Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 744-754
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144202

Hans汉斯

自激点过程在相依事件中的建模及其应用

戴　乐*，高　犇，李晨龙，郭　平#

太原理工大学数学学院，山西　晋中

收稿日期：2025年3月24日；录用日期：2025年4月17日；发布日期：2025年4月25日

摘　要

相依事件序列广泛存在于金融、社交网络等诸多领域，其事件间具有强烈的依赖性，使得传统事件序列建模方法难以准确刻画其动态特征。为了有效地描述事件间的依赖关系，基于自激点过程构建了扩展模型，将自激函数推广到指数函数和形式，并提出了分布式统计推断方法。通过将数据划分至多个节点进行并行计算，再聚合估计结果，解决了传统极大似然估计在大数据处理中计算成本高和效率低的问题，为自激点过程在大数据中的应用提供了新的解决方案。仿真实验结果表明，分布式估计在有限样本下与传统全局估计表现一致，同时将运行时间缩短约70%。在实证分析中，自激点过程模型有效刻画了Boston犯罪数据和IPTV用户点播行为数据的趋势，并将计算时间分别提升了约95%和64%。

关键词

点过程，大数据建模，分布式推断，相依事件序列

Modeling and Application of Self-Exciting Point Process in Dependent Events

Le Dai*, Ben Gao, Chenlong Li, Ping Guo#

School of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi

Received: Mar. 24th, 2025; accepted: Apr. 17th, 2025; published: Apr. 25th, 2025

Abstract

The sequence of dependent events is widely present in various fields, such as finance and social networks, where strong dependencies between events make it difficult for traditional event sequence modeling methods to accurately capture their dynamic characteristics. To effectively describe these dependencies, an extended model based on the self-exciting point process is constructed,

*第一作者。
#通讯作者。

文章引用: 戴乐, 高犇, 李晨龙, 郭平. 自激点过程在相依事件中的建模及其应用[J]. 应用数学进展, 2025, 14(4): 744-754. DOI: 10.12677/aam.2025.144202

generalizing the self-exciting function to a sum of exponential functions. A distributed statistical inference method is proposed, which divides the data across multiple nodes for parallel computation and then aggregates the estimation results. This approach addresses the high computational cost and low efficiency of traditional maximum likelihood estimation in big data processing, providing a new solution for applying self-exciting point processes to big data. Simulation experiments show that the distributed estimation performs consistently with traditional global estimation on limited samples while reducing runtime by approximately 70%. In empirical analysis, the self-exciting point process model effectively captures trends in Boston crime data and IPTV user on-demand behavior data, improving computation time by approximately 95% and 64%, respectively.

Keywords

Point Process, Big Data Modeling, Distributed Inference, Dependent Event Sequence

1. 引言

在众多领域，如金融、社交网络、流行病学等，相依事件序列广泛存在，它是按时间顺序发生的离散事件集合。与独立事件序列不同，相依事件序列中的事件呈现出强烈的相依性特征。对其建模旨在描述事件间相互作用、影响关系及时空分布特征，为实际问题提供分析与预测方法，但面临着对复杂时间依赖性和因果关系进行建模困难的挑战。传统建模方法如马尔可夫过程[1] [2]和泊松过程[3] [4]，假设事件独立发生，难以满足现实需求。

随着自激点过程模型的提出[5]，相依事件序列的建模取得了重要进展。自激点过程通过引入事件发生后的自激机制，有效地描述了事件间的依赖性，在地震学[6] [7]、金融市场[8] [9]、社交媒体[10] [11] [12]等领域广泛应用。相比于传统的独立事件模型，自激点过程能够更准确地刻画事件之间的相互依赖和自激效应，尤其是在事件发生具有明显群体效应或连锁反应时，自激点过程表现出了良好的建模能力。在实际应用中，研究人员往往需要根据具体场景调整自激点过程的结构，以刻画不同的事件序列特征。

在自激点过程模型中，参数估计至关重要。常用的极大似然估计虽具有一致性，但大多数情况下，自激点过程的似然函数没有解析解，使得利用似然函数进行极大似然估计成为一个非线性优化问题，导致计算成本随着数据量的增加而变大，限制了参数型自激点过程在大数据场景下的应用和推广。特别是在高维数据和大规模事件序列的情境下，极大似然估计方法在计算上较为复杂，计算量急剧增加。因此，如何在保证估计精度的同时减少计算成本，是自激点过程参数估计中的一个重要挑战。

针对这一问题，提出了一种创新性的分布式统计推断方法。常用的分布式推断方法[13]是基于数据分割的并行计算方法。通常将数据划分为若干部分，并在不同的计算节点上分别进行模型参数的估计，最后通过一定的算法将各个节点的估计结果进行合并。该方法通过分布式计算的方式，在保证估计精度的同时，并且能够显著提高计算效率，特别是在数据量大的情况下，能够有效解决传统集中式方法所面临的计算问题。在仿真实验以及对 Boston 犯罪行为分析和互联网协议电视用户点播行为的实证分析中，结果表明，当数据量较大时，分布式统计推断不仅能够保证估计精度，还在计算时间上表现出显著的优势。

