統計数理(2021)
第 69 巻 第 2 号 165–180

特集「Hawkes 過程の新展開と応用」
[原著論文]

Hawkes型計数時系列モデル

小山 慎介†

(受付 2020 年 12 月 25 日；改訂 2021 年 4 月 17 日；採択 5 月 6 日)

要　　旨

本論文では，離散時間の計数時系列に対して Hawkes 過程に類似する時系列モデルを提案する．特に Hawkes 過程と分岐過程の対応に着目し，同様な性質を計数時系列モデルで構築する．このときに，計数時系列モデルの構成を Hawkes 過程とパラレルに展開し，両者を対比する．最後に，提案するモデルを感染症のモデリングに応用する．Wallinga と Teunis による実効再生産数の推定量を Hawkes 型計数時系列モデルから導出し，過分散の場合に一般化する．

キーワード：Hawkes 過程，分岐過程，計数時系列モデル，負の二項分布，実効再生産数．

1. はじめに

イベント(事象)の発生が引き金となって更なるイベントが引き起こされる，そのような現象は至るところで観察される．地震や感染症は典型的な例である．Alan G. Hawkes は，不規則に発生するイベントの続発性や相互誘発作用を記述する点過程を導入した (Hawkes, 1971a, 1971b)．

Hawkes 過程はベースラインの発生率に過去のイベントの影響を足し合わせた強度を持つ点過程である．線形性のため，Hawkes 過程は分岐過程からなるクラスターの重ね合わせで表現することができる (Hawkes and Oakes, 1974)．イベントの連鎖的な発生やクラスター性は様々な現象で共通にみられる特徴であり，地震 (Ogata, 1988)，神経活動 (Chornoboy et al., 1988; Pernice et al., 2011)，感染症 (Meyer et al., 2012; Chiang et al., 2020; Koyama et al., 2021)，ファイナンス (Bacry et al., 2015; Hawkes, 2018)，ソーシャル・ネットワーキング・サービス (Fox et al., 2016; Kobayashi and Lambiotte, 2016; Koyama and Shinomoto, 2020)，犯罪や紛争 (Mohler et al., 2011; Lewis et al., 2012; Zhuang and Mateu, 2019)，交通事故 (Kalair et al., 2021) など多くの分野のモデリングに Hawkes 過程は応用されている．

Hawkes 過程の普及が急速に進んでいる背景のひとつに，計測データの精密化を挙げることができる．個々のイベントが発生する場所や時刻が正確に記録され，発生のタイミングや相関に興味があるとき，連続時間上で定義された Hawkes 過程は適当なモデルである．

一方で，個々のイベント発生時刻は計測されず，イベント発生件数が区間毎に集計されたデータも多く存在する．例えば，本稿で取り上げる新型コロナウィルス感染症(COVID-19)では，1 日の新規陽性者数が毎日報告されている．感染もしくは発病した時点をイベントとすれば，感染症の伝播を Hawkes 過程でモデル化することはできる．しかし，実際には 1 日の発生件数だけがデータとして記録されるので，個々のイベントの発生時刻は観測されない潜在変数とみなされる．このようなデータに Hawkes 過程を当てはめるアプローチも考えられる

† 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

が (Cheysson and Lang, 2020)，ここでは Hawkes 過程をデータ生成モデルとして想定するのではなく，離散時間の計数時系列 “そのもの” をモデリングするアプローチをとる．

本論文では，Hawkes 過程に類似する離散時間の計数時系列モデルを提案する．計数時系列に対するよく知られたモデリング法として，動的一般化線形モデル (West and Harrison, 1997; Fahrmeir and Tutz, 2001)（あるいは一般状態空間モデル，Kitagawa, 2010）や整数値自己回帰モデル (Kirchner, 2016, 2017; 中嶋 他, 2017)が挙げられるが，これらとは異なるアプローチを展開する．ここでは特に Hawkes 過程と分岐過程の対応に着目し，計数時系列モデルで同様の性質を構成する．第 2 節では Hawkes 過程と分岐過程の対応関係をまとめ，第 3 節で Hawkes 型計数時系列モデルを構築する．第 4 節では Hawkes 型計数時系列モデルから実効再生産数の推定量を導き，COVID-19 のデータに応用する．第 5 節で他の方法との関連を議論する．

2. Hawkes 過程

時刻 $t \in \mathbb{R}$ までに生起したイベント数を $N(t)$ とし，対応するイベント発生時刻を t_i $(i = 1, 2, \ldots)$ とする．時刻 t までのイベントの発生履歴 $H_t = \{t_i | t_i < t\}$ が与えられた下で次の瞬間にイベントが発生する確率が，条件付き強度関数 $\lambda(t)$ を用いて

$$P\{N(t+\Delta t) - N(t) = 1 | H_t\} = \lambda(t)\Delta t + o(\Delta t)$$
$$P\{N(t+\Delta t) - N(t) > 1 | H_t\} = o(\Delta t)$$

で与えられるとする．Hawkes 過程は条件付き強度関数が

$$\lambda(t) = \mu + \int_0^t g(t-u)dN(u) \tag{2.1}$$

で与えられる点過程である (Hawkes, 1971b)．式(2.1)の右辺第 1 項の μ はベースラインの発生率を表し，第 2 項は過去のイベントの影響を表す．過去のイベントの影響の時間変化を表すカーネル関数 $g(\tau)$ は $g(\tau) \geq 0$ および $g(\tau) = 0$ $(\tau \leq 0)$ を満たすとする．期間 $(0, T]$ に n 個のイベントが時刻 $\{t_1, \ldots, t_n\}$ に発生する確率密度関数は，条件付き強度関数(2.1)を用いて

$$\begin{aligned} p_{(0,T]}(t_1, \ldots, t_n) &= \left[\prod_{i=1}^n \lambda(t_i)\right] \exp\left[-\int_0^T \lambda(t)dt\right] \\ &= \prod_{i=1}^n \left[\mu + \sum_{j<i} g(t_i - t_j)\right] \exp\left[-\mu T - \sum_{j=1}^n \int_{t_j}^T g(t-t_j)dt\right] \end{aligned} \tag{2.2}$$

