Estimating Network Structure from Incomplete Event Data

Benjamin Mark
University of Wisconsin-Madison

Garvesh Raskutti
University of Wisconsin-Madison

Rebecca Willett
University of Chicago

Abstract

Multivariate Bernoulli autoregressive (BAR) processes model time series of events in which the likelihood of current events is determined by the times and locations of past events. These processes can be used to model nonlinear dynamical systems corresponding to criminal activity, responses of patients to different medical treatment plans, opinion dynamics across social networks, epidemic spread, and more. Past work examines this problem under the assumption that the event data is complete, but in many cases only a fraction of events are observed. Incomplete observations pose a significant challenge in this setting because the unobserved events still govern the underlying dynamical system. In this work, we develop a novel approach to estimating the parameters of a BAR process in the presence of unobserved events via an unbiased estimator of the complete data log-likelihood function. We propose a computationally efficient estimation algorithm which approximates this estimator via Taylor series truncation and establish theoretical results for both the statistical error and optimization error of our algorithm. We further justify our approach by testing our method on both simulated data and a real data set consisting of crimes recorded by the city of Chicago.

1 Introduction

Discrete event data arises in a variety of forms including crimes, health events, neural firings, and social media posts. Frequently each event can be associated with a node in a network, and practitioners aim to learn the relationships between the nodes in the network from the event data. For example, one might observe a sequence of crimes associated with different gangs and seek to learn which crimes are most likely to spark retaliatory violence from rival gangs.

Such problems have attracted widespread interest in recent decades and a variety of point process models have been proposed to model such data. A central assumption of many of these works is that *all* the events are observed. However, in many cases we may observe only a subset of the events at random. For example, while point process models have been widely used to model crime incidence (Mohler (2014); Mohler et al. (2016); Linderman and Adams (2014)), frequently one only has access to *reported* crime data. For many crimes the true number of incidents can be substantially higher. The gap between the reported and true crime rates is referred to as "the dark figure of crime" by researchers in Sociology and Criminology who have studied this issue extensively (Biderman (1991); Langton et al. (2012)). Unobserved events also pose a challenge in inference from Electronic Health Record (EHR) data which can be incomplete for a number of different reasons (Wells et al. (2013); Weiskopf and Weng (2013)).

The unobserved events still play a role in the dynamical system governing the time series, making network estimation with incomplete data particularly challenging. In this paper, we contribute to the growing literature on modeling in the presence of unobserved events by proposing a novel method for network estimation when we only observe a subset of the true events.

1.1 Problem Formulation

Many point process models of time series of discrete events are temporally discretized either because event times were discretized during the data collection process or for computational reasons. In such contexts, the temporal discretization is typically such that either one or zero events are observed in each discrete time block for each network node. With this in mind, we model the true but unobserved observations $X_1, \dots X_T$ using a Bernoulli autoregressive process:

$$\begin{aligned} Y_t &= \nu + A^* X_{t-1} \\ X_t &\sim \text{Bernoulli}\left(\frac{1}{1+\exp(-Y_t)}\right). \end{aligned} \tag{1.1}$$

Here $X_t \in \{0,1\}^M$ is a vector indicating whether events

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

occurred in each of the M nodes during time period t. The vector $\nu \in \mathbb{R}^M$ is a constant bias term, and the matrix $A^* \in \mathbb{R}^{M \times M}$ is the weighted adjacency matrix associated with the network we wish to estimate. We assume that each row a of A^* lies in the ℓ_1 ball of radius r, which we denote $\mathbb{B}_1(r)$. We generally consider a case where a is sparse and the magnitude of all its entries are bounded, so that r is a universal constant which is independent of M.

We observe $Z_1, \dots Z_T$, a corrupted version of (1.1) where only a fraction $p \in (0, 1]$ of events are observed as follows:

$$\begin{aligned} W_{t,i} &\overset{iid}{\sim} \text{Bernoulli}(p) \\ Z_t &= W_t \odot X_t. \end{aligned} \tag{1.2}$$

Here $\odot$ denotes the Hadamard product and $W_t \in \{0, 1\}^M$ is a vector where each entry is independently drawn to be one with probability p and zero with probability $1 - p$.

Our analysis of (1.1) and (1.2) can be naturally extended to several more complex variants. Instead of assuming each $X_{t,i}$ is observed with probability p, we can assume events from each node i are observed with a unique probability p_i. We consider only a first order AR(1) process but our framework can be extended to incorporate more sophisticated types of memory as in Mark et al. (2018). We omit discussion of these extensions in the interest of clarity.

2 Related Work

There is an extensive literature on the analysis of complete discrete event data using various point process models. One framework, the Hawkes process (Hawkes (1971)) is a popular continuous time approach which has been applied in a number of different contexts (e.g., Zhou et al. (2013); Yang et al. (2017); Chen et al. (2017); Xu et al. (2018)). In addition, other works have used a discrete time framework to model time series event data (e.g., Linderman et al. (2016); Fletcher and Rangan (2014); Bao et al. (2017)).

Corrupted or missing data in high-dimensional data sets is a problem which appears in a number of different domains and has attracted widespread interest over the past few decades (see Graham (2009) for an application-focused overview). Our focus is on a particular type of corrupted data: partially observed sequences of discrete events. In recent years researchers have started to focus on this problem (Xu et al. (2017); Shelton et al. (2018); Le (2018)). The prior works of which we are aware use a Hawkes process framework and assume knowledge of the time periods when the data is corrupted. In the context of (1.2) this amounts to knowledge of $W_1, \dots, W_T$. Our method can operate in a setting where the researcher cannot differentiate between periods when no event actually occurred, and when events potentially occurred but were not recorded. Moreover, because we use a discrete-time framework, we are able to derive sample complexity bounds for the estimation procedure proposed in Section 3. Our theoretical results complement the empirical nature of much of the past work in this area.