目前，自激点过程分布式统计推断研究稀缺，现有研究多集中于特定类型，如文献[14]中探讨的指数

型自激点过程的分布式统计推断。但指数型自激点过程强度函数形式单一，应用范围受限，难以适应复杂多变的事件序列。在此基础上，将自激函数推广到指数和形式，即自激函数为

$$g(t;\mu)=\sum_{q=1}^{m}a_q\exp(-b_q t).$$

通过引入复杂函数形式，能更好捕捉事件间相互依赖关系的多样性和非线性特征，提升自激点过程在复杂数据分析中的应用范围和准确性。

2. 模型背景

2.1. 自激点过程

自激点过程 N_t 的**条件强度函数** $\lambda(t;\theta)$ 定义为：

$$\lambda(t;\theta)=v+\int_0^t g(t-s;\mu)\mathrm{d}N_s. \tag{1}$$

其中 $\theta=(v,\mu^{\mathrm{T}})^{\mathrm{T}}$ 表示条件强度函数的**参数向量**，μ 表示自激函数的参数向量。v 表示**基本背景强度**，$g(t;\mu)$ 表示**自激函数**，μ 表示自激函数的参数，且对 $t\in\mathbb{R}_+$ 有 $\int_0^t g(t-s;\mu)\mathrm{d}N_s=\sum_{t_i<t}g(t-t_i;\mu)$。$g(t;\mu)$ 需满足 $\sup_{\mu}\int_0^{+\infty}g(t;\mu)\mathrm{d}t<1$ [15]。(1)式在犯罪行为和电视点播行为的分析中具有很好的可解释性。例如在犯罪行为的研究中，参数 v 通常表示犯罪活动的背景事件(例如，一次重大犯罪事件或社会事件)的影响，而自激函数则描述了该事件与后续犯罪行为之间的相互影响过程，即，当前及历史犯罪行为如何影响未来犯罪发生的可能性。

对于自激函数 $g(t;\mu)$，在不同的应用背景中往往采用不同的参数形式。常见的应用背景及其自激函数类型如表 1 所示：

Table 1. Common types of self-exciting functions
表 1. 常见的自激函数类型

应用背景	函数名称	函数形式
社交媒体中消息的传播分析； 金融市场中的交易分析[16] [17]；	指数型自激函数	$g(t)=a\cdot\exp(-bt)$
地震序列等自然灾害研究[18]；	幂律性自激函数	$g(t)=a\cdot(t+k)^{-l}$
神经科学中的脑电图分析、信号处理中的滤波分析[19]；	高斯型自激函数	$g(t)=a\cdot\exp\left\{-\left(\frac{t}{b}\right)^2\right\}$

2.2. 极大似然估计

观测区间 $[0,T]$ 上自激点过程的**对数似然函数**为：

$$L(\theta)=\int_0^T\log\lambda(t;\theta)\mathrm{d}N_t-\int_0^T\lambda(t;\theta)\mathrm{d}t. \tag{2}$$

有 $\int_0^T\log\lambda(t;\theta)\mathrm{d}N_t=\sum_{i=1}^{n}\left\{\log\left[v+\sum_{t_j:t_j<t_i}g(t_i-t_j;\mu)\right]\right\}$，$t_i$ 为第 i 个事件的发生时间点。参数向量 θ 在观测区间[0, T]上的极大似然估计为：

$$\hat{\theta}=\arg\max_{\theta}L(\theta). \tag{3}$$

在一些正则条件下，$\hat{\theta}$ 是一致估计，并且 $\sqrt{T}(\hat{\theta}-\theta)$ 具有渐近正态性[20]。极大似然估计操作简便，

能够提供给定数据下的一致性估计，具有良好的统计性质。然而，当观测数据量较大时，使用(3)式进行求解往往没有解析解，此时极大似然估计变成一个非线性优化问题的数值解。随着数据量的增加，计算成本也显著上升。这限制了传统集中式的极大似然估计在大数据场景中的应用和推广。

3. 模型介绍及其理论结果

3.1. 模型设置

文献[14]研究的是在固定的观测区间$[0,T]$内，事件呈爆炸性增长，即T是固定的，其主要探讨高频数据下自激点过程的统计性质。高频数据能捕捉到事件短期内的精细变化，在一些对时效性要求高的场景，如高频金融交易分析中，可及时反映市场瞬间波动。然而在现实应用场景中，观测时间并非一成不变。特别是在长期观察或分析动态过程行为时，低频数据更为常见。低频数据涵盖的时间跨度大，能展现事件在较长时期内的整体趋势。若继续采用固定时间的模型，难以精准描述真实的事件发生规律。因此，模型考虑了观测区间长度趋于无穷的情况，以更全面地刻画事件在不同时间尺度下的行为特征。实际中，许多现象(如地震、金融市场波动、社交网络传播等)往往没有明确的结束时间，因此研究长时间尺度上的事件动态更加贴近实际应用。对于自激函数，指数型自激函数由于其依赖的事件模式相对固定，不能灵活适应实际中复杂多样的依赖方式。幂律性自激函数缺乏对短期波动的适应性，无法很好描述过程中的短期动态特征。高斯型自激函数对非高斯数据适应性差，可能会产生不符合实际的结果。基于自激点过程构建了扩展模型，将自激函数推广到指数和形式，即自激函数为

$$g(t;\mu)=\sum_{q=1}^{m}a_q\cdot\exp\left(-b_q t\right).$$

记$\mu=\left(a_1,a_2,\cdots,a_m,b_1,b_2,\cdots,b_m\right)^{\mathrm{T}}$，即通过多个指数型函数的加和来建模系统中事件的动态演化。这一创新的自激函数设计能够更灵活地描述不同类型的事件依赖关系，克服了单一自激函数在复杂动态系统中的不足。