で与えられる (Daley and Vere-Jones, 2003, Chap.7)．

条件付き強度関数(2.1)は過去のすべてのイベントの影響を受けるので，各々のイベントが過去のどのイベントによって引き起こされたのかは定まらない．ここで，各イベントに対してそれを引き起こした “親イベント” を割り当てることで，Hawkes 過程に対応する分岐過程を構成することができる (Hawkes and Oakes, 1974; 近江・野村, 2019, 第 5 章)．i 番目のイベントを引き起こした親イベントの番号を $z_i \in \{0, 1, \ldots, i-1\}$ とする．$z_i = 0$ の場合は親イベントを持たないとする．イベント発生時刻と親イベントの番号を合わせた時系列 $\{(t_i, z_i) | i = 1, \ldots, n\}$ は以下のルールに従って生成されるとする．

(a) 親を持たないイベント $\{(t_i, z_i) | z_i = 0\}$ は強度 μ の Poisson 過程に従って発生する．
(b) j (≥ 1) 番目のイベントを親に持つイベント $\{(t_i, z_i) | z_i = j\}$ は強度 $g(t_i - t_j)$ の Poisson 過程にしたがって発生する．

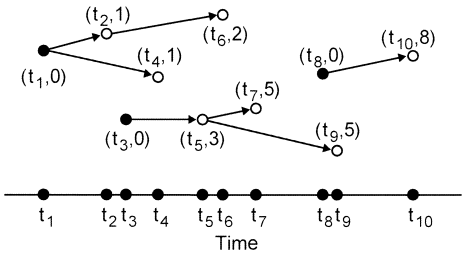

図 1. 上：分岐過程の系統図．黒丸と白丸はそれぞれ親を持たないイベント(a)と子イベント(b)を表し，矢印は親子関係を表す．下：イベント発生時刻だけを見ると Hawkes 過程に従う．

このようにして発生したイベントの親子関係を繋げていくと，親を持たないイベントを祖先とする分岐過程の系統図を描くことができる(図 1 上)．

イベント時系列全体は(a)と(b)の重ね合わせで与えられるとすると，期間 $(0,T]$ における $\{(t_i,z_i)|i=1,\ldots,n\}$ の確率密度関数は

$$p_{(0,T]}(\{t_i,z_i|i=1,\ldots,n\}) = p(\{t_i,z_i|z_i=0\})\prod_{j=1}^{n} p(\{t_i,z_i|z_i=j\}|t_j) \tag{2.3}$$

で与えられる．ここで

$$p(\{t_i,z_i|z_i=0\}) = \left[\prod_{i:z_i=0}\mu\right]\exp(-\mu T) \tag{2.4}$$

および

$$p(\{t_i,z_i|z_i=j\}|t_j) = \left[\prod_{i:z_i=j} g(t_i-t_j)\right]\exp\left[-\int_{t_j}^{T} g(t-t_j)dt\right] \tag{2.5}$$

はそれぞれ(a)と(b)のイベントの寄与を表す．式(2.4)と式(2.5)を式(2.3)に代入して整理すると確率密度関数は

$$p_{(0,T]}(\{t_i,z_i|i=1,\ldots,n\}) = \left[\prod_{i=1}^{n}\psi_{iz_i}\right]\exp\left[-\mu T - \sum_{j=1}^{n}\int_{t_j}^{T} g(t-t_j)dt\right] \tag{2.6}$$

と求められる．ここで

$$\psi_{ij} = \begin{cases} \mu, & j=0 \\ g(t_i-t_j), & 1\le j<i \end{cases}$$

とおいた．

式(2.6)はイベント発生時刻と親イベントの情報を含む分岐過程のモデルである．これを親イベント $\{z_1,\ldots,z_n\}$ について周辺化すると

$$\begin{aligned} p_{(0,T]}(t_1,\ldots,t_n) &= \sum_{z_1=0}^{0}\sum_{z_2=0}^{1}\cdots\sum_{z_n=0}^{n-1} p_{(0,T]}(\{t_i,z_i|i=1,\ldots,n\}) \\ &= \left[\prod_{i=1}^{n}\sum_{z_i=0}^{i-1}\psi_{iz_i}\right]\exp\left[-\mu T - \sum_{j=1}^{n}\int_{t_j}^{T} g(t-t_j)dt\right] \end{aligned} \tag{2.7}$$

$$= \prod_{i=1}^{n} \left[\mu + \sum_{j<i} g(t_i - t_j) \right] \exp \left[-\mu T - \sum_{j=1}^{n} \int_{t_j}^{T} g(t - t_j) dt \right]$$

となり Hawkes 過程の確率密度関数(2.2)が導かれる．すなわちイベント間の因果関係は観測されず，イベント発生時刻だけ観測されたものが Hawkes 過程であると解釈することができる(図 1 下)．