This paper is also related to a variety of works on regularized estimation in high-dimensional statistics, including Bickel et al. (2009); Raskutti et al. (2010); Basu and Michailidis (2015) and Jalali and Willett (2018). Many of these works have derived sample complexity guarantees using linear models, and some of these results have been extended to autoregressive generalized linear models (Negahban et al. (2010); Hall et al. (2016)). Another line of research (Loh and Wainwright (2012); Agarwal et al. (2012); Loh and Wainwright (2015); Loh (2017); Negahban et al. (2010)) has formalized a notion of *Restricted Strong Convexity* (RSC) which we leverage in Section 4.2. While many loss functions of interest in high-dimensional statistics are not strongly convex, these works have shown that they frequently satisfy an RSC condition which is sufficient to provide statistical and optimization error guarantees. The main technical challenges in our setting lie in establishing results similar to these RSC conditions.

2.1 Missing Data in a High-Dimensional Linear Model

Loh and Wainwright (2012) straddles the missing data literature and high-dimensional statistics literature. The authors consider a missing data linear model

$$\begin{aligned} Y_i &= X_i^\top \beta^* + \epsilon_i \\ Z_i &= W_i \odot X_i \end{aligned} \tag{2.1}$$

where $W_i \overset{iid}{\sim} \text{Bernoulli}(p)$ and one observes pairs (Y_i, Z_i) and aims to estimate β^*. The authors propose minimizing a loss function $L_{\text{missing},Z,W}$ of the observed data Z which satisfies the property

$$\mathbb{E}[L_{\text{missing},Z,W}(\beta)|X] = L_{\text{Lasso},X}(\beta)$$

for any β. Here

$$L_{\text{Lasso},X} := \frac{1}{2} \sum_{i=1}^{T} (Y_i - X_i^\top \beta)^2 + \lambda \|\beta\|_1$$

denotes the classical Lasso loss function with the unobserved data X and regularization parameter $\lambda > 0$. In other words, the missing data loss function is an unbiased estimator for the full data Lasso loss function we would ideally like to minimize. This idea motivates our construction of a loss function for the observed process (1.2).

Our problem can be viewed as an extension of Loh and Wainwright (2012) to *autoregressive BAR models without knowledge of* W.[1] In particular, we cannot distinguish

[1] Note that Loh and Wainwright (2012) does consider AR processes, but in a different context from our setting. Specifically, we wish to estimate the AR process parameters, where as they consider a special case of (2.1) where $X_{t+1} = AX_t + \epsilon_t$ but where A is known and one aims to estimate β^*.

events that were missed ($X_{t,j} = 1$ and $W_{t,j} = 0$) from correctly observed periods with no events ($X_{t,j} = 0$). Loh and Wainwright (2012) are able prove sample complexity bounds as well as optimization bounds which are consistent with the high-dimensional statistical literature in that they scale with $\|\beta^*\|_0$ rather than the dimension of β^*. We are able to prove analogous bounds for our estimator in Section 4.

2.2 Contributions

This paper makes the following contributions.

- We propose a novel method for network estimation when only a subset of the true events are observed. In contrast to previous work, our methods do not rely on knowledge of when the data is potentially missing. Our procedure uses Taylor series approximations to an unbiased loss function, and we show that these approximations have controlled bias and lead to accurate and efficient estimators.
- We prove bounds on both the statistical error and optimization error of our proposed estimation method. The results hinge on showing that our loss function satisfies a restricted strong convexity (RSC) condition. Past work on linear inverse problems with corrupted designs also establish RSC conditions, but these conditions do not carry over to the autoregressive BAR setting.
- We demonstrate the effectiveness of our methodology on both simulated data and real crime data.

3 Proposed Estimation Procedure

Given the full data $X = [X_1, \ldots, X_T]$, the negative log-likelihood function $L_X(A)$ is decomposable in the M rows of A. In other words, if

$$A = \begin{bmatrix} a_1^\top & a_2^\top & \cdots & a_M^\top \end{bmatrix}^\top$$

then $L_X(A) = \sum_{m=1}^M L_X(a_m)$ where a_m is the mth row of A and $L_X(a_m)$ denote the loss function restricted to a specific row. Throughout the paper we slightly abuse notation and let $L_X(A)$ refer to the entire loss function when A is a matrix, and let $L_X(a_m)$ refer to the loss function for a specific row when a_m is a row vector.

The loss function for the mth row takes the form

$$L_X(a_m) := \frac{1}{T}\sum_{t=1}^{T} f(a_m^\top X_t) - X_{t+1,m}(a_m^\top X_t)$$

where $f(x) = \log(1 + \exp(x))$ is the partition function for the Bernoulli GLM.

We do not have access to X and instead we aim to estimate A using the corrupted data $Z = [Z_1, \ldots, Z_n]$. As discussed in Section 2.1, our strategy will be to construct a loss function of Z which is an unbiased estimator for L_X. In other words, we want to find some function $L_{Z,p}$ such that for any $a_m \in \mathbb{B}_1(r)$,

$$\mathbb{E}[L_{Z,p}(a_m)|X] = L_X(a_m). \tag{3.1}$$

In contrast to the Gaussian case discussed in Section 2.1, the Bernoulli partition function $f(x) = \log(1 + \exp(x))$ is not a polynomial and constructing a function satisfying (3.1) directly is challenging. We adopt a strategy of computing unbiased approximations to truncated Taylor series expansions of L_X and arriving at $L_{Z,p}$ as a limit of such approximations.

To do this, we first rewrite f using its Taylor series expansion around zero

$$f(a_m^\top X_t) = \log(2) + \frac{a_m^\top X_t}{2} + \frac{(a_m^\top X_t)^2}{8} - \frac{(a_m^\top X_t)^4}{192} + o((a_m^\top X_t)^6).$$

The constant factor $\log(2)$ does not effect estimation in any way so we ignore it for the remainder of our discussion in the interest of simplicity. We let $L_X^{(q)}$ denote the degree q Taylor truncation to L_X. The X_t are binary vectors and we assume each row a_m is sparse, so $a_m^\top X_t \leq \|a_m\|_1$ will not be too far from zero. Thus it is reasonable to hope that for small q, $L_X^{(q)}(a_m)$ is a good approximation for $L_X(a_m)$ whenever $a_m \in \mathbb{B}_1(r)$. We bound the approximation error in Lemma B.3 in the supplement.