3.2. 分布式估计

在本小节，给出模型参数估计的方法。根据极大似然估计方法，分布式局部极大似然估计和分布式全局极大似然估计的构造具体步骤如下：

1) 在观测区间$[0,T]$上，考虑区间长度为h_T的小观测区间，即将$[0,T]$分成规则的非重叠的$B_T=Th_T^{-1}$个小区间，第i小区间为$\left((i-1)h_T,ih_T\right]$，$i\in\{1,2,\cdots,B_T\}$，这里取$h_T=T^{1/\delta}$，$\delta\geq 1$。

2) 在每个区间上求出θ的局部估计$\hat{\theta}_i$：$\forall i\in\{1,2,\cdots,B_T\}$，$\theta\in\Theta$，其中$\Theta$为参数空间，定义第$i$个区间上的**条件强度函数**：

$$\lambda_i(t;\theta)=v+\int_{(i-1)h_T}^{t}g(t-s;\mu)\mathrm{d}N_s, \tag{4}$$

其中$t\in\left((i-1)h_T,ih_T\right]$。定义第$i$个区间上的**局部对数似然函数**：

$$L_i(\theta)=\int_{(i-1)h_T}^{ih_T}\log\lambda_i(t;\theta)\mathrm{d}N_t-\int_{(i-1)h_T}^{ih_T}\lambda_i(t;\theta)\mathrm{d}t. \tag{5}$$

计算第i个区间上θ的**局部极大似然估计(分布式局部估计)**：

$$\hat{\theta}_i=\arg\max_{\theta}L_i(\theta). \tag{6}$$

3) 对分布式局部估计$\hat{\theta}_i$，$i\in\{1,2,\cdots,B_T\}$进行聚合得到**分布式全局估计**$\hat{\theta}$，即，

$$\hat{\theta}:=\frac{1}{B_T}\sum_{i=1}^{B_T}\hat{\theta}_i. \tag{7}$$

分布式估计相比传统集中估计在计算效率上具有明显优势。通过将数据分布到多个计算节点并行处理，分布式估计避免了单节点计算瓶颈，显著加速了计算过程。此外，分布式方法能够按需扩展计算资源，提升了对大数据的处理能力，并降低了对存储和内存的需求。因此，分布式估计在大规模数据环境下能够有效提高计算效率，减少计算时间的压力。

3.3. 理论结果

本小节给出自激函数为指数函数和型自激点过程的分布式统计推断理论结果。对于参数 $\theta=\left(v,\mu^{\mathrm{T}}\right)^{\mathrm{T}}$，假设：

1) 存在两个非负向量 $\underline{\theta}$ 和 $\overline{\theta}$，使得 $\theta\in K:=\left\{w\in\mathbb{R}^{2m+1}:0<\underline{\theta}\le w\le\overline{\theta}\right\}$，其中 $\underline{\theta}$ 和 $\overline{\theta}$ 为 $\underline{\theta}=\left(\underline{v},\underline{a}_1,\cdots,\underline{a}_m,\underline{b}_1,\cdots,\underline{b}_m\right)^{\mathrm{T}}$，$\overline{\theta}=\left(\overline{v},\overline{a}_1,\cdots,\overline{a}_m,\overline{b}_1,\cdots,\overline{b}_m\right)^{\mathrm{T}}$；

2) $\sum_{q=1}^{m}a_q<\underline{b}$，其中 $\underline{b}$ 表示 $\underline{\theta}$ 中关于 b 的最小值，即 $\underline{b}=\min\limits_{i\in\{1,2,\cdots,m\}}\underline{b}_i$。类似的有 $\overline{b}=\max\limits_{i\in\{1,2,\cdots,m\}}\overline{b}_i$、$\underline{a}=\min\limits_{i\in\{1,2,\cdots,m\}}\underline{b}_i$ 和 $\overline{a}=\max\limits_{i\in\{1,2,\cdots,m\}}\overline{a}_i$。

基于上述假设下，有以下两个理论结果。

定理 1 [20]：由(6)式得到的分布式局部估计 $\hat{\theta}_i$ 满足：$\forall i\in\{1,2,\cdots,B_T\}$，在 $T\to+\infty$ 时，一致的有

$$\sqrt{h_T}\left(\hat{\theta}_i-\theta^*\right)\xrightarrow{\mathrm{d}}\Gamma^{-1/2}\xi. \tag{8}$$

