{T}\right).$$

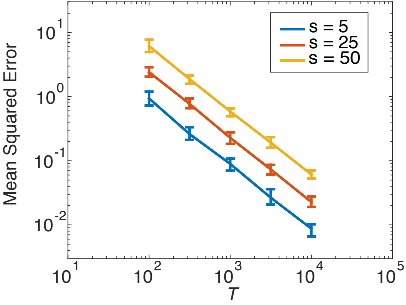
Theory does not depend on all observations coming from stationary distribution!

For constant ρ, error grows linearly in s, but only polylogarithmically in M, showing the benefit of sparsity

Up to log factors, error decreases like $1/T$, which will dictate how much data needs to be collected for a desired accuracy



MSE vs s



MSE vs. T

In both plots the median value of 100 trials is shown, with error bars denoting the $25^{\text{th}} - 75^{\text{th}}$ percentile. $M = 20$, 100 trials, $\lambda = 0.1/\sqrt{T}$.

Beyond inhibitory interactions

We needed to bound two key terms in our analysis:

$$\left\| \sum_t x_{t-1} \left(x_{t,m} - \mathbb{E}[x_{t,m} | x_{t-1}] \right) \right\|_\infty \qquad \text{and} \qquad \mathbb{E}[x_t x_t^\top | x_{t-1}]$$

This is tractable when all the elements of A^* are non-negative (i.e. inhibitory interactions).

Can theory admit stimulatory interactions? Challenge is that processes become non-stationary and observations unbounded.