We now consider the problem of constructing a function $L_{Z,p}^{(q)}$ such that

$$\mathbb{E}[L_{Z,p}^{(q)}(a_m)|X] = L_X^{(q)}(a_m) \text{ for all } a_m \in \mathbb{B}_1(r). \tag{3.2}$$

Once we construct $L_{Z,p}^{(q)}$ we can estimate the mth row of A^* by attempting to solve

$$\hat{a}_m = \operatorname*{arg\,min}_{a \in \mathbb{B}_1(1)} L_{Z,p}^{(q)}(a) + \lambda\|a\|_1.$$

Key question we need to address with this approach include (a) can we (approximately) solve this optimization problem efficiently? (b) will the solution to this optimization problem be robust to initialization? (c) will it be a strong estimate of the ground truth?

3.1 Definition of $L_{Z,p}^{(2)}$

We first derive an unbiased estimator of the degree-two Taylor series expansion $L_X^{(2)}(a_m)$.

$$L_X^{(2)}(a_m) = \frac{1}{T}\sum_{t=1}^{T} \frac{a_m^\top X_t}{2} - X_{t+1,m}(a_m^\top X_t) + \frac{(a_m^\top X_t)^2}{8}.$$

Note that there are straightforward unbiased estimates of the first two terms:

$$\mathbb{E}\left[\frac{1}{p}\frac{a_m^\top Z_t}{2}|X\right] = \frac{a_m^\top X_t}{2}$$
$$\mathbb{E}\left[\frac{1}{p^2}Z_{t+1,m}(a_m^\top Z_t)|X\right] = X_{t+1,m}(a_m^\top X_t). \quad (3.3)$$

For the third term, $\frac{(a_m^\top X_t)^2}{8} = \sum_{i,j} a_{m,i}a_{m,j}X_{t,i}X_{t,j}$, note that

$$\mathbb{E}[Z_{t,i}Z_{t,j}|X] = \begin{cases} p^2 X_{t,i}X_{t,j} & \text{if } i \neq j \\ pX_{t,i}X_{t,j} & \text{if } i = j \end{cases}. \quad (3.4)$$

Thus we must estimate the monomials with repeat terms ($i = j$) differently from the monomials with all distinct terms ($i \neq j$). Using Equations (3.3) and (3.4) we can define $L_{Z,p}^{(2)}$:

$$L_{Z,p}^{(2)}(a_m) := \frac{1}{T}\sum_t \left[\frac{a_m^\top Z_t}{2p} - \frac{Z_{t+1,m}(a_m^\top Z_t)}{p^2} + \sum_{i\neq j}\frac{a_{m,i}a_{m,j}Z_{t,i}Z_{t,j}}{8p^2} + \sum_i \frac{a_{m,i}^2 Z_{t,i}}{8p}\right]. \quad (3.5)$$

3.2 Higher-Order Expansions

The construction of $L_{Z,p}^{(2)}$ in the previous section suggests a general strategy for constructing $L_{Z,p}^{(q)}$ satisfying (3.2). Take any monomial

$$X_{t,m_1} \cdot \ldots \cdot X_{t,m_d}$$

depending on the counts in nodes $m_1, \ldots m_d$ during time period t. Wherever this monomial appears in $L_X^{(q)}(a_m)$, our unbiased loss function will have a term

$$\frac{1}{p^k}Z_{t,m_1} \cdot \ldots \cdot Z_{t,m_d}$$

where k denotes the number of unique terms in the monomial. For example, in Equation (3.3) each degree two monomial was unique so we scaled everything by $\frac{1}{p^2}$. However, in (3.5) some of the degree two monomials had repeated terms and so they were scaled by $\frac{1}{p}$. In order to formalize these ideas and generalize our estimator to $q > 2$, we first need to introduce additional notation.

3.2.1 Notation

Let $\mathcal{U}_d$ denote the set of all monomials of degree d. We represent an element $U \in \mathcal{U}_d$ as a list containing d elements. An element in the list corresponds to the index of a term in the monomial. For an example, the monomial $x_1^2 x_3$ can be represented as the list $(1, 1, 3)$.

For a polynomial function h we let $c_{U,h}$ denote the coefficient of the monomial U in h. Finally we define the order of a list to denote the number of unique elements in the list, so $|(1, 2)| = 2$ whereas $|(1, 1)| = 1$.

Example *Consider the function* $h(x_1, x_2) = x_1^2 + 4x_1x_2$. *We can decompose* h *as*

$$h(x_1, x_2) = \sum_{U\in\mathcal{U}_2} c_{U,h}\prod_{u\in U} x_u$$

where $\mathcal{U}_2 = \{(1, 1), (1, 2), (2, 2)\}$ *with corresponding coefficients* $c_{(1,1),h} = 1$, $c_{(1,2),h} = 4$ *and* $c_{(2,2),h} = 0$.

