INFERRING HIGH-DIMENSIONAL POISSON AUTOREGRESSIVE MODELS

Eric C. Hall[1], *Garvesh Raskutti*[1,2,3], *and Rebecca Willett*[1,3,4]

[1]Wisconsin Institute of Discovery, Optimization
[2]Department of Statistics, University of Wisconsin-Madison
[3]Department of Computer Science, University of Wisconsin-Madison
[4]Department of Electrical and Computer Engineering, University of Wisconsin-Madison

ABSTRACT

Consider observing a series of events associated with a group of interacting nodes in a network, where the interactions among those nodes govern the likelihood of future events. Such data are common in spike trains recorded from biological neural networks, interactions within a social network, and pricing changes within financial networks. Vector autoregressive point processes accurately model these settings and are widely used in practice. This paper addresses the inference of the network structure and autoregressive parameters from such data. A sparsity-regularized maximum likelihood estimator is proposed for a Poisson autoregressive process. While sparsity-regularization is well-studied in the statistics and machine learning communities, common assumptions from that literature are difficult to verify here because of correlations and heteroscedasticity inherent in the problem. Novel performance guarantees characterize how much data must be collected to ensure reliable inference depending on the size and sparsity of the autoregressive parameters, and these bounds are supported by several simulation studies.

Index Terms— Autoregressive Processes, Poisson Point Processes, Sample Complexity, Statistical Learning

1. INTRODUCTION

Time series count data arise in a variety of applications (*cf.* [1, 2, 3]), one of the most notable being biological neural networks [4, 5, 6, 7, 8]. In this application, we record the times at which each neuron in the network fires or "spikes" and wish to infer the structure of the underlying network. Action potentials or neuron spikes can trigger or inhibit spikes in connected neurons, so understanding excitation and inhibition among neurons provides key insight into the structure and operation of the neural network. A central question in the design of this experiment is *"for how long must I collect data before I can be confident that my inference of the neural network is accurate?"* Clearly the answer to this question will depend not only on the number of neurons being recorded, but also on what we may assume *a priori* about the network. Unfortunately, existing statistical and machine learning theory does not address this problem, which is the focus of this paper.

This example of a biological neural network can be modeled as an auto-regressive point process. That is, at each time t, we observe a high-dimensional vector of counts, and the distribution of those counts depends on previous observations. Inferring these dependencies is a key challenge in many settings because a precise understanding of these dependencies facilitates more accurate predictions and interpretable models of the forces that determine the distribution of each new observation.

This paper focuses on multivariate settings, particularly where the vector observed at each time is high-dimensional relative to the duration of the time series. We conduct a detailed investigation of a particular time series data model: the *vector log-linear Poisson autoregressive* (PAR) model. The PAR model has been explicitly studied in [9, 10, 11], is closely related to the discrete-time INGARCH model [12, 13], and can be considered a discretized version of the continuous-time Hawkes point process model [14, 15]. We focus specifically on estimating the parameters of a vector PAR model from a time series of count data by using a regularized maximum likelihood estimation approach that generalizes past work on Poisson inverse problems (*cf.* [16, 17, 18]). While similar algorithms have been proposed in the above-mentioned PAR literature, little is known about their *sample complexity* or *how inference accuracy scales with the key parameters such as the size of the network, the time spent collecting observations, and the density of edges within the network or dependencies among entities.* The temporal dependence among events can make such analyses particularly challenging and beyond the scope of much current research in high-dimensional statistical inference (see [19] for an overview).

That said, there has been a large body of work providing theoretical results for certain high-dimensional models under low-dimensional structural constraints (see *e.g.,*[18, 20, 21, 22, 23, 24, 25]). The majority of prior work has focussed on the setting where samples are independent and/or follow a Gaussian distribution; this work exploits many properties of linear systems and Gaussian random variables that can not be applied to non-Gaussian and non-linear auto-regressive models. In the Poisson auto-regressive model, we have dependent count data samples and signal-dependent Poisson noise. [18, 20, 26] provide results for non-Gaussian noise but still rely on independent observations. Another method [27] studies a general framework for point processes (including the Hawkes process) and provides estimation bounds for a LASSO-type estimator. Our work emphasizes the high-dimensional setting and bounds for short-duration time series.

