INTERFACE

rsif.royalsocietypublishing.org

研究

投稿至期刊的研究文章

学科领域：
计算生物学、计算数学、医疗建模

关键词：
镰状细胞病，随机过程，霍克斯过程，最大似然估计

通讯作者：
Kumar Utkarsh
电子邮件：krutkarsh@u.northwestern.edu

arXiv:2601.04538v1 [stat.ME] 8 Jan 2026

一种增强短时间序列的新方法及其在镰状细胞病疼痛事件中的应用

KumarUtkarsh [1], Nirmish R. Shah [4], Tanvi Banerjee [5], Daniel M. Abrams [1,2,3]

[1]西北大学工程科学与应用数学系，美国伊利诺伊州埃文斯顿[2]西北大学复杂系统研究所，西北大学，美国伊利诺伊州埃文斯顿[3]西北大学物理与天文学系，美国伊利诺伊州埃文斯顿[4]杜克大学医学院，美国北卡罗来纳州达勒姆[5]莱特州立大学计算机科学与工程系，美国俄亥俄州代顿

包括但不限于生态学、生物学和医疗保健在内的不同领域的研究人员，常常面临数据稀疏这一挑战。这种稀疏性可能导致不确定性、估计困难以及建模中的潜在偏差。在此，我们提出一种新颖的数据增强方法，在多个稀疏时间序列数据集具有相似统计性质时将其结合起来，从而提高参数估计和模型选择的可靠性。我们通过比较霍克斯过程与泊松过程的验证研究，展示了该方法的有效性；随后将其应用于镰状细胞病（SCD）患者的主观疼痛动态。镰状细胞病是一种影响全球数百万人的疾病，尤其多见于非洲人、地中海地区人群、中东人群和印度人群。

1. 引言

患者报告数据（包括来自可穿戴设备的数据）近年来在医疗保健环境中日益受到重视 [1–9]。然而，这类数据往往具有内在的变异性，包括报告频率和内容的波动，这使分析变得复杂。此外，与健康事件相关的时间动态（包括症状波动或治疗反应）带来了诸多分析挑战。在数据足够丰富时，随机模型可为分析此类数据提供工具，使研究人员和临床医生能够从随时间观察到的模式中获取有用信息 [10–12]。然而，若试图从过于有限的一组观察结果中得出结论，可能会导致过拟合、估计不可靠以及难以进行泛化。 [13–15]

为解决这些问题，研究人员开发了多种方法，例如贝叶斯方法，将先验信息和不确定性纳入考量[16]，；层次模型则汇集相关群组的数据，以改进稀疏数据条件下的估计 [17]。此外，自助法等增强技术可用于生成合成数据点，从而提高统计分析的稳健性 [18]。然而，鲜有研究

© 作者。英国皇家学会依据知识共享署名许可协议 http://creativecommons.org/licenses/by/4.0/ 发布，依据该协议，在注明原作者和来源的前提下，可不受限制地使用。

THE ROYAL SOCIETY PUBLISHING

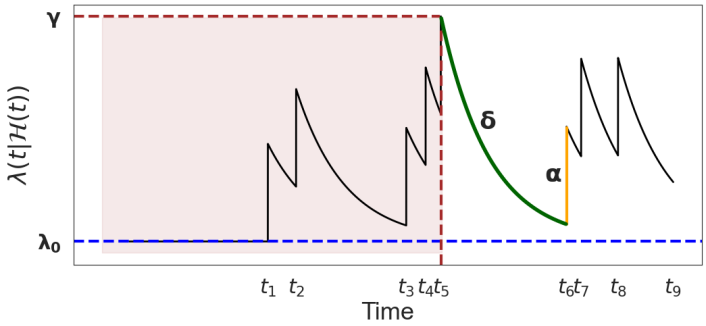


图 1. 移位霍克斯过程参数与强度动态示意图。式 (2.2) 中引入的参数说明（另见表 1）。峰值表示实时事件到达。阴影区域表示观测数据中未捕获的历史记录。在此示例中，可观测的 $t=0$ 为 t_5。蓝色水平虚线表示基线强度 λ_0 的值，红色水平虚线表示初始强度值 γ。影响幅度 α 以黄色垂直线段的长度表示，而衰减率 δ 则刻画指数衰减曲线，例如图中绿色曲线。

参数	说明	单位
λ_0	底层齐次泊松过程的基线强度值。	$[T^{-1}]$
α	单个事件到达对强度的影响幅度。	$[T^{-1}]$
δ	衰减至基线强度的速率（δ^{-1} 设定记忆长度）。	$[T^{-1}]$
γ	在数据收集期开始时记录的强度测量值。	$[T^{-1}]$

表 1。式 (2.2) 中引入的模型参数说明（另见图 1）。

此前的研究是在动力系统背景下针对稀疏采样开展的，而这类系统本身具有很强的时间相关性。本文提出了一种适用范围广泛的新方法，用于分析动力过程中稀疏采样的数据；我们尤其着重于在真实场景中测试该方法，所用数据记录了镰状细胞病患者的疼痛事件。

2. 模型与方法

(a) 模型与假设

我们的方法受到这样一类数据集集合的启发：它们的采样不规则且长度有限，但集合中至少有一部分可以用单一模型很好地解释。我们以镰状细胞病（SCD）患者疼痛事件组成的集合为例，正是如此。

特别是，镰状细胞病患者早期再入院的频率很高（近 90% 的患者会在 30 天内再次入院[19, 20]），这表明疼痛事件可能会自然地在时间上聚集。因此，我们将疼痛危象事件的发生视为一种自激过程，也就是说，患者在一次事件发生后不久，发生后续事件的风险会升高。这类过程最初由 Alan G. Hawkes 在地震活动建模的背景下以数学方式加以描述 [21–23],地震会引发余震。这一类比自然也适用于镰状细胞病：血流最初受阻会引起炎症反应，进而可能诱发后续疼痛危象。在犯罪活动 [24] 和金融市场 [25]中也曾报告过类似的动力学。因此，我们将疼痛危象事件的发生视为霍克斯过程。

霍克斯过程是一种计数过程 $\{N(t)|t\geq 0\}$，它具有相关历史记录 $\mathcal{H}(t):=\{t_i|t_i<t\}$，其中 t_i 表示事件的发生时间，条件强度函数 $\lambda(t|\mathcal{H}(t))$ 的形式为

$$\lambda(t|\mathcal{H}(t))=\lambda_0(t)+\sum_{i:t>t_i}\Phi(t-t_i), \tag{2.1}$$

其中，$\lambda_0>0$ 是基线强度，而 $\Phi\geq 0$ 是一个单调递减的函数，称为记忆核。在此类过程中，事件发生后的一段时间内，初始

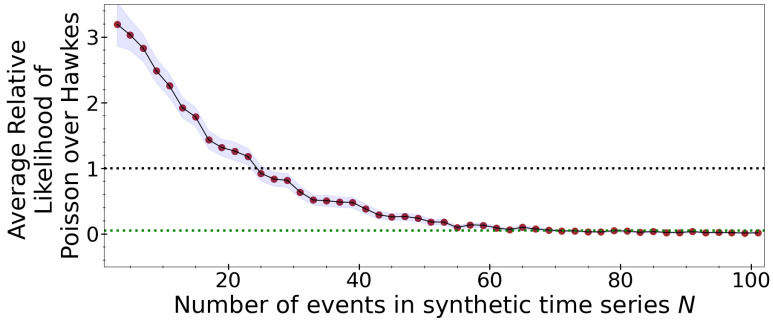


图 2. 可靠进行霍克斯模型与泊松模型判别所需的最小数据集大小。 区分霍克斯模型与泊松模型所需的数据点数量。黑色虚线表示基本偏好（$\mathcal{L}=1$），绿色虚线表示95%置信度（$\mathcal{L}=0.05$）。对于每个 N，我们计算50次相对似然。随后计算平均值，并以红色标记表示。紫色阴影区域表示95%置信区间。似然值和临界 N 取决于参数选择。我们采用（λ_0，α，δ）=（1、3、6）。