且有 $\mathbb{E}\left\{f\left[\sqrt{h_T}\left(\hat{\theta}_i-\theta^*\right)\right]\right\}\longrightarrow\mathbb{E}\left[f\left(\Gamma^{-1/2}\xi\right)\right]$，其中 ξ 服从标准正态分布，Γ 是参数为 $\theta^*=\left(v^*,\mu^*\right)$ 的自激点过程的 Fisher 信息阵，f 是任意多项式增长的连续函数，“$\xrightarrow{\mathrm{d}}$”表示依分布收敛。

定理 2 [20]：由(7)式得到的 $\hat{\theta}$ 满足

$$\sqrt{T}\left(\hat{\theta}-\theta^*\right)\xrightarrow{\mathrm{d}}\Gamma^{-1/2}\xi. \tag{9}$$

其中 Γ 是参数为 $\theta^*=\left(v^*,\mu^*\right)$ 的自激点过程的 Fisher 信息阵，ξ 服从标准正态分布。

定理 1 和定理 2 的证明可参考文献[20]。由上述定理可知，分布式局部估计和全局估计都具有渐近正态性，并以不同的速率收敛。特别地，由定理 2 可知，分布式全局估计具有与直接使用全部数据得到的极大似然估计相同的渐近方差。也就是说定理 2 保证了通过合理的聚合策略能够有效地减少计算成本，并保证了估计的收敛速度和估计精度。

4. 仿真实验

4.1. 仿真参数设置

本节给出自激点过程的极大似然估计仿真实验及结果。为评估分布式估计有效性与可行性，开展 3 组仿真实验，与传统全局估计对比，分析不同阈值对参数估计精度的影响，讨论模型定阶问题，验证大数定律。仿真实验主要是针对二阶指数函数和型自激点过程，仿真实验的真实参数设置为 $\theta=\left(v,a_1,a_2,b_1,b_2\right)=\left(0.05,1,3,4,8\right)$，观测时长为 T = 2,000,000，产生的仿真数据大约为 260,000 个。下面给出各仿真实验结果。在仿真实验中，total_samples 表示传统式全局估计，local_10 表示观测区分成 10 个小区间的分布式全局估计，local_20 表示观测区分成 20 个小区间的分布式全局估计。

4.2. 截断估计

对于点过程 N_t，全局极大似然函数表达式为 $L(\theta)=\int_0^T\log\lambda\left(t;\theta\right)\mathrm{d}N_t-\int_0^T\lambda\left(t;\theta\right)\mathrm{d}t$，其中有积分项 $\int_0^T\log\lambda\left(t;\theta\right)\mathrm{d}N_t=\sum_{i=1}^{n}\left\{\log\left[v+\sum\limits_{t_j:t_j<t_i}g\left(t_i-t_j;\mu\right)\right]\right\}$。若按照定义计算，大样本全局估计的计算成本将会变

得很大，注意到当指数项足够大时，指数函数能够快速收敛到 0，因此在本组仿真实验中，对于任意时间点 t_i，将不再计算满足条件 $\{t_j : t_j < t_i\}$ 所有的 t_j，只取满足条件 $\{t_j : t_j < t_i\}$ 最大的 m 个，其中截断阈值分别取 50、100、200、400、600、800、1000、1200、1400、1600、1800、2000。在不同的截断阈值下，传统式全局估计与分布式全局估计的不同参数的偏差和估计的计算时间关系如图 1 所示：