In this paper, we develop performance guarantees for the vector PAR model that provide sample complexity guarantees in the high-dimensional setting under low-dimensional structural assumptions such as sparsity of the underlying auto-regressive parameter matrix. In particular, our main contribution is the derivation of squared-Frobenius-error bounds on the proposed estimator as a function of the problem dimension, sparsity, and the number of observations in time.

We gratefully acknowledge the support of awards NSF CCF-1418976, NSF IIS-1447449, NIH 1 U54 AI117924-01 and NSF DMS-1407028.

978-1-4673-7802-4/16/$31.00 ©2016 IEEE

2. PROBLEM FORMULATION

We consider the log-linear vector Poisson autoregressive model:

$$X_{t+1}|X_t \sim \text{Poisson}(e^{\nu - A^* X_t}), \tag{1}$$

where $(X_t)_{t=0}^T$ are M-variate observation vectors, $A^* \in [0, A_{\max}]^{M\times M}$ is some unknown parameter matrix, and $\nu \in [\nu_{\min}, \nu_{\max}]^M$ is a known rate parameter[1]. A similar model appears in [11], but that work focuses on maximum likelihood and weighted least squares estimators in univariate settings that are known to perform poorly in high-dimensional settings (as is our focus).

Under this model, the conditional likelihood can be expressed explicitly as:

$$\mathbb{P}(X_{t+1} = y|X_t = x, A) = \prod_{m=1}^{M} \frac{\exp(-e^{\nu_m - A_m^\top x}) e^{(\nu_m - A_m^\top x) y_m}}{y_m!},$$

where x and y are M-variate vectors and $A_m^\top$ is the m^{th} row of a candidate parameter matrix A.

In general, we observe T samples $(X_t)_{t=1}^T$ and our goal is to infer the matrix A^*. In the setting where M is large, we need to impose additional assumptions on A^* in order to have strong performance guarantees. In particular we assume that the matrix A^* is s-sparse, meaning that A^* belongs to the following class:

$$\mathcal{M}_S = \{A \in [0, A_{\max}]^{M\times M} \mid \sum_{\ell=1}^{M}\sum_{m=1}^{M} \mathbf{1}(|A_{\ell,m}| \neq 0) \leq s\}.$$

Finding an optimal estimator within this parameter class would use an ℓ_0 penalty to ensure that the estimator has at most s non-zero entries. However, this is a difficult optimization problem due to the non-convexity of the ℓ_0 function. Therefore, we instead find an estimator using the element-wise ℓ_1 decomposable regularizer, the convex relaxation of the ℓ_0 function, along with the negative log likelihood (using the known, constant vector ν):

$$\begin{aligned}\ell(A) =& \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\left(e^{\nu_m - A_m^\top X_t} + A_m^\top X_t X_{m,t+1}\right)\\ \widehat{A} =& \underset{A\in[0,A_{\max}]^{M\times M}}{\arg\min}\ \ell(A) + \lambda\|A\|_{1,1}\end{aligned} \tag{2}$$

where $\|\cdot\|_{1,1}$ is the element-wise ℓ_1 norm:

$$\|A\|_{1,1} = \sum_{\ell=1}^{M}\sum_{m=1}^{M}|A_{\ell,m}|.$$

The above is the regularized maximum likelihood estimator (RMLE), which attempts to find an estimate of A which both fits the data and has many zero valued elements. The goal is to derive bounds for $\|\widehat{A} - A^*\|_F^2$, the difference between the regularized maximum likelihood estimator, $\widehat{A}$, and the true generating network, A^* in Frobenius norm, under the assumption that the true network is sparse, which is our main theorem. As we will see in the main theorem, a crucial quantity in the regret bounds is the value ρ which is the maximum number of non-zeros in any row of A^*. If ρ is considered a constant independent of M and s, then good error bounds can be determined, but if ρ is on the order of s, meaning there's a single, dense row of non-zeros, poor error rates will be observed.

[1] Our framework easily extends to unknown ν but the notation is cumbersome; we focus on known ν for simplicity of presentation.