到达会增加后续事件发生的可能性（这一“记忆时间尺度”由 Φ的衰减率决定）。此后，“条件强度函数”和“强度函数”将交替使用，且 $\lambda(t) \equiv \lambda(t|\mathcal{H}(t))$.

我们进一步简化霍克斯过程模型，假设基线强度恒定为 $\lambda_0(t) = \lambda_0$，并采用指数记忆核 $\Phi(t-t_i)$ =exp[$-\delta$（$t-t_i$）]，其中 $\delta>0$。我们出于三个原因选择指数核：（a）它具有解析可处理性，因此可以进行闭式似然计算；（b）它具有明确的参数可解释性，其中 δ^{-1} 表示特征记忆时间尺度；（c）它在生理学上具有合理性，因为炎症级联反应很可能呈现指数衰减动力学。

使用霍克斯模型时，一个常见的假设是从过程开始时就对其进行观测。然而，实际收集数据时，可能是在一个更长过程中的任意时间段内进行的——我们的SCD数据集就是如此，它们始于SCD症状出现数月至数年之后。考虑到这一点，我们引入了一个额外的补偿项，以捕捉数据集中未记录的过去事件，$(\gamma-\lambda_0)e^{-\delta t}$。由此得到我们提出的模型强度函数

$$\lambda(t) = \lambda_0 + \alpha \sum_{i:t>t_i} e^{-\delta(t-t_i)} + (\gamma-\lambda_0)e^{-\delta t}, \tag{2.2}$$

其中 $\lambda_0>0$，$\alpha\geq 0$，$\delta>0$，以及 $\gamma\geq 0$ 均为常数，而 t_i 是 i^{th} 事件发生的时间。各参数的说明见表 1，参数示意图见图 1。

需要指出的是，该模型适用于在 $t=0$记录事件的时间序列数据集；该时间点作为初始条件 γ的参考点。决定模型行为的关键参数是 λ_0、α 和 δ，而 γ应视为校正因子，用于补偿数据集中未捕捉到的过去历史所产生的影响。

（b）模型拟合与选择

在对数据进行拟合之前，我们希望明确什么样的模型才可称为“成功”的模型。具体而言，我们希望选择一个合理的零模型作为检验基准，并希望准确推断出区分不同模型所需的数据量。

将 $\alpha=0$ 和 $\gamma=\lambda_0$ 代入公式（2.2）可得到齐次泊松过程（以下简称“泊松过程”），但这只是记忆效应不存在时霍克斯模型的一种特殊情况。[1] 反过来，霍克斯模型可以看作是在基线泊松模型上扩展出具有自激记忆核的模型。因此，泊松过程是零模型的自然候选。除了它代表了一种简单且广泛用于对点过程建模的框架之外，我们还可以评估，霍克斯过程增加的复杂性是否带来了统计学显著改进。

我们使用最大似然估计来拟合模型参数，其中数据集 $\{t_i\}_{i=1}^N$ 的似然函数为

$$L(\theta|t_1,t_2,...,t_N) = \left(\prod_{i=1}^{N}\lambda(t_i)\right) e^{-\int_0^{t_N}\lambda(s)ds}. \tag{2.3}$$

[1]当 $\delta\to\infty$ 时，也会出现相同的渐近行为，这意味着每次事件发生后，记忆都会立即消失。

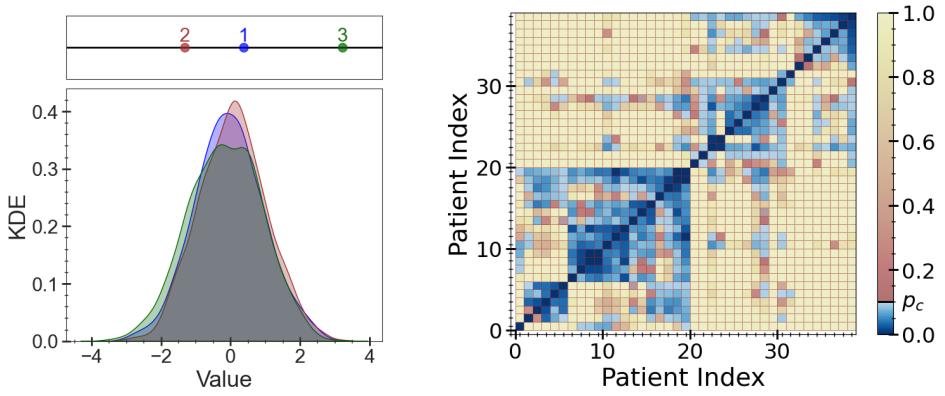


图3. KS检验展示了非传递相似性，并识别出可匹配的患者对。（左）对三个均值和方差略有不同的正态分布（1、2 和 3，分别以蓝色、红色和绿色表示）进行双样本KS检验。（1 对 2）、（1 对 3）和（2 对 3）的三个 p-值分别为 0.03、0.02 和 10^{-5}。此示例展示了该检验的非传递性：上方面板中各点之间的相互距离与KS统计量成正比；1 与 2 和 3 都相似，而 2 与 3 相距较远，因此彼此不相似。（右）使用双样本KS检验（显著性水平为 0.1）在我们的数据集集合中寻找相似患者。蓝色色调表示匹配的患者对，黄粉色色调表示未匹配的患者对。

对于霍克斯模型，如果我们知道完整的历史记录（如在模拟中那样），该过程从基线强度 $\gamma = \lambda_0$ 开始，其定义如下：

$$\lambda(t) = \lambda_0 + \alpha \sum_{i:t>t_i} e^{-\delta(t-t_i)}. \tag{2.4}$$

我们使用赤池信息准则（赤池信息准则）进行比较，因为它在模型的拟合优度与复杂度之间提供了有理论依据的平衡 [26]（不过，我们注意到，在动力系统中使用信息准则可能存在问题 [27]）。我们的候选模型有四个自由度（λ_0，α，δ，γ），而零模型只有一个（λ_P）。由于这两个模型是嵌套的，且泊松模型是霍克斯模型的简化版本，复杂度上的差异可能使我们难以找回真实模型，尤其是在数据有限时。