式(2.2)と式(2.6)から，イベント時刻列 $\{t_1, \ldots, t_n\}$ が与えられた下での親イベント $\{z_1, \ldots, z_n\}$ の条件付き確率は

$$\begin{aligned} p_{(0,T]}(z_1, \ldots, z_n | t_1, \ldots, t_n) &= \frac{p_{(0,T]}(\{t_i, z_i | i = 1, \ldots, n\})}{p_{(0,T]}(t_1, \ldots, t_n)} \\ &= \prod_{i=1}^{n} \frac{\psi_{iz_i}}{\lambda(t_i)} = \prod_{i=1}^{n} p(z_i | t_1, \ldots, t_i) \end{aligned}$$

と求められる．ここで

$$p(z_i | t_1, \ldots, t_i) = \frac{\psi_{iz_i}}{\lambda(t_i)} \tag{2.8}$$

は $\{t_1, \ldots, t_i\}$ が与えられた下での z_i の条件付き確率である．

親イベントの条件付き確率は様々な用途に応用されている．Zhuang et al. (2002)はこの確率に基づいて親イベントをサンプリングするアルゴリズムを提案した．また，親イベントを潜在変数と見なして，パラメータ推定のための EM アルゴリズムを構成することもできる（Veen and Schoenberg, 2008）．

3. Hawkes 型計数時系列モデル

ここからは，イベント発生件数が区間毎に集計された離散時間の計数時系列を考える．このようなデータに対して，Hawkes 過程と同様の性質を持つ計数時系列モデルを構成する．

3.1 モデルの定義

時点 $i \in \{1, 2, \ldots\}$ におけるイベント数を n_i とする．期待値 $\lambda = \mathrm{E}(n)$ をパラメータに持つイベント数の確率分布 $p(n, \lambda)$ を用いて，過去のイベント数 $\{n_1, \ldots, n_{i-1}\}$ が与えられた下での n_i の条件付き確率が

$$P(n_i | n_1, \ldots, n_{i-1}) = p(n_i, \lambda_i)$$

で与えられる計数時系列モデルを考える．Hawkes 過程の条件付き強度関数(2.1)からの類推で期待値パラメータを

$$\lambda_i = \mu + \sum_{j=1}^{i-1} g_{i-j} n_j \tag{3.1}$$

で与える．右辺第 1 項の μ はベースラインの頻度であり，第 2 項は過去のイベントの影響を表す．g_τ は過去のイベントの影響の時間変化を表し，$g_\tau \geq 0$ および $g_\tau = 0$ $(\tau \leq 0)$ を満たすとする．L 時点までの時系列 $\{n_1, \ldots, n_L\}$ の結合確率分布は

$$P(n_1, \ldots, n_L) = P(n_1) \prod_{i=2}^{L} P(n_i | n_1, \ldots, n_{i-1}) = \prod_{i=1}^{L} p(n_i, \lambda_i) \tag{3.2}$$

で与えられる．

イベント数の確率分布 $p(n,\lambda)$ には Poisson 分布もしくは負の二項分布を用いることにする．Poisson 分布の確率質量関数は

$$p(n,\lambda) = \frac{\lambda^n}{n!}e^{-\lambda} \tag{3.3}$$

で与えられる．平均と分散は等しく $\mathrm{E}(n) = \mathrm{Var}(n) = \lambda$ で与えられる．集計区間内の各々のイベント発生が互いに独立であるときイベント数は Poisson 分布に従うので，このような状況を近似的にも想定できる場合は Poisson 分布は妥当な選択である．

一方，集計区間内でイベントの自励効果を無視できないとき，イベント数の分散は平均よりも大きくなる．これを過分散(over-dispersion)という．このような場合には過分散を持つ確率分布を用いることが望ましい．負の二項分布はそのような分布のひとつである．ここでは以下の確率質量関数を持つ負の二項分布を考える（Koyama and Fujiwara, 2019）：

$$p(n,\lambda,\rho) = \frac{\Gamma(n+\frac{\lambda}{\rho})}{\Gamma(n+1)\Gamma(\frac{\lambda}{\rho})}\left(\frac{\rho}{1+\rho}\right)^n\left(\frac{1}{1+\rho}\right)^{\frac{\lambda}{\rho}} \tag{3.4}$$

平均と分散はそれぞれ $\mathrm{E}(n) = \lambda$ および $\mathrm{Var}(n) = (1+\rho)\lambda$ で与えられる．$\rho(>0)$ は過分散の度合いを表すパラメータであり，$\rho \to 0$ で式(3.4)は Poisson 分布(3.3)に収束する．負の二項分布については補足 A にまとめた．

イベント数の確率分布が Poisson 分布(3.3)もしくは負の二項分布(3.4)で与えられる時系列モデル(3.1)–(3.2)を “Hawkes 型” 計数時系列モデルと呼ぶことにする．他の確率分布でも時系列モデルは定義されるが，特にこれら二つの分布を用いる理由は第 3.3 節で明らかになる．

3.2 親イベントの割り当て

Hawkes 過程では，各イベントに親イベントをひとつ割り当てることで分岐過程と対応付けることができた．同様に Hawkes 型計数時系列モデルに対しても親イベントを割り当ててみよう．時点 i のイベント数 n_i のうち，過去の時点 $j(<i)$ のイベントに引き起こされたイベント数を y_{ij} とし，親を持たないイベント数を y_{ii} とする．つまり $\{y_{i1},\ldots,y_{ii}\}$ は n_i の発生要因についての内訳であり

$$n_i = \sum_{j=1}^{i} y_{ij} \tag{3.5}$$

を満たす．イベント数の内訳 y_{ij} の確率分布について以下を仮定する．

(a') 親を持たないイベント数 y_{ii} は期待値が $\mathrm{E}(y_{ii}) = \mu$ の確率分布 $P(y_{ii})$ に従う．
(b') 過去の時点 $j(<i)$ のイベントに引き起こされたイベント数 y_{ij} は期待値が $\mathrm{E}(y_{ij}|n_j) = g_{i-j}n_j$ の確率分布 $P(y_{ij}|n_j)$ に従う．