Using this notation we can write

$$L_X^{(q)}(a_m) = \frac{1}{T}\sum_{t=1}^T \left[\sum_{i=1}^q \sum_{U\in\mathcal{U}_i}\left(c_{U,f}\prod_{u\in U} X_{t,u}\prod_{u\in U} a_{m,u}\right) - X_{t+1,m}(a_m^\top X_t)\right].$$

3.2.2 Definition of $L_{Z,p}^{(q)}$

The degree q likelihood is constructed as follows

$$L_{Z,p}^{(q)}(a_m) := \frac{1}{T}\sum_{t=1}^T \left[\sum_{i=1}^q \sum_{U\in\mathcal{U}_i}\left(\frac{c_{U,f}}{p^{|U|}}\prod_{u\in U} Z_{t,u}\prod_{u\in U} a_{m,u}\right) - \frac{Z_{t+1,m}(a_m^\top Z_t)}{p^2}\right].$$

Recall that $|U|$ denotes the number of unique terms in the monomial U. In other words, we adjust $L_X^{(q)}$ by scaling each monomial according the the number of unique terms rather than the number of overall terms. This definition clearly satisfies (3.2). We show in the supplement that if $r = 1$ and $p > \frac{1}{\pi}$ then $\lim_{q\to\infty} L_{Z,p}^{(q)}(a_m)$ converges uniformly on $\mathbb{B}_1(r)$ to a function we denote as $L_{Z,p}(a_m)$. Extending this loss function on individual rows to an entire matrix, we can define $L_{Z,p}(A) = \sum_{i=1}^M L_{Z,p}(a_m)$. An additional technical discussion in the supplement guarantees that $L_{Z,p}$ actually satisfies the desired property in Equation (3.1).

3.3 Proposed Optimization

In practice we can only compute $L_{Z,p}^{(q)}$ for finite q. To estimate A^* we consider the following constrained optimization:

$$\widehat{A} \in \underset{A\in\mathbb{B}_{1,\infty}(r)}{\arg\min}\ L_{Z,\hat{p}}^{(q)}(A) + \lambda\|A\|_1 \quad (3.6)$$

where $\hat{p}$ is an estimate of the missingness parameter p and

$$\mathbb{B}_{1,\infty}(r) = \{A \in \mathbb{R}^{M\times M} : ||a_m||_1 \leq 1 \text{ for all } m\}.$$

In general, $L_{Z,p}^{(q)}$ is not a convex function. However, we show in Section 4.2 that under certain assumptions all stationary points of (3.6) must lie near A^*. Thus we can approximately solve (3.6) via a simple projected gradient descent algorithm.

In order to apply our algorithm it is necessary to have an estimate $\hat{p}$ of the frequency of missed data. In many cases one may have prior knowledge available. For example, social scientists have attempted to quantify the frequency of unreported crimes (Langton et al. (2012); Palermo et al. (2014)). Moreover, a simulation study in Section 5 suggests our strategy is robust to misspecification of p.

4 Learning Rates

In this section we address both the statistical and optimization aspects of our proposed estimation procedure. Throughout the section we assume $p > \frac{1}{\pi}$ and $A^* \in \mathbb{B}_{1,\infty}(1)$. All results in the section apply for the loss functions $L_{Z,p}^{(q)}$ for $q \in \mathbb{N} \cup \{\infty\}$. In the $q = \infty$ case we recover the idealized loss function $L_{Z,p}$.

We use $a \lesssim b$ to mean $a \leq Cb$ and $a \asymp b$ to mean $a = Cb$ where C is a universal constant. Define $s := \|A^*\|_0$ and $\rho := \max_m \|a_m^*\|_0$.

4.1 Statistical Error

We first address the statistical error of our estimator. The following theorem controls the statistical error of our proposed estimator.

Theorem 4.1 (Accuracy of $L_{Z,p}^{(q)}$)**.** *Suppose*

$$\widehat{A} \in \underset{A \in \mathbb{B}_{1,\infty}(1)}{\arg\min}\ L_{Z,p}^{(q)}(A) + \lambda \|A\|_1$$

where $\lambda \asymp \frac{\log(MT)}{\sqrt{T}(p\pi - 1)} + \frac{1}{(p\pi)^q}$*. Then*

$$\|\widehat{A} - A^*\|_F^2 \lesssim \frac{s \log^2(MT)}{T(\pi p - 1)^2} + \frac{s}{(p\pi)^{2q}}$$

for $T \gtrsim \rho^2 \log(MT)$ *with probability at least* $1 - \frac{1}{T}$.

When $q = \infty$ the rate $\|\widehat{A} - A^*\|_F^2 = O\left(\frac{slog^2(MT)}{T}\right)$ matches the minimax optimal rate for sparse high-dimensional linear regression up to log factors (Raskutti et al. (2011)). The two terms in the upper bound of Theorem 4.2 have a natural interpretation. The first represents the error for the idealized estimator $L_{Z,p}$, while the second represents the error due to the Taylor series truncation. Our error scales as $(\pi p - 1)^{-2}$ which is reasonable in the context of our algorithm because $L_{Z,p}(A) := \lim_{q \to \infty} L_{Z,p}^{(q)}(A)$ is only well-defined when $p > \frac{1}{\pi}$ (see Remark 2 in the supplement). An interesting open question which arises from Theorem 4.1 is whether the process described in (1.1) and (1.2) is unidentifiable for $p \leq \frac{1}{\pi}$ or something specific to our methodology fails for p below this threshold.

The proof of Theorem 4.1 uses ideas from the analysis of high-dimensional GLMs (Hall et al. (2016); Mark et al. (2018)) as well as ideas from the analysis of missing data in the linear model (Loh and Wainwright, 2012) and Gaussian linear autoregressive processes (Jalali and Willett, 2018). The key technical challenge in the proof lies in controlling the gradient of the error term $R^{(q)}(A) := L_X(A) - L_{Z,p}^{(q)}(A)$. This is done in Lemmas B.2-B.6 in the supplement.

4.2 Optimization Error

We next focus on the optimization aspects of Equation (3.6). Our loss function $L_{Z,p}^{(q)}$ is non-convex, so at first glance it may appear to be a poor proxy loss function to optimize. However, a body of research (see Agarwal et al. (2012); Loh and Wainwright (2015, 2012); Loh (2017)) has studied loss functions satisfying a properties known as restricted strong convexity (RSC) and restricted smoothness (RSM). These works have shown that under certain conditions, the optimization of non-convex loss functions may be tractable. The formal definitions of the RSC and RSM conditions we use are as follows.