Theorem 1. *Let $\widehat{A}$ be the RMLE as in Equation 2, and assume $\lambda \geq \|\frac{2}{T}\sum_{t=0}^{T-1}(X_{t+1} - e^{\nu - A^* X_t})X_t^\top\|_{\infty,\infty}$ where $\|.\|_{\infty,\infty}$ is the element-wise ∞-norm then,*

$$\|\widehat{A} - A^*\|_F^2 \leq O(e^\rho s\lambda^2)$$

with probability at least $1 - 2\exp(-\min\left(c_3 T/\rho^2 - c_4\rho\log(2M), c_2 MT\right))$ where c_2, c_3 and c_4 are independent of M, T, ρ and s. Further,

$$\left\|\frac{2}{T}\sum_{t=0}^{T-1}(X_{t+1} - e^{\nu - A^* X_t})X_t^\top\right\|_{\infty,\infty} \leq \frac{8C_1^2 e^{\nu_{\max}}\log^3(MT)}{\sqrt{T}}$$

with high probability yielding the overall error rate of

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

with probability at least $1 - \exp(-c_6\min\left(T/\rho^2 - s\log(M), \log(MT)\right))$ for some c_6 independent of M, T, ρ and s.

The first interesting fact about the bound comes from looking at the high probability statement, we see that we require $T \geq \rho^3\log(M)$. If ρ grows slowly with increasing M, this tells us that T needs to be on the order of $\log(M)$, which is significantly less than the total M^2 parameters which are being estimated, and therefore including the sparsity assumption has lead to a significant gain. The next fact is that the theorem provides guidance in the setting of the regularization parameter. We see that we would like to set λ generally as small as possible, since the error ends up scaling approximately like λ^2, but that it also needs to be at least as large as $\tilde{O}(T^{-1/2})$ for the bounds to hold. The balance between setting λ small enough to have low error, while maintaining that it is large enough is an equivalent argument to needing to set λ large enough for it to take effect, but not too large to cause over-smoothing. The next important fact about the error is that it scales like T^{-1}, which tells us how long we need to observe the process in order to achieve a desired level of accuracy. Finally, we notice that the error scales linearly with the sparsity level s but only logarithmically with the dimension M in order to estimate M^2 parameters. This fact enforces the idea that doing inference in sparse settings can greatly reduce the needed amount of sensing time especially when $s \ll M^2$.

The exponential scaling with the maximum number of non-zeros in a row, ρ, at first seems unsatisfying. However, we can imagine a worst-case scenario where a large ρ compared to s and M would actually lead to very poor estimation. Consider the case of a large star-shaped network, where every node in the network influences and is influenced by a single node, and there are no other edges in the network. This would correspond to a matrix with a single, dense row and corresponding column. Therefore we would have $\rho = s/2 = M - 1$. In this network $M-1$ of the observations would be drawn independently and identically as Poisson random variables at every time with mean ν, but the central node of the network would be constantly inhibited, almost completely. In a large network, it would be very difficult to know if this inhibition was coming from a few strong connections or from the cumulative effect of all the inhibitions. Additionally, since the central node would almost never have a positive count, it would also be difficult to learn about the influence that node has on the rest of the network. Because of networks like this, it is important that not only is the overall network sparse, but each row also needs to be sparse. This requirement might seem inhibitive, but it has been shown in many real world networks

that the degree of a node in the network follows a power-law which is independent of the overall size of the network [28], and ρ would grow slowly with growing M.

3. SKETCH PROOF OF MAIN THEOREM

We first list the important assumptions and lemmas about the process that we need in order to prove the error bounds. It is important in the derivation of the bounds that observations are bounded by a term that grows logarithmically in T, and additionally that a large fraction of the data is bounded by some constant. Intuitively, these bounds are important because if the elements of X_t are very large, then the conditional expectation $\exp(\nu - A^* X_t)$ will be very small. If X_{t+1} is being drawn from a Poisson distribution with mean close to zero, then we do not learn anything about the network at that time. While we can allow some time counts to be very large, it becomes a problem when it becomes too frequent. It is these boundedness requirements that require that the values in A^* to be non-negative, otherwise we could have an unstable process which could grow very large at times.