時点 i に発生するイベント数の内訳 $\{y_{i1},\ldots,y_{ii}\}$ は，過去のイベント数 $\{n_1,\ldots,n_{i-1}\}$ が与えられた下で互いに独立であるとする：

$$P(y_{i1},\ldots,y_{ii}|n_1,\ldots,n_{i-1}) = P(y_{ii})\prod_{j=1}^{i-1}P(y_{ij}|n_j) \tag{3.6}$$

すると，時点 1 から L までのイベント数の全内訳 $Y_{1:L} := \{y_{ij}|i=1,\ldots,L,\ j=1,\ldots,i\}$ の結合確率分布は

$$P(Y_{1:L}) = P(y_{11}) \prod_{i=2}^{L} P(y_{i1}, \ldots, y_{ii} | n_1, \ldots, n_{i-1}) = \prod_{i=1}^{L} P(y_{ii}) \prod_{j=1}^{i-1} P(y_{ij} | n_j) \tag{3.7}$$

で与えられる．式(3.7)は各時点のイベント数の内訳情報を含むモデルであり，Hawkes 過程に対する分岐過程(2.6)に対応している．

式(3.7)を $Y_{1:L}$ について式(3.5)を満たすように周辺化すると計数時系列 $\{n_1, \ldots, n_L\}$ の確率分布が得られる．すなわち $Y_i = \{y_{i1}, \ldots, y_{ii}\}$ とし，式(3.5)を満たす Y_i の値の集合を $\mathcal{Y}_i$ とすると

$$\begin{aligned} P(n_1, \ldots, n_L) &= \sum_{Y_1 \in \mathcal{Y}_1} \cdots \sum_{Y_L \in \mathcal{Y}_L} P(Y_{1:L}) \\ &= \sum_{Y_1 \in \mathcal{Y}_1} \cdots \sum_{Y_L \in \mathcal{Y}_L} \prod_{i=1}^{L} P(y_{ii}) \prod_{j=1}^{i-1} P(y_{ij} | n_j) \\ &= \prod_{i=1}^{L} \sum_{Y_i \in \mathcal{Y}_i} P(y_{ii}) \prod_{j=1}^{i-1} P(y_{ij} | n_j) \\ &= P(n_1) \prod_{i=2}^{L} P(n_i | n_1, \ldots, n_{i-1}) \end{aligned} \tag{3.8}$$

となる．ここで

$$P(n_i | n_1, \ldots, n_{i-1}) = \sum_{Y_i \in \mathcal{Y}_i} P(y_{ii}) \prod_{j=1}^{i-1} P(y_{ij} | n_j) \tag{3.9}$$

の期待値は，仮定(a')と(b')より

$$\begin{aligned} \mathrm{E}(n_i | n_1, \ldots, n_{i-1}) &= \mathrm{E}(y_{ii}) + \sum_{j=1}^{i-1} \mathrm{E}(y_{ij} | n_j) \\ &= \mu + \sum_{j=1}^{i-1} g_{i-j} n_j \end{aligned}$$

となり式(3.1)に一致する．すなわち式(3.7)をイベント数の内訳について周辺化することで Hawkes 型計数時系列モデルが得られた．これは，分岐過程(2.6)を親イベントについて周辺化することで Hawkes 過程が得られることに対応している．

3.3 加法性とイベント内訳の条件付き確率分布

Hawkes 過程に対する親イベントの条件付き確率(2.8)と同様に，各時点のイベント数 n_i が与えられた下での内訳 y_{ij} の条件付き確率を導こう．そのために，y_{ij} に対して以下の “加法性” の条件を加える．

定義. n 個の独立な確率変数の和

$$y = y_1 + \cdots + y_n, \quad y_i \sim p(y_i, \lambda_i)$$

が同一の確率分布族 $p(y, \lambda_1 + \cdots + \lambda_n)$ に従うとき加法的であるという．

イベント数の内訳 y_{ij} が加法的な確率分布 $p(y_{ij}, \psi_{ij})$ に従うとする．ここで

$$\psi_{ij} = \mathrm{E}(y_{ij}) = \begin{cases} \mu, & j = i \\ g_{i-j} n_j, & 1 \le j < i \end{cases}$$

とおいた．すると，式(3.5)および式(3.9)より，n_i は互いに独立な y_{ij} $(j = 1, \ldots, i)$ の和で与えられるから，加法性より同一の確率分布族

$$P(n_i | n_1, \ldots, n_{i-1}) = p(n_i, \lambda_i) \tag{3.10}$$

に従う．ここで λ_i は式(3.1)で与えられる．つまり加法性を仮定すると，各時点のイベント数 n_i もその内訳 y_{ij} $(j = 1, \ldots, i)$ も同一の確率分布族に従うのである．

式(3.6)と式(3.10)より，時点 i までのイベント数 $\{n_1, \ldots, n_i\}$ が与えられた下での時点 i のイベント数の内訳 $Y_i = \{y_{i1}, \ldots, y_{ii}\}$ の条件付き確率分布は

$$\begin{aligned} p(Y_i | n_1, \ldots, n_i) &= \frac{p(Y_i | n_1, \ldots, n_{i-1})}{p(n_i | n_1, \ldots, n_{i-1})} \\ &= \frac{\prod_{j=1}^{i} p(y_{ij}, \psi_{ij})}{p(n_i, \lambda_i)} \end{aligned} \tag{3.11}$$