为了说明这一点，我们使用方程（2.4）模拟多个不同长度的时间序列。图2 显示，对于较短的时间序列，赤池信息准则通常更倾向于泊松模型，这表明存在一个阈值长度（取决于参数选择），低于该长度时，这种方法无法检测出真实的依赖结构。这与以下原则一致：样本量较大时，赤池信息准则通常表现更佳，因为此时拟合与复杂度之间的权衡会更加清晰 [28]。[2]

（c）数据增强

我们增强稀疏数据集的思路如下：（1）对数据集进行两两配对，检验其统计相似性；（2）将每个单独的数据集替换为由已证实相似的数据集组成的集成；（3）将模型参数拟合到每个完整集成上，并将其视为由从单一过程抽取的、不相连的片段构成。

具体而言，我们使用双样本 Kolmogorov-Smirnov（KS）检验，识别分布相似的数据集。KS检验是一种非参数方法，用于比较两个独立样本的累积分布，并量化它们之间的距离。这使其成为识别集合内部相似性的有效工具，因为它不依赖于对底层分布的假设，并且对位置和形状上的差异都很敏感。

通过应用KS检验，我们可以系统地将特征相近的数据集分组，从而进行集体分析。该检验的零假设是两个样本来自同一分布。检验统计量和样本量用于计算检验的p值；如果该p值超过我们设定的阈值 p_c，则接受该假设。其应用示例如图3 所示。

需要指出的是，由于 KS检验比较的是从分布中抽取的样本，因此无法直接应用于完整的时间序列。为解决这一问题并适用于我们的患者数据，我们转而比较事件之间到达间隔时间的分布，从而捕捉其潜在的时间结构。

[2]图 2 使用修正赤池信息准则（AICc） [29] 的版本见附录C。有限数据校正并不能解决稀疏数据问题，反而会使其更具挑战性。

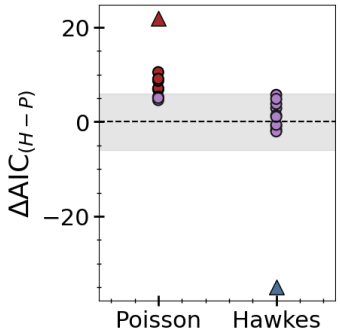


图4. 数据增强使模型选择的倾向从难以确定转向明确。我们生成了 20 条合成时间序列，其中 10 条来自泊松过程，10 条来自霍克斯过程，然后考察增强前后每个模型的相对统计支持 $\Delta\text{AIC}^{(H-P)} = \text{AIC}(H) - \text{AIC}(P)$。每个单序列实现以彩色圆圈表示，增强数据集则以较大的三角形表示，并采用相同的配色方案。蓝色标记表示倾向于霍克斯模型，红色标记表示倾向于泊松模型，紫色标记表示无法确定的情况。虚线以上的点表示倾向于泊松模型，虚线以下的点则表示倾向于霍克斯模型。此处 $N_{\text{aug}} = 10$ 和 $N_{\text{e}} = 30$（每条时间序列的事件数）。灰色带以外的区域表示对模型的支持置信度为 $> 95\%$。参数：$(\lambda_0 = 1,\ \alpha = 2,\ \delta = 3.5)$ 和 $\lambda_P = 7/3$。

确定相似的数据集后，我们定义一个“集体似然”，整合这些匹配组中的信息，从而提高模型选择的可靠性，尤其适用于单个数据集过于稀疏、无法进行稳健分析的情形。以下分步骤说明这一策略，并将其应用于由少量或稀疏时间序列构成的集合；在这种情况下，模型选择原本可能并不可靠：

步骤 1：考虑由 m 时间序列数据集组成的集合 $\mathcal{C} = \{t_i^j : i = 1 \dots n_j, j = 1 \dots m\}$。计算到达间隔时间集合 $\Delta\mathcal{C} = \{\Delta t_i^j : i = 1 \dots n_j - 1, j = 1 \dots m\}$。
步骤2：对每一对到达间隔时间数据集使用双样本KS检验计算 p 值，并定义矩阵 $\mathbf{P} = \{p(i,j) : i,j = 1 \dots m\}$。将 p 值低于给定阈值 p_c 的数据集视为相似。˜步骤 3：我们将模型对数据集 j 的集体似然（L）定义为

$$\tilde{L}_j = L_j \prod_{i:\mathbf{P}_{ij} < p_c, i \neq j} L_i, \tag{2.5}$$

其中，L_i 是给定模型对数据集 i 的个体似然，而 p_c 是相似性阈值。这样，我们就将相似的数据集视为同一过程的不同实现。
步骤 4：我们使用集体似然计算每个模型的最佳拟合参数和赤池信息准则值，并以此进行模型选择。

请注意，这种聚合方法不具传递性。如果数据集 A 与 B 在统计上相似，且 B 与 C 也相似，并不意味着 A 与 C 也相似。因此，最终拟合到 A、B 和 C 的模型参数可能各不相同。

3. 模型分析与方法验证

（a）区分霍克斯过程与泊松过程

我们考虑两个过程：一个参数为 $\lambda_{\text{P}} = 2$ 的泊松过程，以及一个具有可比平稳行为的霍克斯过程（式（2.4））。具体而言，我们选择参数，使其平衡强度的期望值 $\mathbb{E}[\lambda(t)] = \lambda_0(1 - \alpha/\delta)^{-1}$ 也为 2（该公式的简要推导见附录 A ，参数可辨识性问题的简要讨论见附录 B ）。

本次数值实验旨在检验所提出的数据增强策略是否通过提升区分这两个过程的能力来改善模型选择。具体而言，我们评估泊松模型和霍克斯模型分别拟合到 $N_{\text{aug}} = 10$ 个原始单序列数据集（每个数据集包含 $N_{\text{e}} = 30$ 个事件）以及每类模型的单个增强数据集后所得的赤池信息准则差异。