Lemma 1. *There exists constants C_1 and c_1 which depend on the value $\nu_{\max}$, but are independent of T and M, such that $0 \leq X_{t,m} \leq C_1 \log(MT)$ with probability at least $1 - e^{-c_1 \log(MT)}$ for all $1 \leq t \leq T$ and $1 \leq m \leq M$. Additionally, for any $\alpha \in (0,1)$ such that αMT is an integer, there exist constants C_2 and c_2 which depend on the values of $\nu_{\max}$ and α, but independent of T and M, such that with probability at least $1 - e^{-c_2 MT}$, $0 \leq X_{t,m} \leq C_2$ for at least αMT of the indices.*

We define the set $\mathcal{A}_T = \{t \in (0, 1, .., T-1) |\ A_m^{*\top} X_t \leq \rho A_{\max} C_2 \ \forall\ m\}$, which is the time indices where all the inner products of the rows of A^* and the observations in the vector X_t are less than or equal to the constant $\rho A_{\max} C_2$, which can be controlled by setting C_2 and α appropriately, thus ensuring that $|\mathcal{A}_T| \geq \xi T$ for $\xi \in (0,1)$. For observations at times in $\mathcal{A}_T$ we have the additional following Lemmas.

Lemma 2. *Let the sequence $X_{a_1}, X_{a_2}, ..., X_{a_N}$ where $\xi T \leq N = |\mathcal{A}_T| \leq T$, be randomly drawn according to the Poisson autoregressive model. Define the matrix $\Gamma_t \in \mathbb{R}^{M \times M} = \mathbb{E}[X_{a_t} X_{a_t}^\top | X_{a_{t-1}}]$. Then the smallest eigenvalue of Γ_t is lower bounded by $\omega \triangleq \exp(\nu_{\min} - \rho A_{\max} \max(e^{\nu_{\max}}, C_2))$*

Lemma 2 states that the smallest eigenvalue of Γ_t is bounded away from zero. This is analogous to the restricted eigenvalue condition [29] used in Gaussian noise settings. Typically, in proving sample complexity bounds the restricted eigenvalue condition is assumed to hold for a given (potentially random) design or sensing matrix that is *independent of the data and hence the unknown variable A^*, but dependence introduced in the autoregressive setting makes this a more difficult condition to satisfy.* In much of the literature, the columns of sensing matrices are assumed independent, so we would only need to bound $E[X_t X_t^\top]$. In previous work on correlated Gaussian design matrices, there may be correlation and hence dependence among columns, but they are still signal-independent. In contrast, we bound the conditional expectation $E[X_t X_t^\top | X_{t-1}]$, which is signal dependent. We now move to the sketch proof of the main theorem.

Proof. We start the proof by making an important observation about the estimator defined in Equation 2: this loss function can be completely decoupled by a sum of functions on rows. Therefore the RMLE is the matrix of the row-wise RMLEs. Therefore, we can use a standard method in empirical risk minimization and use the definition of the minimizer of the regularized likelihood for each row:

$$\frac{1}{T}\sum_{t=0}^{T-1} e^{\nu_m} \exp(-\widehat{A}_m^\top X_t) + \widehat{A}_m^\top X_t X_{t+1,m} + \lambda ||\widehat{A}_m||_{1,1}$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1} e^{\nu_m} \exp(-A_m^{*\top} X_t) + A_m^{*\top} X_t X_{t+1,m} + \lambda \|A_m^*\|_{1,1}.$$

By rearranging terms and using the strong convexity property of the Bregman divergence induced by the exponential function, we get the following important property:

$$\|\Delta_{m,\mathcal{S}^C}\|_{1,1} \leq 3\|\Delta_{m,\mathcal{S}}\|_{1,1}$$

where $\Delta_m = \widehat{A}_m - A_m^*$. Thus Δ must lie in the cone

$$\mathcal{B}_S := \{\Delta \in [-A_{\max}, A_{\max}]^{M \times M} |\ \|\Delta_{m,\mathcal{S}^c}\|_{1,1} \leq 3\|\Delta_{m,\mathcal{S}}\|_{1,1}\ \forall m \in [1,2,...,M]\},$$

and we restrict ourselves to studying properties of matrices in that set. From the definition of the RMLE, we can also show that

$$\|\Delta\|_T^2 \quad \leq \quad \frac{3}{C_3}\lambda\sqrt{s}\|\Delta\|_F =: \delta_T \|\Delta\|_F, \tag{3}$$

where $\|\Delta\|_T^2 = \frac{1}{T}\sum_{t \in \mathcal{A}_T} \|\Delta X_t\|_2^2$ for any $\Delta \in \mathbb{R}^{M \times M}$, and $\delta_T = \frac{3}{C_3}\lambda\sqrt{s}$. In this equation C_3 is a constant which depends on $C_2, A_{\max}, \rho$ and the entries of ν. By focusing on $\mathcal{A}_T$, we ensure the strong convexity of the Bregman divergence induced by the exponential function, which leads to Equation 3. We will now use Equation 3 to show that

$$\max(\|\Delta\|_F^2, \|\Delta\|_T^2) \leq \frac{144}{C_3^2 \omega^2 \xi^2} s\lambda^2 = O(e^\rho s\lambda) \tag{4}$$

with probability at least $1 - 2\exp(-\min(c_3 T/\rho^2 - c^4 s\log(2M), c_2 MT))$, which is the result of the theorem.