と求められる．ここまで来ると，イベント数の分布に Poisson 分布(3.3)と負の二項分布(3.4)を採用した理由が明らかになる．すなわち，これらはともに加法的である．それぞれの確率質量関数を式(3.11)に代入することで，Y_i の条件付き確率を具体的に導くことができる．

Poisson 分布の場合　式(3.3)を式(3.11)に代入すると，Y_i の条件付き確率分布は多項分布

$$p(Y_i | n_1, \ldots, n_i) = \frac{n_i!}{\prod_{j=1}^{i} y_{ij}!} \prod_{j=1}^{i} \left(\frac{\psi_{ij}}{\lambda_i} \right)^{y_{ij}} \tag{3.12}$$

として求められる．平均と分散はそれぞれ

$$\mathrm{E}(y_{ij} | n_1, \ldots, n_i) = \frac{n_i \psi_{ij}}{\lambda_i} \tag{3.13}$$

$$\mathrm{Var}(y_{ij} | n_1, \ldots, n_i) = \frac{n_i \psi_{ij}}{\lambda_i} \left(1 - \frac{\psi_{ij}}{\lambda_i} \right) \tag{3.14}$$

で与えられる．

負の二項分布の場合　式(3.4)を式(3.11)に代入すると，Y_i の条件付き分布確率分布は

$$p(Y_i | n_1, \ldots, n_i, \rho) = \frac{\Gamma(n_i + 1) \Gamma(\frac{\lambda_i}{\rho})}{\Gamma(n_i + \frac{\lambda_i}{\rho})} \prod_{j=1}^{i} \frac{\Gamma(y_{ij} + \frac{\psi_{ij}}{\rho})}{\Gamma(y_{ij} + 1) \Gamma(\frac{\psi_{ij}}{\rho})} \tag{3.15}$$

と導かれる．これはディリクレ多項分布と呼ばれ，平均と分散はそれぞれ以下で与えられる：

$$\mathrm{E}(y_{ij} | n_1, \ldots, n_i) = \frac{n_i \psi_{ij}}{\lambda_i} \tag{3.16}$$

$$\mathrm{Var}(y_{ij} | n_1, \ldots, n_i) = \kappa_i \frac{n_i \psi_{ij}}{\lambda_i} \left(1 - \frac{\psi_{ij}}{\lambda_i} \right) \tag{3.17}$$

ここで分散に掛かる係数 κ_i は

$$\kappa_i = \frac{\lambda_i + \rho n_i}{\lambda_i + \rho} \ (\ge 1) \tag{3.18}$$

で与えられる．多項分布の分散(3.14)よりも大きいことは，元になる負の二項分布が過分散であることに対応している．

4. 感染症モデルへの応用

前節で構成した Hawkes 型計数時系列モデルを感染症の実効再生産数 (effective reproduction number) の推定に応用する．再生産数は 1 人の感染者が引き起こす 2 次感染者数の平均である．実効再生産数はすでに感染が広がっている状況における再生産数の総称で，幾つかの異なる定義がある（Fraser, 2007; Nishiura and Chowell, 2009）．ここでは Wallinga and Teunis（2004）による実効再生産数を Hawkes 型計数時系列モデルから導き，過分散の場合に一般化する．

4.1 Wallinga and Teunis（2004）の方法

感染者 k が感染者 l から感染したとし，それぞれの感染時刻を t_k, t_l とする．世代時間が確率密度関数 $\phi(\tau)$ に従うとすると，感染源候補の中で k が l から感染した確率は

$$p_{(k,l)} = \frac{\phi(t_k - t_l)}{\sum_{m \neq k} \phi(t_k - t_m)}$$

で与えられる．全ての感染者が独立に 2 次感染者を感染させていたと仮定すると，感染者 l の 2 次感染者数は

$$R_l \sim \sum_k \mathrm{Bernoulli}(p_{(k,l)}) \tag{4.1}$$

に従う．ここで $\sum_k$ は時刻 t_l 以降の感染者すべてについての和である．t 日目の実効再生算数 R_t を，t 日目に感染した患者の 2 次感染者数の平均 $R_t = \frac{1}{n_t} \sum_{t_l = t} R_l$ で定義する．ここで n_t は t 日目の感染者数である．式(4.1)から t 日目の実効再生算数の平均と分散はそれぞれ

$$\hat{R}_t = \mathrm{E}(R_t) = \frac{1}{n_t} \sum_{t_l = t} \sum_k p_{(k,l)} \tag{4.2}$$

$$s^2 = \mathrm{Var}(R_t) = \frac{1}{n_t^2} \sum_k \left[\sum_{t_l = t} p_{(k,l)}(1 - p_{(k,l)}) - \sum_{t_l = t} \sum_{t_m = t, m \neq l} p_{(k,l)} p_{(k,m)} \right] \tag{4.3}$$

で与えられる（Cowling et al., 2008）．これが Wallinga and Teunis（2004）によって与えられた実効再生産数の推定量である．