Definition 1 (Restricted Strong Convexity)**.** Let $T_L(v, w)$ denote the first order Taylor approximation to a loss function L centered at w. A loss function L satisfies the RSC condition with parameters α, τ if

$$T_L(v, w) \geq \frac{\alpha}{2}\|v - w\|_2^2 - \tau \|v - w\|_1^2$$

for all $v, w \in \mathbb{B}_1(1)$.

Definition 2 (Restricted Smoothness)**.** A loss function L satisfies the RSM condition with parameters α, τ if

$$T_L(v, w) \leq \frac{\alpha}{2}\|v - w\|_2^2 + \tau \|v - w\|_1^2$$

for all $v, w \in \mathbb{B}_1(1)$.

We are able to show these conditions are satisfied for $\alpha \asymp 1$ and $\tau \asymp \sqrt{\frac{\log(MT)}{T}} + (p\pi)^{-q}$. This in turn gives the following result. As in Theorem 4.1 we assume $p > \frac{1}{\pi}$.

Theorem 4.2. *Suppose* $A^* \in \mathbb{B}_{1,\infty}(1)$ *and* $\|a_m^*\|_0 > 0$ *for at least* $\frac{M}{C}$ *rows of* A^* *where* C *is a universal constant. Let* $\tilde{A} \in \mathbb{B}_{1,\infty}(1)$ *be any stationary point of* $L_{Z,p}^{(q)}(A) + \lambda\|A\|_1$ *where* $\lambda \asymp \frac{\log(MT)}{\sqrt{T}(p\pi - 1)} + \frac{1}{(p\pi)^q}$*. Then*

$$\|\tilde{A} - A^*\|_F^2 \lesssim \frac{s}{p\pi - 1}\sqrt{\frac{\log(MT)}{T}} + \frac{s}{(p\pi)^q}$$

with probability at least $1 - \frac{\log(T)}{T^2}$ *for* $T \gtrsim \rho^2 \log(MT)$ *and* $q \gtrsim \frac{\log(\rho)}{\log(\pi p)}$.

As in Theorem 4.1 the first term in our bound can be interpreted as the error for the idealized estimator $L_{Z,p}$ while the second term can be thought of as the error due to the Taylor

series truncation. The assumption that $\|a_m^*\|_0 > 0$ for at least $\frac{M}{C}$ rows of A^* says that at least a constant fraction of nodes are influenced by other nodes in the network. This assumption allows us to state Theorem 4.2 in terms of s - the support of A^*. In extreme cases where almost all nodes in the network fire independently of the other nodes it is possible for the optimization error to have a slower scaling than s.

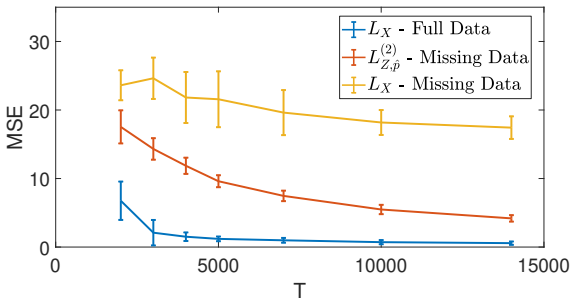
The RSC and RSM conditions are closely related to ideas used in our statistical error bounds in Theorem 4.1. Lemma D.2 shows that the conditions are satisfied for τ on the order of $\frac{1}{\sqrt{T}}$ which leads to an overall optimization error bound of the same order. This is a slower convergence rate than in the linear case; whether stronger rates can be obtained in the autoregressive GLM setting is an open question.

In order to prove Theorem 4.2 we first establish that the RSC/RSM conditions hold for reasonable constants in Lemma D.2. This proofs relies on the technical machinery built up in Lemmas B.2-B.6. We then combine our RSC/RSM results with Theorem 2 in Agarwal et al. (2012) to conclude that all stationary points of $L_{Z,p}^{(q)}(A) + \lambda\|A\|_1$ lie in a small neighborhood of A^* with high probability.

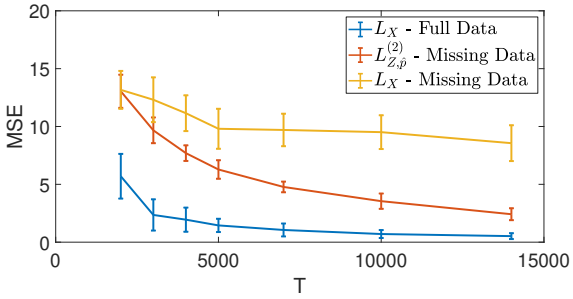
5 Simulations

In this section we evaluate the proposed method on synthetic data. We generate $A^* \in \mathbb{R}^{50\times 50}$ with $s = 50$ nonzero entries with locations chosen at random. Each nonzero entry is chosen uniformly in $[-1, 1]$ and $\nu = 0$. We then generate a "true" data set X and an "observed" data set Z according (1.1) and (1.2) with $\lambda = \frac{.75}{\sqrt{T}}$. We perform projected gradient descent with a random initialization and show a median of 50 trials.

Figure 1 shows mean squared error (MSE) vs T for $p = .6$ (top) and $p = .75$ (bottom). Our method is shown in red. It uses the loss function $L_{Z,p}^{(2)}$ on the partially observed data Z. Our method is compared to the loss function L_X using both the full data X (i.e., an oracle estimator with access to the missing data) and the partially observed data Z (i.e. a naive estimator that ignores the possibility of missing data). As expected, with access to the complete data one can get a more accurate estimate of A^* than either method using the partially observed data. However, our method outperforms the full data likelihood when given the partially observed data. In particular, note that the accuracy for the full data likelihood stalls after some time due to the inherent bias in using the corrupted data on the true data likelihood. In contrast our unbiased method continues to converge to the solution, as suggested by the results in Section 4. Finally, observe that for large T there is little variation between trials when using L_Z^2 even though each trial was initialized randomly. This agrees Theorem 4.2 which states that all stationary points of $L_{Z,p}^{(q)}$ lie near one another.