To get (4) we need to consider three cases: if $\|\Delta\|_T \geq \|\Delta\|_F$, then $\max(\|\Delta\|_T, \|\Delta\|_F) \leq \delta_T$. On the other hand if $\|\Delta\|_T \leq \|\Delta\|_F$ and $\|\Delta\|_F \leq \delta_T$, then $\max(\|\Delta\|_T, \|\Delta\|_F) \leq \delta_T$.

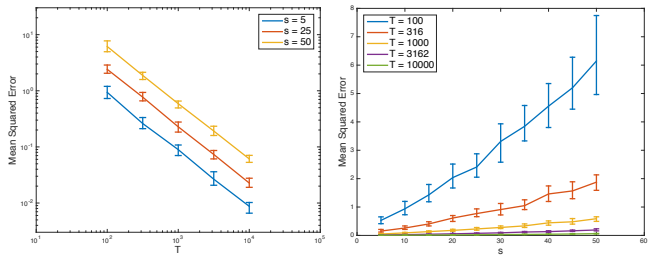
Hence the final case we need to consider is $\|\Delta\|_T \leq \|\Delta\|_F$ and $\|\Delta\|_F \geq \delta_T$. We follow a similar proof technique to that used in [24] for this scenario. Let us define the following set:

$$\mathcal{B}(\delta_T) = \{\Delta \in \mathcal{B}_\mathcal{S} \mid \|\Delta\|_T \leq \|\Delta\|_F, \|\Delta\|_F \geq \delta_T\}.$$

Further, let us define the alternative set:

$$\mathcal{B}'(\delta_T) = \{\Delta \in \mathcal{B}_\mathcal{S} \mid \|\Delta\|_T \leq \|\Delta\|_F, \|\Delta\|_F = \delta_T\}.$$

We wish to show that for $\Delta \in \mathcal{B}(\delta_T)$, we have $\|\Delta\|_T^2 \geq \kappa\|\Delta\|_F^2$ for some $\kappa \in (0,1)$ with high probability, and therefore Equation 3 would imply that $\max(\|\Delta\|_T, \|\Delta\|_F) \leq \delta_T/\kappa$. We claim that it suffices to show that $\|\Delta\|_T^2 \geq \kappa\|\Delta\|_F^2$ is true on $\mathcal{B}'(\delta_T)$ with high probability. In particular, given an arbitrary non-zero $\Delta \in \mathcal{B}(\delta_T)$, consider the re-scaled matrix $\tilde{\Delta} = \frac{\delta_T}{\|\Delta\|_F}\Delta$. Since $\Delta \in \mathcal{B}(\delta_T)$, we have $\tilde{\Delta} \in \mathcal{B}(\delta_T)$ and $\|\tilde{\Delta}\|_F = \delta_T$ by construction. Therefore, if $\|\tilde{\Delta}\|_T^2 \geq \kappa\|\tilde{\Delta}\|_F^2$ is true, then $\|\Delta\|_T^2 \geq \kappa\|\Delta\|_F^2$ is also true. Alternatively if we define the random variable $\mathcal{Z}_T(\mathcal{B}') = \sup_{\Delta \in \mathcal{B}'(\delta_T)}\{\delta_T^2 - \|\Delta\|_T^2\}$, then it suffices to show that $\mathcal{Z}_T(\mathcal{B}') \leq (1-\kappa)\delta_T^2$.



(a) MSE vs. T (b) MSE vs s

Fig. 1: Plot (a) shows the MSE behavior over a widely varying range of T values, from 100 to 10000 behaving as $1/T$. Plot(b) shows the MSE behavior over a range of s values to show that the MSE is linear is s. In all plots the median value of 100 trials is shown, with error bars denoting the middle 50 percentile.