结果汇总于图4中，其中 Δ展示了单条序列（彩色圆点）和增强数据集（较大的三角形）的赤池信息准则值。蓝色标记表示更倾向于霍克斯模型，红色标记表示更倾向于

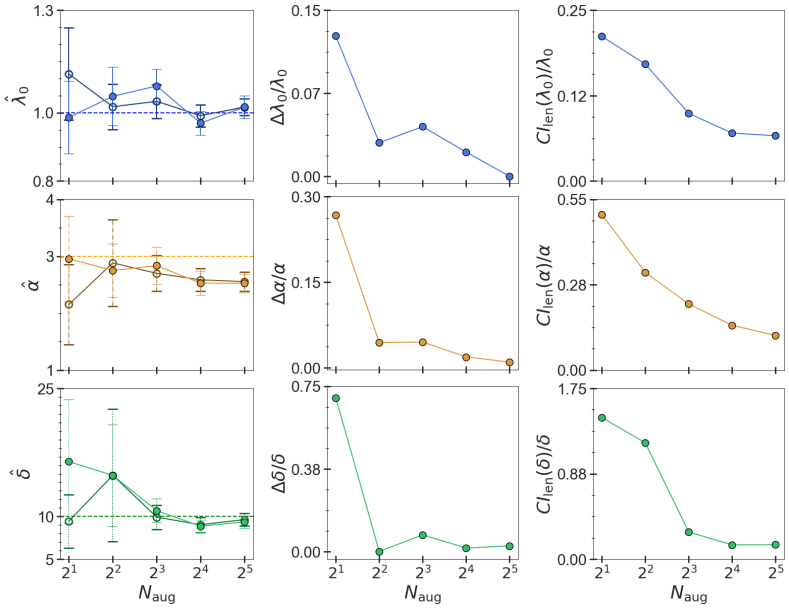


图 5. 增强后的稀疏序列能够恢复参数，效果与等长度的连续数据相当。比较我们的数据增强策略（实心标记）与总长度相同的单条连续时间序列（空心标记）的参数估计性能 $N_{\mathrm{e}}N_{\mathrm{aug}}$。所有面板均绘制结果随 N_{aug} （增强后序列的数量）的变化，其中每条增强后序列包含 N_{e} 个事件。顶行： λ_0；中行： α；底行： δ。左列：估计的参数值，水平虚线表示真实值（$\lambda_0 = 1$、 $\alpha = 3$、 $\delta = 10$），标记表示 30 次试验的均值，误差线表示标准差。中列：两种方法之间的相对误差，定义为 $|\hat{\theta}_{\mathrm{full}} - \hat{\theta}_{\mathrm{aug}}|/\theta_{\mathrm{true}}$，其中 $\hat{\theta}_{\mathrm{full}}$ 和 $\hat{\theta}_{\mathrm{aug}}$ 分别是完整序列和增强策略得到的参数估计值。右列：95% 置信区间的长度，按参数值归一化，以便比较不同尺度。

泊松模型，紫色标记则表示结果不确定。图中的灰色阴影表示不确定区域，即在95%置信区间内两个模型均未明显占优的结果。该不确定区域的边界可根据 Δ赤池信息准则与相对似然$\mathcal{L}_{\mathrm{relative}} = \exp(-\Delta\text{赤池信息准则}/2)$之间的关系推导得出。在95%置信水平下，可得 $\Delta\text{赤池信息准则}^{(\mathrm{crit})} = -2\ln(0.05) \approx 6$（因此，任何位于不确定区域之外的 $|\Delta\text{赤池信息准则}| \gtrsim 6$ 都表明有强有力的证据支持其中一个模型而非另一个模型）。

增强数据集始终会使两种过程的结果移出不确定区域，这表明该策略确实能够实现更稳健的模型选择。

(b) 增强数据集的参数估计

我们希望检验所提出的数据增强策略对参数估计的影响。为此，我们生成一个时间序列集合，每条序列由从霍克斯过程中任意时间区间内抽取的 N_{e} 个事件组成[3]。我们将通过增强 N_{aug} 条时间序列得到的最佳拟合估计量，与从总计包含 $N_{\mathrm{aug}}N_{\mathrm{e}}$ 个事件的单条较长时间序列中得出的最佳拟合参数进行比较。图 5展示了在每条序列的事件数固定为 $N_{\mathrm{e}} = 16$ 时，增强次数变化所产生的影响。

如图所示，数据增强方法能够恢复出与等长不间断时间序列所得结果相当的参数。也就是说，我们的方法可以得出稳健的参数估计值，有效弥补数据稀疏性。

4. 真实世界数据的结果

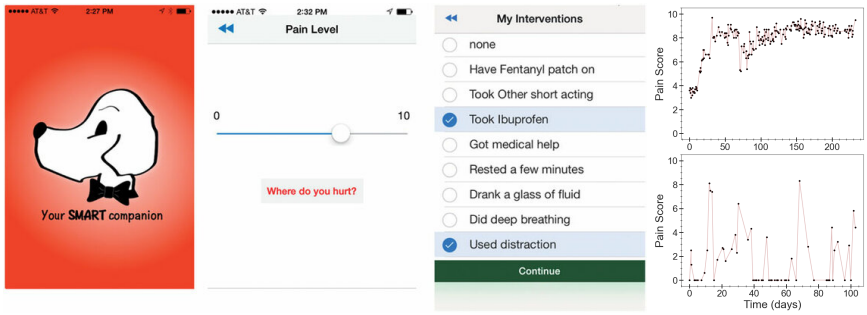


图6.真实患者数据。SMART 应用程序的示例截图，以及使用该应用程序收集的典型患者时间序列数据 [30]。请注意患者间变异性、时间不规则性、报告疲劳以及其他数据质量挑战。

(a) 数据

镰状细胞病（SCD）是一种终身性遗传性疾病，会影响血红蛋白。《血红蛋白》是红细胞（RBCs）中负责携带氧气的成分。患有 SCD 时，红细胞会变形成“镰刀”状，阻碍正常血流，并可能引发危及生命的问题。SCD 影响着美国超过 10 万人，全球则有 800 万人 [31, 32]。SCD 患者约 90% 的急性护理就诊与严重且频繁的疼痛发作有关。了解并建模这些疼痛发作，对于改善患者护理和治疗策略至关重要，因为它们会显著影响生活质量和医疗成本 [33]。

我们的研究采用了通过自研的镰状细胞移动应用程序“利用技术记录症状”（SMART 应用程序）收集的数据 [30, 34, 35]。图6 展示了通过该应用程序收集的两组患者报告的主观疼痛数据示例。这些数据来自一个由39名患者组成的小型队列，他们每天都需要报告自己的疼痛程度。尽管主观疼痛报告可能无法完全对应生理指标，但由于疼痛本质上具有主观性，且自我报告目前是评估疼痛严重程度的唯一经验证的方法（并且已被证明是镰状细胞病临床结局的可靠指标），因此它们仍是镰状细胞病疼痛管理的核心 [36–39]。该数据集包括疼痛程度及其对应的时间戳等内容。

在我们的分析中，我们将事件时间视为仅对应非零疼痛程度的时间戳，其中每个报告的高于零的疼痛程度都构成一个事件。我们假设两次报告之间的疼痛实际上为零，并且疼痛事件的发生具有时间依赖关系的随机行为。尽管我们假设某一天未报告等同于未发生事件，但我们承认，考虑到可能存在的报告疲劳或漏填，这一假设可能会引入一些误差。

（b）模型拟合与区分

在图 7中，我们展示了使用数据增强方法，对来自镰状细胞病患者队列的现实世界数据集进行霍克斯模型与泊松零模型比较的结果。我们通过 Δ赤池信息准则（AIC）展示模型比较，其中圆圈表示未进行数据增强的结果，三角形表示进行数据增强后的结果。颜色表示置信度：对于 **红色**的不确定区域， ΔAIC 介于-6 和 6 之间；对于置信度超过 95% 的霍克斯拟合，**蓝色**表示（ΔAIC < -6）。[4]

单序列拟合中，偏好泊松零模型而非霍克斯模型的情况占多数（39 例中有 28 例）；而增强拟合显示，绝大多数情况下更偏好霍克斯模型（39 例中有 36 例）——见表2。值得注意的是，即使在许多未增强数据集的 ΔAIC ≈ 6（这是支持泊松零模型的最有力证据）情况下，增强数据集的霍克斯模型选择仍然出现。

这些结果有力地证明，利用基于患者相似性的集体似然，有助于提升模型选择和参数估计的效果。霍克斯模型偏好度的提高，与时间依赖关系在数据中发挥关键作用这一假设高度吻合。