(a) $p = \hat{p} = .6$



(b) $p = \hat{p} = .75$

Figure 1: MSE vs T for $p = .6$ (top) and $p = .75$ (bottom). The blue line uses regularized MLE on full data – i.e. data unavailable in our setup – and represents a kind of oracle estimator. The red line uses incomplete data with $L_{Z,p}^{(2)}$ (our proposed method). The yellow lines corresponds to minimizing the full data likelihood over the incomplete data – that is, this estimator naively ignores the issue of missing data. Median of 50 trials is shown and error bars denote sample standard deviations.

In practical applications one may have strong reason to believe some events are unobserved, but pinning down a precise estimate of the missingness parameter p might be unrealistic. Therefore it is important to see how our algorithm performs as a function of the misspecification $\hat{p} - p$. We examine this in Figure 2. We generate data as in the previous section but with $p = .7$. We then apply our algorithm with the loss function $L_{Z,\hat{p}}^{(2)}$ and varying values of $\hat{p}$.

Figure 2 shows that our method is highly robust to small misspecification of the missingness parameter p. Interestingly, underestimating p by more than 10% leads to poor results but our method is still robust to overestimation of over 10%. This suggests there is value in applying our techniques with a conservative estimate of the amount of missed data, even when one has only a rough estimate of the frequency of missed events.

As a final experiment we measure how MSE varies as a function of the Taylor series truncation level q. Calculating $L_{Z,p}^{(4)}$ takes a significant amount of time for high-dimensional problems, so we randomly generate $A^* \in \mathbb{R}^{20\times 20}$ with $s = 20$ nonzero entries compared to 50 in the previous simulations and run 30 trials. We set $p = \hat{p} = .7$.

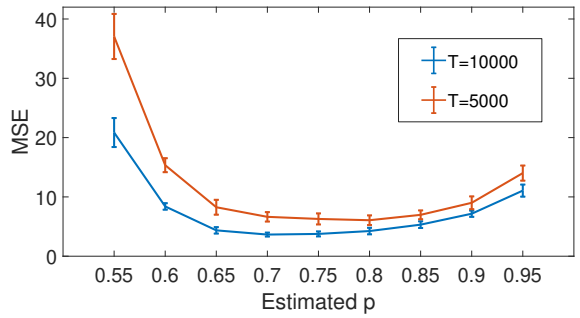


Figure 2: Robustness to missestimation of p using a true value of $p = .7$. Median of 50 trials is shown and error bars denote sample standard deviations.

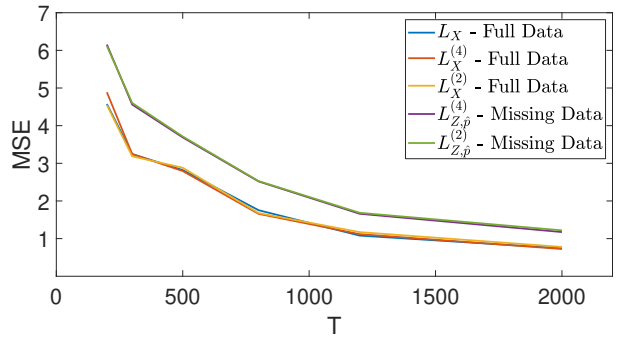


Figure 3: MSE vs T using different loss functions in (3.6). L_X, L_X^4 and L_X^2 use the full data X while $L_{Z,p}^{(4)}$ and $L_{Z,p}^{(2)}$ use the missing data Z. Plots suggest that Taylor series truncations produce nearly identical results to complete loss functions.

In Figure 3 we show MSE as a function of T for the full data loss function at three different truncation levels: L_X^2, L_X^4 and L_X. Recall that L_X and $L_{Z,p}$ has no odd degree terms other than 1, so $L_X^{(3)} = L_X^{(2)}$ and $L_{Z,p}^{(3)} = L_{Z,p}^{(2)}$. We see that the second and fourth degree truncations perform essentially the same as the full data likelihood. We also plot MSE as a function of T for the truncated missing data loss functions $L_{Z,p}^{(2)}$ and $L_{Z,p}^{(4)}$. As expected, using the full data gives stronger results than the partially observed data. We again see that the second and fourth order truncations perform nearly the same. The sample standard deviations are also similar - e.g. when $T = 2000$ the standard deviations using L_X, L_X^2 and L_X^4 are .184, .178 and .186 respectively while the standard deviations using L_Z^2 and L_Z^4 are .322 and .311. The experiments in Figure 3 involving $L_{Z,p}^{(4)}$ take approximately 16 times longer than those involving $L_{Z,p}^{(2)}$. The similarity between the second and fourth order truncation levels suggests that choosing one of these truncation levels will give us a strong approximation to $L_{Z,p}$. Since $L_{Z,p}^{(4)}$ takes significantly longer to compute, we use the second order approximation in the first two experiments. In general we expect the cost of computing $L_{Z,p}^{(q)}$ to scale exponentially in q, so computing $L_{Z,p}^{(q)}$ for large q will not be tractable.

6 Chicago Crime Data

This section studies a data set consisting of crimes committed in Chicago between January 2001 and July 2018 (City of Chicago (2018)). Point process models have been applied to this data set in the past (Linderman and Adams (2014)). In a missing data setting, in order to validate our model it is important to have a "ground truth" data set. For this reason we limit our study to homicides within the data set. For other crimes recorded data is known to be incomplete (Langton et al. (2012); Biderman (1991)), but we assume that nearly every murder is observed. This allows us to create an incomplete data set by removing murders randomly while still maintaining a ground truth data set for validation. The goal of this section is to illustrate how one might apply our method to extract more signal from real data than would be possible using a naive method. However, we want to emphasize that crime data may be corrupted in ways beyond those that are considered in this paper. Using statistical models to guide policy without a more systematic attempt to control for various biases in crime data can lead to a number of pitfalls (see Hao (2019) for further discussion).