In order to bound this quantity, we take a covering approach by defining the proper covering set of $\mathcal{B}'$ in $\|.\|_T$ norm as $\mathcal{N} := N_{\mathbf{Pr}}(\gamma; \mathcal{B}'; \|.\|_T)$, so that for all $\Delta \in \mathcal{B}'$, there is a Δ^k in $\mathcal{N}$ such that $\|\Delta^k - \Delta\|_T \leq \gamma\delta_T$. Using a simple covering argument and the Cauchy-Schwarz inequality we show

$$\mathcal{Z}_T(\mathcal{B}) \leq \max_{1\leq k\leq \mathcal{N}} \{\delta_T^2 - \|\Delta^k\|_T^2\} + 2\gamma\delta_T^2.$$

For any Δ^k in our covering set, a martingale concentration inequality can be used to show that:

$$\mathbb{P}\Big(\delta_T^2 - \|\Delta^k\|_T^2 > \big(1 - \frac{\xi}{2}\omega\big)\delta_T^2\Big) \leq \exp\left(-\frac{c_3 T}{\rho^2}\right).$$

This line assumes $|\mathcal{A}_T| \geq \xi T$ which holds with probability at least $1 - 2e^{-c_2 MT}$. Now we can use a union bound,

$$\begin{aligned}&\mathbb{P}\left(\max_{k=1,2,\ldots,\mathcal{N}} \delta_T^2 - \|\Delta^k\|_T^2 > (1 - \frac{\xi}{2}\omega)\delta_T^2\right)\\ &\leq \exp\left(\log N_{\mathbf{Pr}}\left(\gamma\delta_T, \mathcal{B}', \|.\|_T\right) - c_3 T/\rho^2\right).\end{aligned}$$

What remains is to bound $\log N_{\mathbf{Pr}}(\gamma\delta_T, \mathcal{B}', \|.\|_T)$. Since the proper covering entropy $\log N_{\mathbf{Pr}}(\gamma\delta_T, \mathcal{B}', \|.\|_T)$ is upper bounded by the standard covering entropy $\log N(\frac{\gamma\delta_T}{2}, \mathcal{B}', \|.\|_T)$, it suffices to upper bound this quantity. This can be accomplished by using a technique called the Sudakov Minoration giving the result

$$\log N(\gamma\delta_T/2, \mathcal{B}', \|.\|_T) \leq \frac{C_4 s}{\gamma^2}\log(2M)$$

and therefore by setting $\gamma = \frac{\xi\omega}{8}$ gives us the desired bound on $\mathcal{B}'(\delta_T)$

$$\mathbb{P}\left(\mathcal{Z}_T(\mathcal{B}) > \left(1 - \frac{\xi}{4}\omega\right)\delta_T^2\right) \leq \exp\left(c_4 s\log(2M) - \frac{c_3 T}{\rho^2}\right).$$

Overall this tells us that on the set $\mathcal{B}'(\delta_T)$ we have that $\|\Delta\|_T^2 \geq \frac{\xi\omega}{4}\|\Delta\|_F^2$ with high probability. □

4. EXPERIMENTAL RESULTS

In this section we validate our theoretical results with experimental results performed on synthetically generated data. We do this by generating many trials of synthetic data with known generating parameters and then compare the estimated values. For all trials the constant offset vector ν was set identically at 0, and the 20×20 matrices A were set such that s randomly assigned values in the range $[0, 1]$. Data were then generated according the process described in Equation 1. An initial burn in period of 100,000 time steps were performed in order to assure that the process had time to sufficiently mix. Then the next T data points were taken and used to perform the estimation. The values of s and T were then varied over a wide range of values. For each (s, T) pair 100 trials were performed and the regularized maximum likelihood estimate $\hat{A}$ was calculated with $\lambda = 0.1/\sqrt{T}$ and the MSE was recorded. The MSE curves are shown in Figure 1. We show a pair of plots which compare the MSE versus increasing behavior of T and s. What is plotted in the figure is the median of 100 trials for each (s, T) pair, with error bars denoting the middle 50 percentile. These plots show that setting λ proportional to $T^{-1/2}$ gives us the desired T^{-1} error decay rate. Additionally, we also see that the error increases linearly in the sparsity level s, as predicted by the theory.