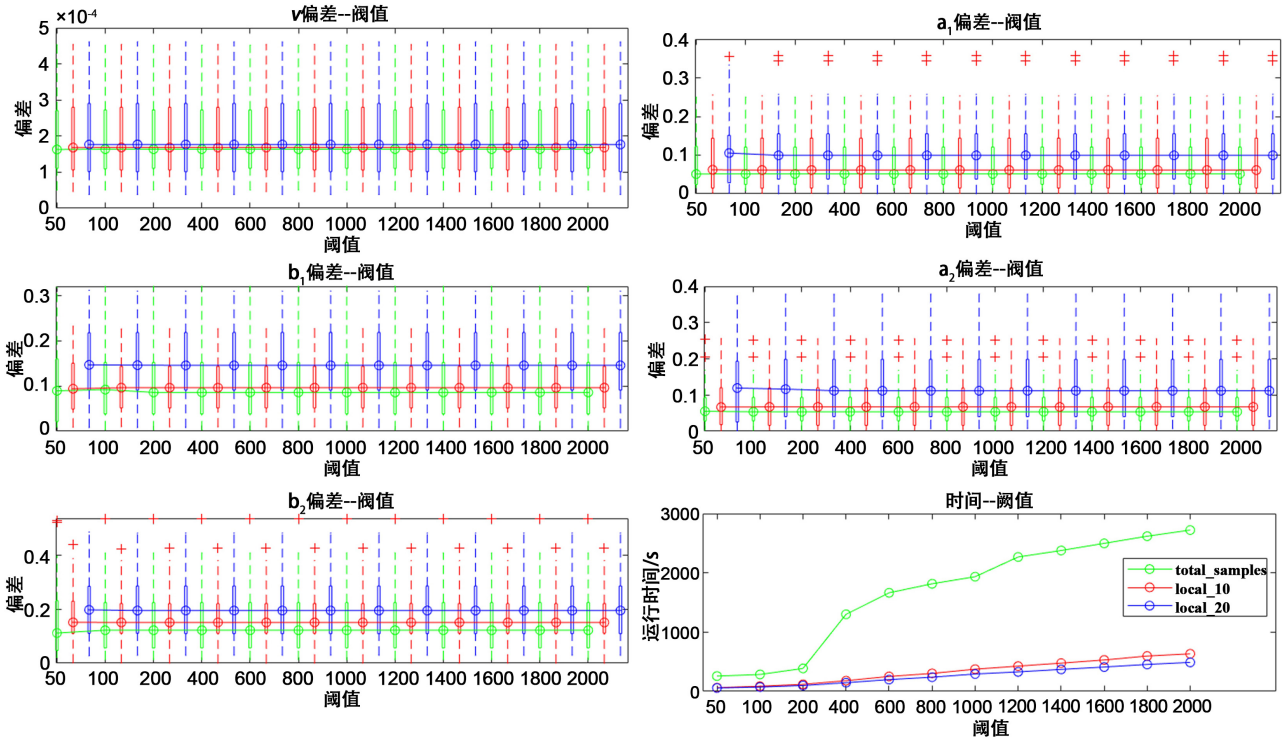

Figure 1. Relationship between deviations of parameter estimates and thresholds and computational time and thresholds
图 1. 各参数估计值的偏差与阈值的关系以及计算时间与阈值的关系

4.3. 模型的定阶

在本小节中，将给出 2 阶指数型自激点过程进行定阶实验的仿真结果。首先通过指定自激点过程参数产生仿真数据，然后根据不同样本量(1000、5000、10,000、50,000、100,000、所有样本)的仿真数据对 1 到 5 阶的指数型自激点过程进行参数估计。主要是采用两种常规的定阶方法：赤池信息准则 $\mathrm{AIC} = 2k - 2\ln(L)$ 和贝叶斯信息准则 $\mathrm{BIC} = \ln(n) * k - 2\ln(L)$，其中 k 为参数个数，n 是数据量数。通过比较 AIC 和 BIC 的值选出定阶结果。若阶数为 k，自激函数为 $g(t;\mu) = \sum_{q=1}^{k} a_q \cdot \exp(-b_q t)$。不同样本量的定阶结果如下表 2 所示：

Table 2. AIC and BIC values for different sample sizes
表 2. 不同样本量的 AIC、BIC 值

阶数 \ 样本量	1000	5000	10,000	50,000	100,000	所有样本
1	2600.3905 2615.1381	13049.0055 13072.6225	26177.4667 26199.0978	130611.9834 130560.4866	261624.6172 261458.2289	696634.1314 776764.6617
2	2600.8014 2625.3402	13040.0364 13068.5572	26155.9715 26192.0232	130495.6294 130539.7283	261388.2269 261435.7915	696045.0156 745065.9191

续表

3	2604.0142 2639.1557	13043.6248 13089.2451	26159.7314 26210.0239	130498.7481 130582.1374	261391.6384 261481.2391	696048.6934 809056.1979
4	2608.8042 2652.9740	13047.6245 13106.2793	26163.7304 26228.6235	130502.7593 130603.7670	261395.6227 809056.1979	696052.6560 841348.0190
5	2612.8410 2666.8263	13051.6246 13123.3138	26167.7357 26247.0495	130506.7494 130638.4428	261399.6200 261653.1560	696056.7039 873639.9254

在表 2 中，第一行为 AIC 值，第二行为 BIC 值，通过比较 AIC 和 BIC 的值，当样本量较少时(1000)，模型选出的阶数是 1 阶，可能原因是样本量比较少，而且高阶模型的复杂度比较高，导致估计不准确，因此计算的 AIC 和 BIC 的值会变大；当样本量较多时(≥5000)，模型选出的阶数是均 2 阶。这表明在适当的样本量下，可以准确选择模型的阶数。

4.4. 仿真结果

在本小节，将验证 3 种方式的全局估计均满足大数定律。在 3.2 节的基础上，令截断阈值 $m=1000$，重复实验 200 次，分别得到 3 种估计方式的各参数的频率直方图，如图 2 所示：