The city is divided into 77 community areas and the community area where each murder occurred is recorded. The majority of these community areas experience few murders so we focus on the nine areas with the most murders since 2001. These areas form two clusters: one on the west side of the city and another on the south side. We discretize the events using one week bins, so $X_{t,i} = 1$ if a murder occurred in community area i during week t and $X_{t,i} = 0$ otherwise. This gives a data matrix $X \in \{0,1\}^{9\times 918}$ which we divide into a train set $X_{\text{train}} \in \{0,1\}^{9\times 600}$ containing the first 600 weeks in the period, and a test set $X_{\text{test}} \in \{0,1\}^{9\times 318}$ containing the final 318 weeks. We then create an incomplete data set $Z_{\text{train}} = W \odot X_{\text{train}}$ where $W \in \{0,1\}^{9\times 318}$ contains independent realizations of a Bernoulli random variable with mean $p = .75$.

We learn parameters $\nu_X \in \mathbb{R}^9$ and $\hat{A}_X \in \mathbb{R}^{9\times 9}$ using the training set X_{train} and the full data likelihood L_X. We also learn parameters $\nu_{Z,\hat{p}} \in \mathbb{R}^9$ and $\hat{A}_{Z,\hat{p}} \in \mathbb{R}^{9\times 9}$ using the incomplete train data Z_{train} and the missing data likelihood $L_{Z,\hat{p}}^2$ for various values of $\hat{p}$.

We compare the log-likelihood of these parameters on the test set X_{test}. The results are shown in Figure 4. The missing data estimates perform nearly as well as the full data estimate when $\hat{p}$ is close to the true value of .75. Note that $L_{Z,1}^2 = L_X^2$ closely approximates the full data likelihood L_X and the hold out likelihood is substantially worse for $L_{Z,1}^2$ compared to $L_{Z,\hat{p}}^2$ for $\hat{p}$ close to .75. In other words, ignoring the missing data entirely gives a weaker estimate than applying the techniques this paper introduces, even when $\hat{p}$ is not a perfect estimate of p. Finally, observe that $L_{Z,\hat{p}}$ is more robust to misspecification when $\hat{p} > p$ compared to when $\hat{p} < p$. This is a trend which also appears in Figure 2

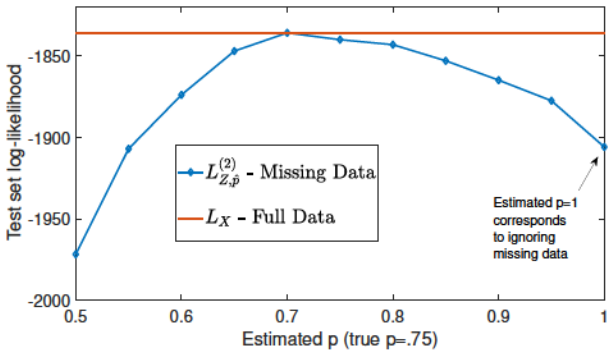


Figure 4: Test performance on Chicago crime data. Log-likelihood of events on hold out set using full data with L_X (yellow) and partial data with $L^2_{Z,\hat{p}}$ for various values of $\hat{p}$, where $p = 0.75$ (blue). For $\hat{p}$ near p, the proposed estimator performs nearly as well as an oracle estimator with full access to the missing data, and significantly better than a naive method that ignores the possibility of missing data.

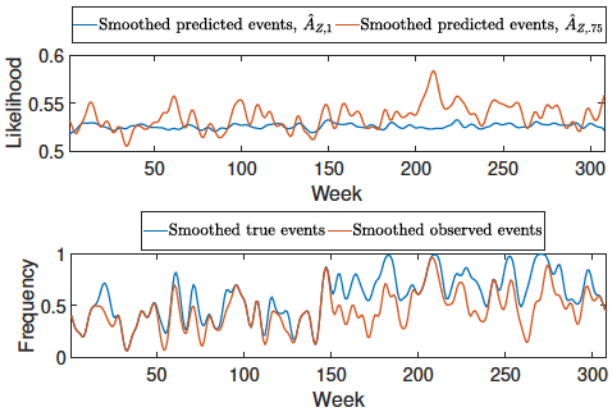


Figure 5: Result of density propagation on Chicago crime data for Community Area 25 (Austin). After a training period used to estimate $\hat{A}_{Z,.75}$ (proposed estimator) and $\hat{A}_{Z,1}$ (naive estimator that doesn't account for missing data), density propagation is run in subsequent test weeks to predict the likelihood of each community area having a homicide at time n based on observations up to time $n-1$. The top panel shows the predicted likelihood of a homicide occurring in the Austin community area of Chicago. The network $\hat{A}_{Z,.75}$ predicts 960 total homicides in the nine community areas during the test period, compared to 748 for $\hat{A}_{Z,1}$. The actual number of homicides was 1035. The bottom panel shows the true events as well as the partially observed events (after Gaussian smoothing used for visualization).

and suggests there is value in using conservative estimates of the amount of missed data in practical applications.

Given estimates of A and p we can use density propagation to predict the likelihood of homicides during week n based on observed homicides up to week $n-1$. We do this for $\hat{A}_{Z,.75}$ learned from the incomplete data Z_{train} with $\hat{p} = .75$ as well as $\hat{A}_{Z,1}$ learned from Z_{train} but with $\hat{p} = 1$, which corresponds to assuming there is no missing data. We use particle filtering to construct estimates

$$p(X_n = 1 | \hat{A}_{Z,.75}, Z_1, \dots Z_{n-1})$$

and

$$p(X_n = 1 | \hat{A}_{Z,1}, Z_1, \dots Z_{n-1}).$$

These probabilities correspond to the likelihood of homicides during the nth week based on the observations over the first $n-1$ weeks. We construct such estimates for each week in the 318 week test set. As expected $\hat{A}_{Z,.75}$ assigns higher likelihoods of homicides, with 960 expected homicides in total compared to 748 for $\hat{A}_{Z,1}$. As a naive method of correcting for missing data, we divide $p(X_n = 1 | \hat{A}_{Z,1}, Z_1, \dots Z_{n-1})$ by a constant scaling factor of 0.75 and report these likelihoods below; by doing this, we ensure that both predictions yield similar *average* numbers of homicides, so differences in performance between the proposed and naive estimator are not due to a simple difference in averages, but rather because the proposed method is capturing the underlying dynamics of the system.