5. CONCLUSIONS

The log-linear Poisson autoregressive process has been used successfully in many settings to learn network structure. However, this model is often used without rigorous non-asymptotic guarantees of accuracy. In this paper we have shown important properties of the regularized maximum likelihood estimator of the PAR process under a sparsity assumption. Namely, we have proven bounds on the MSE of the estimator as a function of sparsity, maximum degree of a node, ambient dimension and time. In order to prove this risk bound, we have incorporated many recently developed tools of statistical learning, including concentration bounds for dependent random variables. Our results on the MSE show that by incorporating sparsity the amount of data needed is on the order of $s\log(M)$ for constant degree networks, which is a significant gain compared to the M^2 parameters being estimated. This $s\log(M)$ dependence is also consistent with rates derived in Gaussian settings.

6. REFERENCES

[1] Kurt Brännäs and Per Johansson, "Time series count data regression," *Communications in Statistics-Theory and Methods*, vol. 23, no. 10, pp. 2907–2925, 1994.

[2] Scott L Zeger, "A regression model for time series of counts," *Biometrika*, vol. 75, no. 4, pp. 621–629, 1988.

[3] Ludwig Fahrmeir and Gerhard Tutz, *Multivariate statistical modelling based on generalized linear models*, Springer Science & Business Media, 2013.

[4] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[5] M. Hinne, T. Heskes, and M. A. J. van Gerven, "Bayesian inference of whole-brain networks," *arXiv:1202.1696 [q-bio.NC]*, 2012.

[6] M. Ding, CE Schroeder, and X. Wen, "Analyzing coherent brain networks with Granger causality," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 5916–8.

[7] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, pp. 995–999, 2008.

[8] M. S. Masud and R. Borisyuk, "Statistical technique for analysing functional connectivity of multiple spike trains," *Journal of Neuroscience Methods*, vol. 196, no. 1, pp. 201–219, 2011.

[9] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim, "Poisson autoregression," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1430–1439, 2009.

[10] Fukang Zhu and Dehui Wang, "Estimation and testing for a poisson autoregressive model," *Metrika*, vol. 73, no. 2, pp. 211–230, 2011.

[11] Konstantinos Fokianos and Dag Tjøstheim, "Log-linear poisson autoregression," *Journal of Multivariate Analysis*, vol. 102, no. 3, pp. 563–578, 2011.

[12] Andréas Heinen, "Modelling time series count data: an autoregressive conditional poisson model," *Available at SSRN 1117187*, 2003.

[13] Fukang Zhu, "Modeling overdispersed or underdispersed count data with generalized poisson integer-valued garch models," *Journal of Mathematical Analysis and Applications*, vol. 389, no. 1, pp. 58–71, 2012.

[14] A. G. Hawkes, "Point spectra of some self-exciting and mutually-exciting point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 83–90, 1971.

[15] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes, Vol. I: Probability and its Applications*, Springer-Verlag, New York, second edition, 2003.

[16] M. Raginsky, R. Willett, Z. Harmany, and R. Marcia, "Compressed sensing performance bounds under Poisson noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 3990–4002, 2010, arXiv:0910.5146.

[17] M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank, "Performance bounds for expander-based compressed sensing in Poisson noise," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, 2011, arXiv:1007.2377.

[18] X. Jiang, R. Willett, and G. Raskutti, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Transactions on Information Theory*, vol. 61, pp. 4458–4474, 2015.

[19] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[20] S. van de Geer, "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, vol. 36, pp. 614–636, 2008.

[21] L. Meier, S. van de Geer, and P. Buhlmann, "High-dimensional additive modeling," *Annals of Statistics*, vol. 37, pp. 3779–3821, 2009.

[22] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2010.

[23] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Transactions on Information Theory*, vol. 57, pp. 6976–6994, 2011.

[24] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," *Journal of Machine Learning Research*, vol. 13, pp. 398–427, 2012.

[25] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Annals of Statistics*, vol. 43, no. 4, pp. 1535–1567, 2015.

[26] X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. Willett, "A data-dependent weighted lasso under poisson noise," *arXiv preprint arXiv:1509.08892*, 2015.

[27] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard, "Lasso and probabilistic inequalities for multivariate point processes," *Bernoulli*, vol. 21, no. 1, pp. 83–143, 02 2015.

[28] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[29] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.