[3]这模拟了现实世界数据集中常见的情况：完整的历史记录往往不可得，只能观测到有限时间区间内的片段。[4]请注意，最大的 Δ赤池信息准则为 $+6$，因为通过适当选择参数，霍克斯模型可以简化为泊松模型，且两者的自由度相差三。

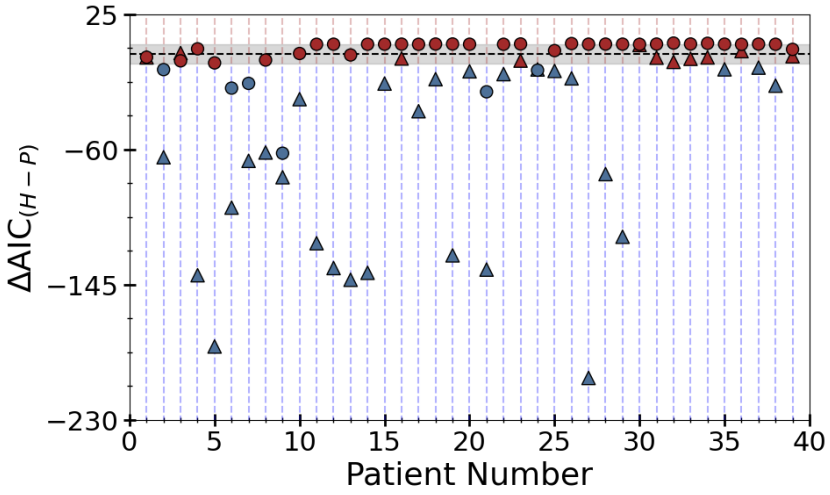


图 7. 应用于患者数据的增强方法。我们绘制了增强前（圆圈）和增强后（三角形）每个时间序列的赤池信息准则差值（霍克斯减去泊松）。灰色带表示不确定区域（红色标记）；灰色带下方的区域表示以至少 95% 的置信度倾向于霍克斯模型（蓝色标记）。有关增强中使用的基于 KS 的相似性矩阵，请参见图 3右侧面板。

序列类型	置信水平	对应的 $\lvert\Delta$赤池信息准则$\rvert$	泊松患者数	霍克斯患者数
单条	基本偏好	0	28	11
	0.05	6	2	6
	0.01	9.2	0	6
增强后	基本偏好	0	3	37
	0.05	0	0	29
	0.01	9.2	0	27

表 2。基于 AIC 的模型偏好比较：单条序列拟合与增强后拟合。“基本偏好”表示 AIC 较低的模型。0.05 的置信水平表示强烈偏好某一模型（$|\Delta$赤池信息准则$| \gtrsim 6$），而 0.01 的置信水平表示非常强烈偏好某一模型（$|\Delta$赤池信息准则$| \gtrsim 9.2$）。

5. 讨论

我们提出了一种数据增强策略，利用稀疏时间序列之间的统计相似性来改善模型选择和参数估计——在许多科学领域中，单个观测有限但预期存在时间依赖关系，因此这一直是一项挑战。

方法学贡献。 增强方法解决了分析稀疏事件数据时的两个关键挑战。首先，当单个时间序列缺乏足够事件、无法做出有结论的模型选择时，该方法能够实现模型区分。其次，通过将统计上相似的数据集视为同一底层过程的多次实现，该方法能够提供稳健的参数估计值。集体似然框架（式 2.5）具有通用性，不仅适用于点过程，也适用于任何特定领域之外的情形。

我们的增强方法在保留时间结构的同时，通过增加有效样本量来提高精度，且无需对依赖关系的具体形式作出假设。

关键局限。 要使增强方法适用，**必须合理预期多个数据集来自相同或非常相似的模型**。这是其适用性的一个关键且严格的限制。例如，如果时间序列来自不同实验，其模型参数存在显著差异（或其控制方程完全不同），该方法将会失效；若仍采用该方法，则可能为一个或多个模型提供错误且具有误导性的统计支持。

对于镰状细胞病，领域专家通常会将患者划分为数量有限的类别 [40, 41], ，这也为一种便于建模的假设提供了依据：患者可以根据动力学进行聚类。

其他局限性。基于 KS 的相似性评估可能无法捕捉到所有时间结构。非传递性（图 3）使不同个体的增强数据集可能存在差异，但也引入了潜在的选择偏差。结合领域特定协变量进行更复杂的聚类，或许能够改进分组 [16, 17]。

若分枝比相同，不同的霍克斯过程参数可能产生相似的到达间隔分布（附录 B）。外部协变量有助于消除这种退化性。指数核假设只有一个记忆时间尺度，而许多过程涉及多个尺度。恒定基线强度忽略了周期模式或趋势。采用时变参数、复合核或协变量的扩展方法可以在保留核心策略的同时解决这些问题 [42, 43]。

相似性检验的计算成本为 $O(m^2)$，优化的计算成本为 $O(mk)$，其中 m是单元数量， k 是平均组大小。对于大型数据集，近似方法或层次聚类可以提高可扩展性。

与现有方法的比较。 针对稀疏数据的传统方法包括自助法 [18]，（生成合成观测值）和贝叶斯方法 [16]，（纳入先验信息）。我们的方法在根本上有所不同：我们不使用合成数据或基于先验的数据扩充单个数据集，而是汇集统计上相似单元的真实观测值。这样既保留了推断的经验性质，又增加了有效样本量。

层次模型 [17, 44] 也会在相关群组间汇聚信息，但需要明确的嵌套结构和共享参数假设。我们基于相似性的方法更加灵活，可根据经验分布特性进行非层次分组。相似性的非传递性（图 3）意味着每个单元都可以与集合中不同的子集进行增强，从而实现标准层次框架无法做到的异质性汇聚。

具体而言，对于点过程，大多数增强策略都侧重于空间汇聚，或假设各单元之间具有同质性 [42]。我们通过到达间隔分布进行时间相似性评估，为识别可汇聚的单元提供了原则性标准，无需空间结构或同质性假设。

在镰状细胞病疼痛动态中的应用。我们将该方法应用于镰状细胞病疼痛事件，展示了其在真实世界临床情境中的实用价值。支持自激过程模型的患者比例从15%上升至74%，这对管理具有重要意义。例如，这表明在急性发作后的高风险时期加强监测或可改善治疗；监测时长由记忆时间尺度决定（δ^{-1}）（我们发现，该时间尺度在数据中为30秒至6分钟，这表明风险应在约30分钟内回归基线）。

现行指南强调反应性治疗 [37, 38]，，但时间依赖关系表明，在易感时期预防初始事件或打破激发循环，可能更为有效。分枝比 α/δ 量化了自激强度——分枝比更高的患者可能受益于更积极的早期干预，以防止级联，从而根据个体的时间动态制定个体化方案 [8, 36]。

未来方向。 接下来的重点工作包括在独立镰状细胞病队列中进行验证，以评估可推广性；并将其应用于真实值已知的其他时间事件数据集（例如，已知自激参数的模拟流行病数据）。对于镰状细胞病，将临床协变量（血红蛋白水平、基因型、治疗）纳入相似性评估，可能有助于在仅依据到达间隔时间的基础上进一步改善患者分组。