4.2 Hawkes 型計数時系列モデルからの導出

感染症の伝播を Hawkes 型計数時系列モデルで表す．n_i を i 日目の新規感染者数とし，頻度が

$$\lambda_i = r_i \sum_{j=1}^{i-1} \phi_{i-j} n_j \tag{4.4}$$

で与えられるとする．ここで ϕ_τ は世代時間の分布であり $\sum_{\tau=0}^{\infty} \phi_\tau = 1$ を満たす．r_i は i 日目の感染力を表すパラメータとする．このとき，イベント数の内訳 y_{ij} は “i 日目の感染者のうち j 日目の感染者から感染した人数” となる．Wallinga and Teunis（2004）に従うと，j 日目の実効再生産数はこの日の感染者 1 人当たりの 2 次感染者数であるから，y_{ij} を用いて

$$R_j = \frac{1}{n_j}\sum_{i>j} y_{ij}$$

で与えられる．

L 日目までの新規感染者数の時系列データ $\{n_1,\ldots,n_L\}$ が与えられているとする．感染者数が Poisson 分布に従うとすると，データが与えられた下での y_{ij} の条件付き確率は多項分布(3.12)に従うので，R_j の条件付き期待値と分散は式(3.13)と式(3.14)を用いて

$$\begin{aligned}\hat{R}_j &= \mathrm{E}(R_j|n_1,\ldots,n_L)\\ &= \frac{1}{n_j}\sum_{i=j+1}^{L}\mathrm{E}(y_{ij}|n_1,\ldots,n_i)\\ &= \sum_{i=j+1}^{L}\frac{n_i\phi_{i-j}}{\sum_{k=1}^{i-1}\phi_{i-k}n_k}\end{aligned} \tag{4.5}$$

および

$$\begin{aligned}s^2 &= \mathrm{Var}(R_j|n_1,\ldots,n_L)\\ &= \frac{1}{n_j^2}\sum_{i=j+1}^{L}\mathrm{Var}(y_{ij}|n_1,\ldots,n_i)\\ &= \frac{1}{n_j}\sum_{i=j+1}^{L}\frac{n_i\phi_{i-j}}{\sum_{k=1}^{i-1}\phi_{i-k}n_k}\left(1-\frac{\phi_{i-j}n_j}{\sum_{k=1}^{i-1}\phi_{i-k}n_k}\right)\end{aligned} \tag{4.6}$$

と求められ，それぞれ式(4.2)と式(4.3)に一致する(証明は付録 B を参照)．すなわち，Poisson 分布を用いた Hawkes 型計数時系列モデルから Wallinga and Teunis (2004)による実効再生産数の推定量と分散が導かれた．

一方，感染者数が負の二項分布(3.4)に従うとすると，実効再生産数の条件付き期待値は式(4.5)と同じであるが，条件付き分散は式(3.17)より

$$s^2 = \frac{1}{n_j}\sum_{i=j+1}^{T}\kappa_i\frac{n_i\phi_{i-j}}{\sum_{k=1}^{i-1}\phi_{i-k}n_k}\left(1-\frac{\phi_{i-j}n_j}{\sum_{k=1}^{i-1}\phi_{i-k}n_k}\right) \tag{4.7}$$

となる．過分散のパラメータ κ_i が掛かることに着目しよう．$\kappa_i = 1$ のときは式(4.6)に帰着される．したがって，式(4.7)は Wallinga and Teunis (2004)の方法を過分散の場合に一般化したものとみなせる．

4.3 COVID-19 への応用

上で導いた実効再生産数の推定量(4.5)–(4.7)を新型コロナウィルス感染症(COVID-19)のデータに応用する．世界各国の 1 日の新規陽性者数は毎日報告され，オープンデータとして公開されている．ここでは，厚生労働省が公開するデータ(https://www.mhlw.go.jp/stf/covid-19/open-data.html)を用いた．

図 2 上に日本の新規陽性者数を示す．陽性者数にみられる 1 週間の周期変動は，平日と週末の検査態勢の違いに起因する．この周期変動の実効再生産数への影響を抑えるため，曜日毎の新規陽性者数の平均を取り，

$$\frac{1}{7}\sum_{\text{Sunday}}^{\text{Saturday}}\beta_i = 1$$

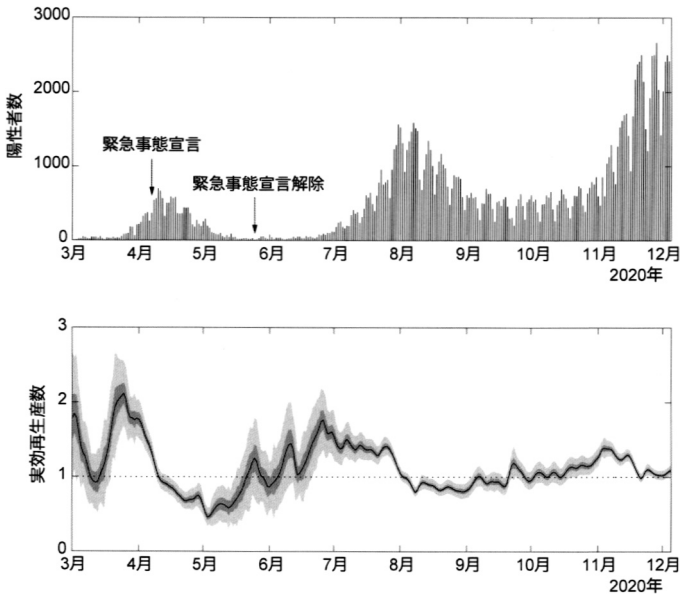

図 2. 上：日本の新規陽性者数．下：推定した実効再生産数(平均値 ±SD)．薄い灰色はデータの過分散を考慮に入れた場合を表し，濃い灰色は考慮に入れない場合を表す．

と規格化して求めた重み β_i を用いて，もとの新規陽性者数 n_i $(i = 1, \ldots, L)$ を $\tilde{n}_i = n_i/\beta_i$ $(i = 1, \ldots, L)$ と変換した(小数部は最も近い整数に丸め込んだ)．

発症間隔(serial interval)分布 ϕ_τ には，Nishiura et al. (2020)の報告に従い，平均 4.7 日，標準偏差 2.9 日の対数正規分布を採用し，累積分布関数

$$\Phi_\tau = \frac{1}{2}\mathrm{erfc}\left(-\frac{\log\tau - \mu}{\sqrt{2\sigma^2}}\right), \tag{4.8}$$