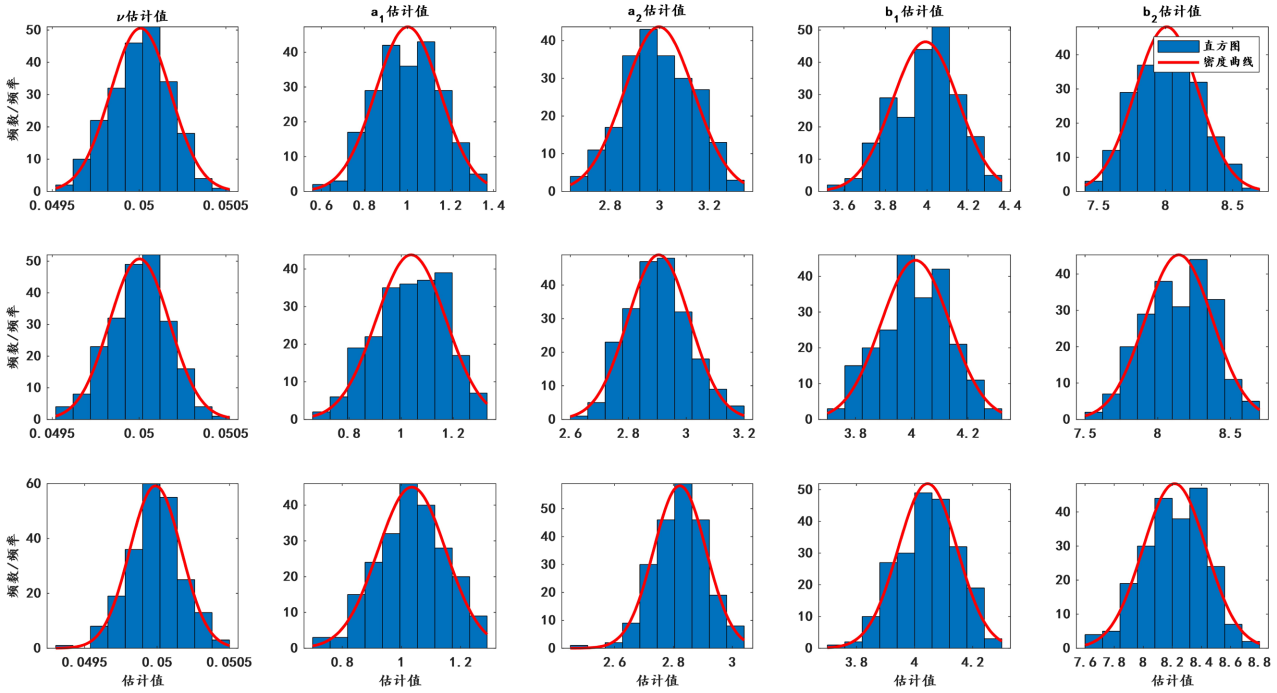

Figure 2. Frequency histogram of parameter estimates
图 2. 各参数估计值的频率直方图

Table 3. Asymptotic variance
表 3. 渐近方差

参数 / 估计方式	v (*10^{-3})	a_1	a_2	b_1	b_2
total_samples	2.9560	0.0132	0.0079	0.0103	0.0475
local_10	2.9942	0.0181	0.0115	0.0150	0.0541
local_20	3.0100	0.0226	0.0198	0.0253	0.0560

图 2 中，第一行为传统式全局估计(total_samples)的各参数频率直方图，第二行为将观测区分成 10 个(local_10)小区间的分布式全局估计的各参数频率直方图，第三行为将观测区间分成 20 个(local_20)小区间的分布式全局估计的各参数频率直方图。不同的估计方式与其参数估计的方差如表 3 所示。

图 2 说明了分布式全局估计和传统式全局估计在有限样本下都表现出渐近正态性，表明本文的估计方法是稳定可靠的，从而确保了估计结果的稳定性和可靠性。在有限样本下，具有与传统式全局估计方法一致的表现。

5. 实证分析

选取了 Boston 市内犯罪数据、互联网协议电视(IPTV)的用户点播行为数据进行实证分析。

5.1. Boston 市内犯罪数据

该数据集包含自 2015 年 6 月至 2018 年 8 月期间 Boston 市内的犯罪数据，选取特定区域(经度范围[−71.1275, −71.0275]，纬度范围[42.2755, 42.3755])内的犯罪数据作为实验数据。该数据共计包含 183,767 条记录。

首先利用 AIC、BIC 确定阶数，具体定阶结果如表 4 所示：

Table 4. Order determination results of Boston crime data
表 4. Boston 犯罪数据定阶结果

准则 \ 阶数		1 阶	2 阶	3 阶	4 阶	5 阶
AIC	total_samples	341575.2436	341579.2436	341583.2432	341587.2434	341591.2436
	local_10	320766.6849	320770.6852	320774.6804	320778.6818	320782.6847
	local_20	320763.6427	320771.5128	320775.6848	320778.5418	320783.1541
BIC	total_samples	341605.6079	341629.8507	341654.0932	341678.3363	341702.5792
	local_10	320797.0491	320821.2923	320845.5303	320869.7741	320894.0204
	local_20	320799.0489	320840.2906	320851.5438	320875.0186	320890.7760

由上表可得模型的阶数为 1 阶。接下来对实验数据进行分布式统计推断实验。将实验数据分成 10 个小区间(local_10)和 20 个小区间(local_20)分别进行分布式统计推断，根据(7)式计算分布式全局估计，再根据(2)式和(3)式，计算出传统式的全局极大似然估计(total_samples)，对传统式的全局估计和分布式全局估计进行对比，其结果如表 5 所示：

Table 5. Results of Boston crime data
表 5. Boston 犯罪数据结果

估计方式 \ 参数	v	a	b	运行时间/s	预测事件数
total_samples	1.5484	0.6208	0.8146	33925.3244	167
local_10	1.5768	0.6203	0.8183	1709.3994	165
local_20	1.6002	0.6204	0.8220	898.2922	140