Figure 5 displays these likelihoods for Community Area 25 (Austin) which has the largest number of homicides recorded during the test period. We use Gaussian smoothing to help visualize the trends. The top panel shows the predicted probability of events using $\hat{A}_{Z,.75}$ (in red) and the scaled predicted probability of events using $\hat{A}_{Z,1}$ (in blue). The bottom panel shows the actual and partially observed events during the test period. The true events generally peak at times in which the predicted events for $\hat{A}_{Z,.75}$ peak. For example, both the predicted event and true event charts have peaks around weeks 60 and 210. In contrast, the predicted events for $\hat{A}_{Z,1}$ are nearly constant over time. Since it does not account for the missing data (except via a uniform scaling factor), the network $\hat{A}_{Z,1}$ is not able to capture the dynamics of the process and so it cannot predict events with as much precision as $\hat{A}_{Z,.75}$.

7 Conclusion

We propose a novel estimator for Bernoulli autoregressive models which accounts for partially-observed event data. This model can be used in a variety of contexts in which discrete event data exhibits autoregressive structure. We provide mean squared error bounds which suggest that our method can accurately capture a network's structure in the presence of missing events. Simulations and a real data experiment show that our method yields significant improvement compared with ignoring the missed events and that our method is robust to misspecification of the proportion of missed events. The framework described in this paper suggests a strategy for addressing regression problems with corrupted data in other GLMs, although further work is needed to extend our theoretical analysis beyond binary observations.

Acknowledgements

This work was supported by NGA HM0476-17-1-2003, NSF CCF-1418976, NIH AI117924-01, NSF CCF-0353079, ARO W911NF-17-1-0357, and AFOSR FA9550-18-1-0166.

References

Agarwal, A., Negahban, S., and Wainwright, M. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482.

Alzer, H. (2000). Sharp bounds for the bernoulli numbers. *Archiv der Mathematik*, 74(3):207–211.

Bao, Y., Kwong, C., Peissig, P., Page, D., and Willett, R. (2017). Point process modeling of adverse drug reactions with longitudinal observational data. In *Proc. Machine Learning and Healthcare*.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4):1535–1567.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732.

Biderman, A. (1991). *Understanding Crime Incidence Statistics*. Springer-Verlag, New York, first edition.

Carlitz, L. (1959). Eulerian numbers and polynomials. *Mathematics Magazine*, 32(5):247–260.

Chen, S., Shojaie, A., Shea-Brown, E., and Witten, D. (2017). The multivariate Hawkes process in high dimensions: Beyond mutual excitation. *arXiv preprint arXiv:1707.04928*.

City of Chicago (2018). Crimes - 2001 to present. https://data.cityofchicago.org.

Fletcher, A. and Rangan, S. (2014). Scalable inference for neuronal connectivity from calcium imaging. In *NIPS*. arXiv:1409.0289.

Graham, J. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576.

Hall, E. C., Raskutti, G., and Willett, R. (2016). Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*.

Hao, K. (2019). Police across the us are training crime-predicting ais on falsified data. *MIT Technology Review*. https://www.technologyreview.com/s/612957/predictive-policing-algorithms-ai-crime-dirty-data/.

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

Jalali, A. and Willett, R. (2018). Missing data in sparse transition matrix estimation for sub-gaussian vector autoregressive processes. In *American Control Conference*. arXiv:1802.09511.

Langton, L., Berzofsky, M., Kerbs, C., and Smiley-McDonald, H. (2012). Victimizations not reported to the police, 2006-2010. Technical report, Bureau of Justice Statistics.

Le, T. (2018). A multivariate Hawkes process with gaps in observations. *IEEE Transactions on Information Theory*, 63(3):1800–1811.

Linderman, S., Adams, R., and Pillow, J. (2016). Bayesian latent structure discovery from multi-neuron recordings. In *Advances in neural information processing systems*.

Linderman, S. W. and Adams, R. P. (2014). Discovering latent network structure in point process data. In *ICML*. arXiv:1402.0914.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics*, 42(5):866–896.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics*, 40:1637–1664.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Mark, B., Raskutti, G., and Willett, R. (2018). Network estimation from point process data. *arXiv preprint arXiv:1802.09511*.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497.

Mohler, G., Short, M., Malinowski, S., Johnson, M., Tita, G., Bertozzi, A., and Brantingham, P. (2016). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Palermo, T., Bleck, J., and Peterman, A. (2014). Tip of the iceberg: Reporting and gender-based violence in developing countries. *American Journal of Epidemiology*, 179(5):602–612.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. on Information Theory*, 57(10):6976–6994.

Shelton, C., Qin, Z., and Shetty, C. (2018). Hawkes process inference with missing data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Weiskopf, N. and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151.

Wells, B., Chagin, K., Nowacki, A., and Kattan, M. (2013). Strategies for handling missing data in electronic health record derived data. *EGEMS*, 1(3):1035.

Xu, H., Luo, D., Chen, X., and Carin, L. (2018). Benefits from superposed Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Xu, H., Luo, D., and Zha, H. (2017). Learning Hawkes processes from short doubly-censored event sequences. In *ICML*. arXiv:1702.07013.

Yang, Y., Etesami, J., He, N., and Kiyavash, N. (2017). Online learning for multivariate Hawkes processes. In *NIPS*.

Zhou, K., Zha, H., and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*.