扩展到标记点过程（例如，参见 [45] 或 [46]）可以纳入事件严重程度，从而解决 SCD 应用中的一个关键局限：疼痛强度存在差异。时变 Hawkes 模型 [43] 与我们的策略相结合，可以捕捉急性和慢性疼痛状态 [47, 48]之间的转变。每项扩展都保留了核心原则：利用相似性克服个体数据的稀疏性。

在本文中，我们为每次数值实验手动选择了合适的数据集增强阈值 p_c 。尽管自动确定合理阈值显然值得研究，但我们将这一挑战留待未来工作；它与加权网络上的聚类／社群检测问题有关（相当于我们的 p 值矩阵 $\mathbf{P}$）。

最后，我们介绍了在两个点过程模型中进行选择的方法，但我们认为，该方法也可扩展到两个以上候选模型之间的选择，以及连续时间机理模型（例如动力系统）；不过，在这些情况下使用赤池信息准则时可能需要谨慎 [27]。

6. 结论

本研究提出一种用于时间事件建模的数据增强策略，以应对个体时间序列稀疏所带来的挑战。通过利用集体似然整合统计特征相似的数据集，该方法能够在单个单位中的事件不足以支持确凿推断时，实现可靠的模型选择和稳健的参数估计。

我们以一组39名镰状细胞病患者的疼痛事件数据为例，展示了该方法的应用。该方法的适用性最终取决于一项关键假设：多个稀疏数据集源自同一个（或几乎相同的）模型。在这一假设合理的情况下，我们预计该框架能够基于碎片化数据实现可靠推断，提升我们理解和预测复杂系统动力学的能力。

数据可获取性。 结果部分所用的数据已随稿件一并共享。

作者贡献。 K.U. 和 D.M.A.：概念构思、形式分析、方法设计、可视化、原稿撰写、审阅与编辑。N.R.S.：临床指导、数据。T.B：数学与计算方面的指导。所有作者均对稿件发表作出最终批准，并同意对其中开展的工作负责。

利益冲突。 K.U. 和 D.M.A. 声明，本研究不存在利益冲突。

经费。 本研究由美国国立卫生研究院资助，项目编号为 5R01AT010413。

致谢。作者谨向 Gary K. Nave Jr. 博士（科罗拉多矿业学院）和 Richard Suhendra（西北大学）表示感谢，感谢他们提供的宝贵讨论意见。

参考资料

[1]E. Pyper、S. McKeown、J. Hartmann-Boyce 和 J. Powell，《使用患者生成数据进行真实世界临床结局测量的数字健康技术：系统范围综述》，《医学互联网研究杂志》25，e46992（2023）。[2]W. B. Nowell 等，《参与者参与及提供智能手表和患者报告结局数据的依从性：类风湿关节炎纵向数字追踪（DIGITAL）真实世界研究》，《JMIR 人因》10，e44034（2023）。[3]H. S. Kang 和 M. Exworthy，《穿戴未来——可穿戴设备助力用户对自身健康与护理承担更大责任：范围综述》，《JMIR 移动健康与泛在健康》10，e35684（2022）。[4]M. Wettstein 等（作者），《将基于消费级可穿戴设备的远程患者监测系统整合到临床工作流程中：可用性与可行性研究》，《JMIR 移动健康与泛在健康》12，e58441（2024）。[5]M.-P. Gagnon 等，《支持慢性病自我管理的可穿戴设备：范围综述》，《交互式医学研究杂志》13，e55925（2024）。[6]H. Zheng 等，《可穿戴技术在全关节置换术后测量和改善患者结局中的作用：文献综述》，《JMIR 围手术期医学》6，e39396（2023）。[7]L. Hassan、A. Milton、C. Sawyer、A. J. Casson、J. Torous、A. Davies、B. Ruiz-Yu 和 J. Firth，《消费级可穿戴设备用于推断严重精神疾病患者身心健康结局的效用：系统综述》，《JMIR 精神健康》12，e65143（2025）。[8]C. Vuong、K. Utkarsh、R. S. Stojancic、A. Subramaniam、O. Fernandez、T. Banerjee、D. M. Abrams、K. Fijnvandraat 和 N. R. Shah，《使用消费级可穿戴设备监测和预测镰状细胞病患者的疼痛》，《数字健康前沿》5，1285207（2023）。[9]R. S. Stojancic、A. Subramaniam、C. Vuong、K. Utkarsh、N. Golbasi、O. Fernandez 和 N. R. Shah，《在日间医院使用商用 Apple Watch 预测镰状细胞病患者的疼痛：可行性研究》，《JMIR 形成性研究》7，e45355（2023）。[10]X. Teng、S. Pei 和 Y.-R. Lin，《Stocast：考虑进展不确定性的随机疾病预测》，《IEEE 生物医学与健康信息学汇刊》25，850–861（2020）。[11]A. D. Kaplan、U. Tipnis、J. C. Beckham、N. A. Kimbrel、D. W. Oslin、B. H. McMahon 等，《纵向电子健康记录的连续时间概率模型》，《生物医学信息学杂志》130，104084（2022）。[12]J. Liu 等，《用于生物医学传感器数据的贝叶斯结构时间序列：评估干预措施的灵活建模框架》，《PLOS 计算生物学》17，e1009303（2021）。[13]M. A. Babyak，《所见未必所得：回归类模型中过拟合的简明非技术性介绍》，《身心医学》66，411–421（2004）。[14]J. P. A. Ioannidis，《为什么大多数已发表的研究发现都是错误的》，《PLoS 医学》2，e124（2005）。[15]R. D. Riley、K. I. E. Snell、J. Ensor、D. L. Burke、F. E. Harrell Jr、K. G. M. Moons 和 G. S. Collins，《开发多变量预测模型的最小样本量：第二部分——二分类和事件发生时间结局》，《医学统计学》38，1276–1296（2019）。[16]A. Gelman、J. B. Carlin、H. S. Stern、D. B. Dunson、A. Vehtari 和 D. B. Rubin，《贝叶斯数据分析》，第三版（Chapman & Hall/CRC，2013）。[17]S. W. Raudenbush 和 A. S. Bryk，《分层线性模型：应用与数据分析方法》（2002）。[18]B. Efron 和 R. J. Tibshirani，《自助法简介》（Chapman & Hall/CRC，1994）。