を用いて $\phi_\tau = \Phi_\tau - \Phi_{\tau-1}$ とした．対数正規分布のパラメータ (μ, σ^2) は，平均 $m = 4.7$，標準偏差 $s = 2.9$ として，$\mu = \log(m^2/\sqrt{s^2 + m^2})$ および $\sigma^2 = \log(1 + s^2/m^2)$ で与えた．

負の二項分布(3.4)のパラメータ ρ の値はデータから

$$\hat{\rho} = \frac{1}{L}\sum_{i=1}^{L}\frac{(\tilde{n}_i - \bar{\lambda}_i)^2}{\bar{\lambda}_i} - 1$$

と推定した．ここで $\bar{\lambda}_i = \sum_{j=-3}^{3}\tilde{n}_{i+j}/7$ は 1 週間の新規陽性者数の平均である．これと式(3.18)より式(4.7)のパラメータ κ_i は

$$\hat{\kappa}_i = \frac{\bar{\lambda}_i + \hat{\rho}\tilde{n}_i}{\bar{\lambda}_i + \hat{\rho}}$$

と求められる．

図 2 下に推定した実効再生産数(平均値 ±SD)を示す．薄い灰色はデータの過分散を考慮に入れた場合を表し，濃い灰色は考慮に入れない場合 (Wallinga and Teunis, 2004 の方法)を表す．実際のデータは過分散であるため，Wallinga and Teunis (2004)の方法では推定量の分散が過少に評価されている．

実効再生産数は 4 月に下がり始め 1 を下回る．緊急事態宣言が発出されたのは 4 月 7 日である．しばらく実効再生産数は 1 を下回るが，5 月 25 日に緊急事態宣言が全国で解除され，夏にかけて再び 1 を超える．8 月に実効再生産数は再び 1 を下回るが，秋から冬にかけて陽性者数は急増している．

5. まとめ

本稿では，Hawkes 過程の性質，特に分岐過程との対応に着目して，同様な性質を持つ計数時系列モデルを提案した．第 3 節の Hawkes 型計数時系列モデルの構成を，第 2 節の Hawkes 過程とパラレルに展開させ，両者の対応をみた．ただし計数時系列モデルに対して，確率分布の加法性という条件を課した．これは Hawkes 過程にはない要請である．第 4 節では，Wallinga and Teunis (2004) による実効再生産数の推定量を Hawkes 型計数時系列モデルをから導出した．さらに感染者数に負の二項分布を仮定することで，実効再生産数の推定量が過分散の場合に一般化された．

Hawkes 過程と計数時系列を関連付けるアプローチとして，Kirchner (2016, 2017) は整数値自己回帰モデル (Integer-valued Autoregressive Models，INAR モデル) に基づく方法を提案している．この方法では，離散観測された Hawkes 過程を INAR(p) モデルで近似し，条件付き最小 2 乗法で求めた推定量を Hawkes 過程のカーネル関数に対応させている．比較的計算が容易な行列演算でカーネル関数のノンパラメトリックな推定量を与えていることが利点である．パラメトリックな関数を事前に想定できないとき，Kirchner の方法でノンパラメトリックに推定し，それをここで提案した時系列モデルのカーネル関数に使うこともできるであろう．

本論文では，もう一つのアプローチである一般状態空間モデル (あるいは動的一般化線形モデル) を扱わなかったが，Hawkes 型計数時系列モデルを観測モデルとし，パラメータ (3.1) にシステムモデルを組み込むことで，容易に一般状態空間モデルとして定式化することもできる．Koyama et al. (2021) は，このアプローチで非ガウス平滑化アルゴリズムを用いた実効再生産数の推定方法を提案している．

Hawkes 過程の分岐過程による表現は，デクラスタリング (Zhuang et al., 2002)，シミュレーション (Moller and Rasmussen, 2005)，EM アルゴリズム (Veen and Schoenberg, 2008) など様々なアルゴリズムや応用の基礎になる．Hawkes 型計数時系列モデルに対する同様の展開は今後の研究課題である．

付 録

A. 負の二項分布

負の二項分布は通常，成功率 p のベルヌーイ試行で r 回成功する前に失敗する回数 y の確率分布

$$p(y,p,r) = \binom{y+r-1}{r-1} p^r (1-p)^y \tag{A.1}$$

として表される．平均と分散はそれぞれ

$$\mathrm{E}(y) = \frac{r(1-p)}{p}, \quad \mathrm{Var}(y) = \frac{r(1-p)}{p^2}$$

である．パラメータを

$$p = \frac{1}{1+\rho}, \quad r = \frac{\lambda}{\rho}$$

と変換して階乗をガンマ関数 $\Gamma(x+1) = x!$ で表すと式(3.4)の形が得られる．

式(3.4)のキュムラント母関数は

$$K(s) = \mathrm{E}(e^{sy}) = -\frac{\lambda}{\rho}\log[1-(e^s-1)\rho]$$

で与えられる．ρ で展開して主要項を取り出すと

$$K(s) = \frac{\lambda}{\rho}[(e^s-1)\rho + o(\rho)]$$

となるので，$\rho \to 0$ で Poisson 分布のキュムラント母関数 $K(s) = \lambda(e^s-1)$ に収束する．すなわち負の二項分布(3.4)は $\rho \to 0$ で Poisson 分布(3.3)に収束する．

y_i $(i = 1, \ldots, n)$ が互いに独立な負の二項分布(3.4)に従うとすると，$y = y_1 + \cdots + y_n$ の分布のキュムラント母関数は

$$K_n(s) = -\frac{\sum_{i=1}^n \lambda_i}{\rho}\log[1-(e^s-1)\rho]$$

となるので，y は期待値パラメータ $\lambda_1 + \cdots + \lambda_n$ を持つ同一の分布に従う．したがって負の二項分布(3.4)は加法的である．