根据实验结果，犯罪事件的自激效应与背景效应存在显著数量关系，每发生一个背景犯罪事件，平均引发约 4 个自激犯罪事件，表明犯罪事件相互影响、连锁反应，呈现自我增强和蔓延趋势，前一次犯

罪会显著提高后续犯罪发生概率。研究人员用自激点过程模型预测未来 1 小时犯罪事件数量，预测值与实际观测值(154)高度一致，验证了模型有效性和犯罪事件的自激效应。

基于自激点过程的分析为制定精准的犯罪防控措施提供了依据。研究表明，通过调节外部因素，如社会经济条件、警力时空配置效率以及城市空间规划等，可以有效减少犯罪活动的内生驱动强度，从而降低犯罪的蔓延。具体而言，政府可以针对高自激效应的区域和时段进行精准干预。例如，针对犯罪高发时段进行警力集中部署，增强防控力度；通过改善社会经济条件、减缓贫困和失业等问题，缓解社会矛盾，降低犯罪发生的诱因；而合理的城市规划可以有效消除犯罪滋生的“盲区”，减少犯罪的发生。此外，政府还可以利用该模型进行犯罪预测和预警，在犯罪高发的时间和区域采取预防措施，从而提高社会安全管理的精准度。

5.2. 互联网协议电视(IPTV)的用户点播行为

该数据集包含 2012 年 302 个用户在互联网协议电视(IPTV)点播的行为数据，选定用户的点播数据作为实验数据，总共有 4491 条点播行为记录。首先需要确定模型阶数，该数据集的定阶结果如表 6 所示：

Table 6. Order determination results of IPTV on-demand behavior data
表 6. IPTV 点播行为数据定阶结果

准则 \ 阶数		1 阶	2 阶	3 阶	4 阶	5 阶
AIC	total_samples	11110.9500	11137.9647	11141.9647	11145.9647	11149.9647
	local_10	10985.4917	11023.5218	11058.6351	11084.8657	11097.2463
	local_20	10320.7874	10330.5008	10343.1570	10371.4490	10395.0043
BIC	total_samples	11130.1795	11170.0139	11186.8335	11203.6531	11220.4728
	local_10	11030.3605	11042.7513	11090.6842	11127.3871	11142.5542
	local_20	10340.0169	10375.2062	10395.6959	10435.2693	10442.1957

由上表可知，对于该数据集模型的定阶结果为 1 阶。其估计结果如表 7：

Table 7. Results of IPTV on-demand behavior data
表 7. IPTV 点播行为数据结果

估计方式 \ 参数	v	a	b	运行时间/s	预测事件数
total_samples	0.0457	4.1902	4.5612	33.9636	156
local_10	0.0472	3.6815	4.6571	12.7093	122
local_20	0.0478	3.7363	4.5612	2.3761	137

通过自激点过程对 IPTV 点播行为建模发现，背景强度小于自激强度，观众点播行为存在自我增强效应，每次点播都会增加未来点播概率，表明观众观影行为与其过去行为紧密相关。基于参数估计预测未来 1 天点播次数，真实观测值为 138，对比显示模型有效，能准确捕捉点播行为趋势。

自激点过程模型的引入，为点播行为的分析带来了较高的研究价值。它不仅揭示了点播行为的内在规律，还为平台提供了精确预测和干预的工具，推动了内容推荐、平台运营及用户行为管理等方面的创新。通过对点播行为的精准分析，平台能够在竞争激烈的市场环境中获得竞争优势，确保用户的长期留存和平台的稳定发展。同时，借助这一模型，平台还可以更灵活地调整运营策略，进行更加精准的市场

定位和用户细分，从而提升整体的服务质量和用户体验。因此，基于自激点过程的点播行为分析，不仅具有深远的理论意义，也为实际操作提供了可行的解决方案，是未来 IPTV 平台持续优化和发展的核心工具。

6. 结论

通过探讨利用具有指数和形式的自激点过程对相依事件序列进行建模及相应的参数快速估计方法—分布式估计策略，在理论创新和实践应用两个层面取得了重要进展：1) 在理论层面，构建了完整的分布式估计框架，通过严格的数学推导证明了该方法满足大数定律，并具有渐近正态性，为分布式统计推断的可靠性提供了理论支撑；2) 在实践应用层面，仿真实验和实证分析均表明，与传统集中式方法相比，所提出的分布式策略在计算效率方面具有显著优势，这一优势随着数据规模的扩大和计算阈值的提高而愈加明显。通过对 Boston 犯罪数据集和 IPTV 点播行为数据的实证研究，验证了该方法在大样本和小样本下均能保持较高的时间效率，为分析复杂社会现象提供了新的研究工具。

然而，分布式统计推断方法仍存在着一些局限性。首先，由于数据分割导致的局部信息损失，在数据量有限或分布不均匀时可能引入估计偏差；其次，参数化模型在刻画数据动态特征方面存在固有局限，难以有效捕捉长期趋势和周期性波动。针对这些不足，未来的研究方向将聚焦于以下两个方面：1) 对偏差进行校正，通过引入正则化项或设计加权估计策略来降低分割偏差；2) 探索非参数化自激点过程的分布式推断方法，以增强模型对复杂数据特征的适应能力。