[19]O. Adesina、A. Brunson、M. J. Fisch、T. Wun 和 Q. Shi，《镰状细胞病患者全因30天再入院率及风险因素：基于人群的队列研究》，《美国血液学杂志》98，730–738（2023）。[20]N. Shah、M. Bhor、L. Xie、J. Paulose 和 H. Yuce，《镰状细胞病并发症：患病率与资源利用》，PLoS ONE 14，e0214355（2019）。[21]A. G. Hawkes，《若干自激和互激点过程的谱》，《生物统计学》58，83–90（1971）。[22]A. G. Hawkes 和 D. Oakes，《自激过程的簇过程表示》，《应用概率杂志》11，493–503（1974）。[23]Y. Ogata，《地震发生的统计模型及点过程的残差分析》，《美国统计学会杂志》83，9–27（1988）。[24]G. O. Mohler、M. B. Short、P. J. Brantingham、F. P. Schoenberg 和 G. E. Tita，《犯罪的自激点过程建模》，《美国统计学会杂志》106，100–108（2011）。[25]C. G. Bowsher，《连续时间内证券市场事件的建模：基于强度的多元点过程模型》，《计量经济学杂志》141，876–912（2007）。[26]H. Akaike，《对统计模型识别的新思考》，《IEEE自动控制汇刊》19，716–723（1974）。[27]K. Utkarsh 和 D. M. Abrams，《信息准则不适用于动力系统：采样率与维数依赖性》，arXiv，2511.14931（2025）。[28]C. M. Hurvich 和 C.-L. Tsai，《小样本中的回归和时间序列模型选择》，《生物统计学》76，297–307（1989）。[29]C. M. Hurvich 和 C.-L. Tsai，《小样本中的回归和时间序列模型选择》，《生物统计学》76，297–307（1989）。[30]N. Shah、J. Jonassaint 和 L. De Castro，《患者欢迎使用镰状细胞病移动应用程序，通过技术记录症状（SMART）》，《血红蛋白》38，99–103（2014）。[31]F. B. Piel、M. H. Steinberg 和 D. C. Rees，《新生儿镰状血红蛋白的全球流行病学：当代基于地统计模型的地图与人口估计》，《柳叶刀》381，142–151（2013）。[32]K. L. Hassell，《美国镰状细胞病人口估计》，《美国预防医学杂志》38，S512–S521（2010）。[33]O. S. Platt、B. D. Thorington、D. J. Brambilla、P. F. Milner、W. F. Rosse、E. Vichinsky 和 T. R. Kinney，《镰状细胞病疼痛：发生率与风险因素》，《新英格兰医学杂志》325，11–16（1991）。[34]C. R. Jonassaint、N. Shah、J. Jonassaint 和 L. De Castro，《患者欢迎使用镰状细胞病移动应用程序，通过技术记录症状（SMART）》，《血红蛋白》38，99–103（2014）。[35]N. Shah、J. Jonassaint 和 L. De Castro，《针对镰状细胞病急性疼痛管理的数字健康干预：试点研究》，《JMIR 移动健康与泛在健康》7，e11791（2019）。[36]W. R. Smith、L. T. Penberthy、V. E. Bovbjerg、D. K. McClish、J. D. Roberts、B. Dahman、I. P. Aisiku、J. L. Levenson 和 S. D. Roseff，《镰状细胞病成人患者的每日疼痛评估》，《内科学年鉴》148，94–101（2008）。[37]美国国家科学院、工程院和医学院，《应对镰状细胞病：战略计划与行动蓝图》（美国国家科学院出版社，华盛顿特区，2020）。[38]美国国家科学院、工程院和医学院，《社会保障残疾评估中的镰状细胞病：疼痛与治疗环境》（美国国家科学院出版社，华盛顿特区，2025）。[39]K. A. Stewart、M. Parshad-Asnani、A. Wonkam、J. Bollinger、V. N. Bitoungui、E. Wonkam-Tingang、J. Powell、K. Desronvil、K. R. Benson、A. Clark 等，《“疼痛是主观的”：关于三个国家医疗服务提供者对镰状细胞病疼痛管理的态度与实践的混合方法研究》，《疼痛与症状管理杂志》61，474–487（2021）。[40]N. Shah 等，《镰状细胞病严重程度分类系统的开发》，《临床经济学与结局研究》12，625–633（2020）。[41]S. K. Ballas 等，《镰状细胞病表型表现的定义》，《美国血液学杂志》85，6–13（2010）。[42]A. Reinhart，《自激时空点过程及其应用综述》，《统计科学》33，299–318（2018）。[43]W.-H. Chiang、X. Liu 和 G. Mohler，《结合流动性领先指标和空间协变量的霍克斯过程新冠肺炎建模》，《国际预测杂志》38，505–520（2022）。[44]H. Goldstein，《多层统计模型》，10.1002/9780470973394（2011）。[45]M. Schlather，《关于标记点过程的二阶特征》，《伯努利》7，99–117（2001）。[46]H. Lotwick 和 B. Silverman，《分析多种点空间过程的方法》，《英国皇家统计学会杂志B辑：统计方法学》44，406–413（1982）。[47]C. P. Carroll、S. Lanzkron、C. J. Haywood、K. Kiley、M. Pejsa、G. Moscou-Jackson、J. A. Haythornthwaite 和 C. M. Campbell，《检测镰状细胞病慢性疼痛的出现》，《疼痛与症状管理杂志》55，1131–1139（2018）。

[48]D. S. Darbari, J. P. Hampson, E. Ichesco, N. Kadom, G. Vezina, I. Evangelou, D. J. Clauw, J. G. Taylor 和 R. E. Harris，“镰状细胞病疼痛数据纵向评估的新型度量指标”，《临床疼痛杂志》**33**，517–524（2016）。[49]P. J. Laub、Y. Lee 和 T. Taimre，霍克斯过程要素（Springer，2021）。

A. 霍克斯过程的平稳平均强度

考虑一个基线强度为 λ_0 、核函数为指数核的霍克斯过程

$$\Phi(t) = \alpha e^{-\delta t}, \quad t \geq 0.$$

时刻 t 的条件强度为

$$\lambda(t) = \lambda_0 + \sum_{i:t_i<t} \Phi(t - t_i) = \lambda_0 + \alpha \sum_{i:t_i<t} e^{-\delta(t-t_i)},$$

其中，求和遍历所有在时刻 t 之前发生的事件。每个过去事件都会对当前强度产生贡献，且该贡献会随时间呈指数衰减。

分枝过程解释。 理解霍克斯过程的一种方式是考察其分枝结构。可以将每个事件视为 [49] 根据均值为某值的泊松过程，独立地产生未来的“子代”事件

$$\eta = \int_0^\infty \Phi(t)\, dt = \frac{\alpha}{\delta}.$$

这个量 η 称为分枝比。如果 $\eta < 1$，该过程就是次临界的，这意味着平均而言，每个事件产生的后代少于一个。这可以确保该过程不会爆发，并且存在平稳分布。

平稳平均强度。 令 $\lambda_c = \mathbb{E}[\lambda(t)]$ 表示平稳平均强度。根据平稳性，期望总强度可分解为基线强度与所有后代事件的期望贡献之和。换言之，每个事件对未来强度的平均贡献由分枝比决定：

$$\lambda_c = \lambda_0 + \eta\, \lambda_c.$$

这里，第一项 λ_0 表示自发事件，第二项 $\eta\, \lambda_c$ 则表示该过程的期望自激发效应。求解 λ_c 可得精确的平稳平均强度：

$$\lambda_c = \frac{\lambda_0}{1 - \eta} = \frac{\lambda_0}{1 - \alpha/\delta}. \tag{A 1}$$

分枝比 η 直观地衡量了过程的自激发强度：当 $\eta \to 1$时，平稳平均值发散，反映出过程向临界状态转变。

B. 指数Hawkes过程中 (α, δ) 的可辨识性

在图 5中，我们注意到，估计参数几乎收敛至真实值，但并未完全收敛；这是霍克斯模型参数可辨识性近乎失效的结果。如果参数可辨识性完全失效，则 (α, δ)平面上的一条或多条似然等高线将变成直线（见图 8）。

问题不在于该模型在理论上不可辨识，而在于有限样本下（尤其是观测窗口较短或事件较少时），似然面会沿着（α, δ）平面中的某些曲线变得近乎恒定，从而使不同的参数对产生几乎无法区分的拟合结果。

考虑具有指数记忆核的霍克斯过程

$$\Phi(t) = \alpha e^{-\delta t}, \qquad \alpha > 0,\ \delta > 0, \tag{A 1}$$

及其强度

$$\lambda(t) = \lambda_0 + \sum_{i:t_i<t} \alpha e^{-\delta(t-t_i)}. \tag{A 2}$$

对于$_{i=1}^{N}$在 $[0, T]$上的事件时间 $\{t_i\}$，对数似然为

$$\mathcal{L}(\alpha, \delta) = \underbrace{\sum_{i=1}^{N} \log \lambda(t_i)}_{\text{log–intensity at events}} - \underbrace{\int_0^T \lambda(t)\, dt}_{\text{compensator}}. \tag{A 3}$$

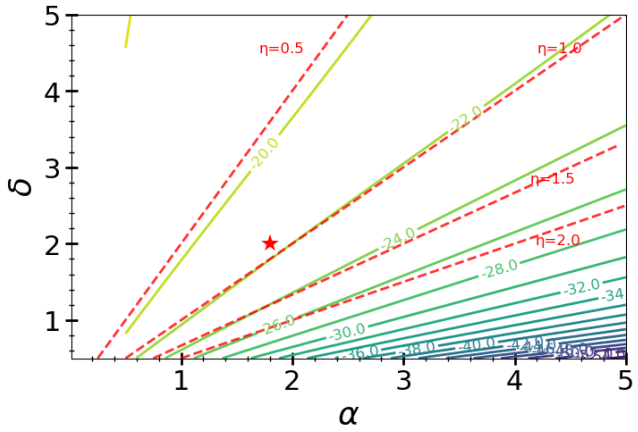


图 8. 沿恒定分枝比的似然脊线揭示了实际不可识别性。 指数Hawkes过程中 $(\alpha,\ \delta)$ 的实际不可识别性。对于包含 $N = 6$ 个事件、观测窗口为 $[0, 10]$，的合成数据，其对数似然曲面展示了等高线以及恒定分枝比 $\eta = \alpha/\delta$（红色虚线）的曲线。真实参数 ($\alpha = 1.8, \delta = 2.0$，红色星号）位于 $\eta = 0.9$ 似然脊线上。似然等高线几乎与恒定-η 线平行，从而形成一个平坦方向。数值计算表明，对于 $\eta = 0.9$，在 α 取值范围较宽的情况下，对数似然仅变化 $\Delta\log\ L \approx 0.8$，这说明尽管参数比值受到强约束，单个参数所受的约束却相对较弱。这种平坦性解释了为什么最大似然估计优化通常会得到接近但并不恰好等于真实值的参数估计值。

对于指数核，补偿子（积分强度函数）具有闭式形式

$$\int_0^T \lambda(t)\,dt = \lambda_0 T + \frac{\alpha}{\delta}\sum_{i=1}^{N}\left(1 - e^{-\delta(T-t_i)}\right). \tag{A 4}$$

当衰减率 δ 足够大，且事件在观测窗口中足够早地发生，使得$\delta(T - t_i) \gg 1$ 对大多数事件都成立时，这些项满足 $e^{-\delta(T-t_i)} \approx 0$。在此情形下，

$$\int_0^T \lambda(t)\,dt \approx \lambda_0 T + \frac{\alpha}{\delta}N, \tag{A 5}$$

因此，积分几乎完全取决于分枝比 $\eta = \alpha/\delta$。这会在似然平面中产生一条 仅由补偿子导致的精确线性等高线。

事件时刻的强度为

$$\lambda(t_i) = \lambda_0 + \sum_{j<i} \alpha\, e^{-\delta(t_i - t_j)}. \tag{A 6}$$

每个过去事件的贡献都涉及乘积 $\alpha e^{-\delta\Delta t}$，其中 $\Delta t = t_i - t_j$。原则上，这一项可以区分 α 和 δ，因为它不会简化为仅依赖 α/δ 的函数。然而，当以下任一情况发生时，激发贡献 $\sum_{j<i} \alpha e^{-\delta(t_i - t_j)}$ 相较于 λ_0 会变得可以忽略：(i) 激发较弱（$\alpha/\lambda_0 \ll 1$），使得该过程仅表现出微弱的自激性；或 (ii) 衰减率 δ 相对于典型的事件间隔时间较大，导致过去事件产生的激发在下一事件发生前便已消退。在这些情况下，对于大多数事件，$\lambda(t_i) \approx \lambda_0$，因此对数强度贡献 $\log\ \lambda(t_i)$ 随 (α, δ) 的变化很小。因此，事件时刻项无法显著改变补偿子所产生的近似线性等高线。

结合两种效应：

- 补偿子项沿着 α/δ恒定的曲线几乎保持不变，并且
- 事件时间强度只能对（α，δ）施加较弱的约束，除非数据包含能够探测指数衰减尺度的事件间隔，

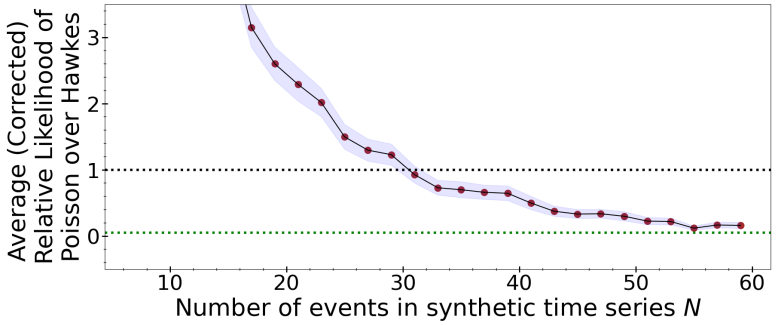


图 9. AICc 修正无法解决稀疏数据模型选择面临的挑战。使用 AICc 区分霍克斯模型与泊松模型所需的数据点数量。与图2一样，黑色虚线表示基本偏好（$\mathcal{L}=1$），绿色虚线表示 95% 置信度（$\mathcal{L}=0.05$），红色标记表示 50 次重复实验的平均值，紫色阴影表示 95%置信区间。参数取值为：（λ_0、α、δ）=（1, 3, 6）。

对数似然沿着接近以下曲线的方向形成一条近乎恒定的脊线

$$\{(\alpha,\delta):\alpha/\delta\approx\text{constant}\}. \tag{A 7}$$

这解释了为什么除非拥有大型数据集或采用极严格的优化容差，否则参数估计值通常会趋近于真实值，但并不完全相等。图8展示了这些脊线。

C. 模型选择的小样本修正

为评估小型数据集对模型选择的影响，我们使用修正后的赤池信息准则（AICc） [29]重新进行了模型比较。图9 展示了与图2类似的结果，说明 AICc 在定性表现上与 AIC 几乎相当。