基金项目

山西省回国留学人员科研资助项目(2024-034)，山西省自然科学基金(No.20210302124081)。

参考文献

[1] Baum, L.E. and Petrie, T. (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, **37**, 1554-1563. https://doi.org/10.1214/aoms/1177699147

[2] Rabiner, L.R. (1990) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Waibel, A. and Lee, K.-F., Eds., *Readings in Speech Recognition*, Elsevier, 267-296. https://doi.org/10.1016/b978-0-08-051584-7.50027-9

[3] Trinh, M. (2018) Non-Stationary Processes and Their Application to Financial High-Frequency Data. University of Sussex.

[4] Cifuentes-Amado, M.V. and Cepeda-Cuervo, E. (2015) Non-Homogeneous Poisson Process to Model Seasonal Events: Application to the Health Diseases. *International Journal of Statistics in Medical Research*, **4**, 337-346. https://doi.org/10.6000/1929-6029.2015.04.04.4

[5] Hawkes, A.G. (1971) Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, **58**, 83-90. https://doi.org/10.1093/biomet/58.1.83

[6] Kwon, J., Zheng, Y. and Jun, M. (2023) Flexible Spatio-Temporal Hawkes Process Models for Earthquake Occurrences. *Spatial Statistics*, **54**, Article ID: 100728. https://doi.org/10.1016/j.spasta.2023.100728

[7] Shah, R., *et al*. (2020) Temporal Point Process Models for Nepal Earthquake Aftershocks. *International Journal of Statistics and Reliability Engineering*, **7**, 275-285.

[8] Ding, X., Shi, J., Duan, J., Qin, B. and Liu, T. (2021) Quantifying the Effects of Long-Term News on Stock Markets on the Basis of the Multikernel Hawkes Process. *Science China Information Sciences*, **64**, Article ID: 192102. https://doi.org/10.1007/s11432-020-3064-4

[9] Zhuo, J., Chen, Y., Zhou, B., *et al*. (2023) A Hawkes Process Analysis of High-Frequency Price Endogeneity and Market Efficiency. *The European Journal of Finance*, **30**, 1-31.

[10] Kobayashi, R. and Lambiotte, R. (2021) Tideh: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. *Proceedings of the International AAAI Conference on Web and Social Media*, **10**, 191-200. https://doi.org/10.1609/icwsm.v10i1.14717

[11] Nie, H.R., Zhang, X., Li, M., Dolgun, A. and Baglin, J. (2020) Modelling User Influence and Rumor Propagation on

Twitter Using Hawkes Processes. 2020 *IEEE 7th International Conference on Data Science and Advanced Analytics* (*DSAA*), Sydney, 6-9 October 2020, 637-656. https://doi.org/10.1109/dsaa49011.2020.00090

[12] Goda, M., Mizuno, T. and Yano, R. (2022) Hawkes Process Marked with Topics and Its Application to Twitter Data Analysis. *Europhysics Letters*, **140**, 61001. https://doi.org/10.1209/0295-5075/aca78c

[13] Gao, Y., Liu, W., Wang, H., Wang, X., Yan, Y. and Zhang, R. (2021) A Review of Distributed Statistical Inference. *Statistical Theory and Related Fields*, **6**, 89-99. https://doi.org/10.1080/24754269.2021.1974158

[14] Clinet, S. and Potiron, Y. (2018) Statistical Inference for the Doubly Stochastic Self-Exciting Process. *Bernoulli*, **24**, 3469-3493. https://doi.org/10.3150/17-bej966

[15] Ozaki, T. (1979) Maximum Likelihood Estimation of Hawkes' Self-Exciting Point Processes. *Annals of the Institute of Statistical Mathematics*, **31**, 145-155. https://doi.org/10.1007/bf02480272

[16] Oakes, D. (1975) The Markovian Self-Exciting Process. *Journal of Applied Probability*, **12**, 69-77. https://doi.org/10.2307/3212408

[17] Hardiman, S.J., Bercot, N. and Bouchaud, J. (2013) Critical Reflexivity in Financial Markets: A Hawkes Process Analysis. *The European Physical Journal B*, **86**, 1-9. https://doi.org/10.1140/epjb/e2013-40107-3

[18] Pratiwi, H., Slamet, I., Saputro, D.R.S. and Respatiwulan, (2017) Self-Exciting Point Process in Modeling Earthquake Occurrences. *Journal of Physics*: *Conference Series*, **855**, Article ID: 012033. https://doi.org/10.1088/1742-6596/855/1/012033

[19] Zhang, R., Walder, C. and Rizoiu, M. (2020) Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 6803-6810. https://doi.org/10.1609/aaai.v34i04.6160

[20] Clinet, S. and Yoshida, N. (2017) Statistical Inference for Ergodic Point Processes and Application to Limit Order Book. *Stochastic Processes and Their Applications*, **127**, 1800-1839. https://doi.org/10.1016/j.spa.2016.09.014