式(A.1)のパラメータを

$$p = \frac{1}{1+\rho\lambda}, \quad r = \frac{1}{\rho}$$

と変換すると一般化線形モデルでよく使われるもう一つのパラメータ表示

$$p(y, \lambda, \rho) = \frac{\Gamma(y+\rho^{-1})}{\Gamma(y+1)\Gamma(\rho^{-1})}\left(\frac{\rho\lambda}{1+\rho\lambda}\right)^y\left(\frac{1}{1+\rho\lambda}\right)^{\frac{1}{\rho}} \tag{A.2}$$

が得られる（Hilbe, 2011）．平均と分散はそれぞれ $\mathrm{E}(y) = \lambda$ および $\mathrm{Var}(y) = \lambda + \rho\lambda^2$ で与えられる．式(3.4)の分散と異なることに注意しよう．Cameron and Trivedi（1986）は式(3.4)と式(A.2)をそれぞれ “Negbin I” および “Negbin II” と呼んで区別している．Negbin I は加法性を持つが，Negbin II は加法的ではない．このため Hawkes 型計数時系列モデルには Negbin I を用いる．

B. Wallinga and Teunis（2004）の方法との等価性の証明

感染者 k と l の感染した日をそれぞれ $t_k = s$, $t_l = t$ とすると $\phi(t_k - t_l) = \phi_{s-t}$ である．$\sum_k \phi(t_k - t_l) = \sum_s n_s\phi_{s-t}$ に注意すると式(4.2)は

$$\hat{R}_t = \frac{1}{n_t}\sum_{t_l=t}\sum_k \frac{\phi(t_k-t_l)}{\sum_{m\neq k}\phi(t_k-t_m)} = \sum_{s=t+1}^{L}\frac{n_s\phi_{s-t}}{\sum_{u=1}^{s-1}\phi_{s-u}n_u}$$

となり式(4.5)に一致する．

同様に式(4.3)の右辺第 1 項と第 2 項はそれぞれ

$$\frac{1}{n_t^2}\sum_k\sum_{t_l=t} p_{(k,l)}(1-p_{(k,l)}) = \frac{1}{n_t^2}\sum_{t_l=t}\sum_k \frac{\phi(t_k-t_l)}{\sum_{m\neq k}\phi(t_k-t_m)}\left(1-\frac{\phi(t_k-t_l)}{\sum_{m\neq k}\phi(t_k-t_m)}\right)$$

$$= \frac{1}{n_t} \sum_{s=t+1}^{L} \frac{n_s \phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \left(1 - \frac{\phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \right)$$

$$\frac{1}{n_t^2} \sum_k \sum_{t_l = t} \sum_{t_m = t, m \neq l} p_{(k,l)} p_{(k,m)}$$

$$= \frac{n_t - 1}{n_t} \sum_k \left(\frac{1}{n_t} \sum_{t_l = t} \frac{\phi(t_k - t_l)}{\sum_{j \neq k} \phi(t_k - t_j)} \right) \left(\frac{1}{n_t - 1} \sum_{t_m = t, m \neq l} \frac{\phi(t_k - t_m)}{\sum_{j \neq k} \phi(t_k - t_j)} \right)$$

$$= \frac{n_t - 1}{n_t} \sum_{s=t+1}^{L} \frac{n_s \phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \frac{\phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u}$$

となるので，これらを式(4.3)の右辺に戻すと

$$s^2 = \frac{1}{n_t} \sum_{s=t+1}^{L} \frac{n_s \phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \left(1 - \frac{\phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \right) - \frac{n_t - 1}{n_t} \sum_{s=t+1}^{L} \frac{n_s \phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \frac{\phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u}$$

$$= \frac{1}{n_t} \sum_{s=t+1}^{L} \frac{n_s \phi_{s-t}}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \left(1 - \frac{\phi_{s-t} n_t}{\sum_{u=1}^{s-1} \phi_{s-u} n_u} \right)$$

となり式(4.6)に一致する．

参 考 文 献

Bacry, E., Mastromatteo, I. and Muzy, J. F. (2015). Hawkes processes in finance, *Market Microstructure and Liquidity*, **1**, 1550005.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests, *Journal of Applied Econometrics*, **1**, 29–53.

Cheysson, F. and Lang, G. (2020). Strong mixing condition for Hawkes processes and application to Whittle estimation from count data (preprint), arXiv:2003.04314.

Chiang, W. H., Liu, X. and Mohler, G. (2020). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates (preprint), medRxiv:2020.06.06.20124149.

Chornoboy, E. S., Schramm, L. P. and Karr, A. F. (1988). Maximum likelihood identification of neural point process systems, *Biological Cybernetics*, **59**, 265–275.

Cowling, B. J., Ho, L. M. and Leung, G. M. (2008). Effectiveness of control measures during the SARS epidemic in Beijing: A comparison of the Rt curve and the epidemic curve, *Epidemiology and Infection*, **136**, 562–566.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes Vol.1*: *Elementary Theory and Methods*, 2nd ed., Springer-Verlag, New York.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*, 2nd ed., Springer-Verlag, New York.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D. and Bertozzi, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes, *Journal of the American Statistical Association*, **111**, 564–584.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic, *PLoS ONE*, **2**, e758.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes, *Journal of the Royal Statistical Society. Series B* (*Methodological*), **33**, 438–443.

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes, *Biometrika*, **58**, 83–90.

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: A review, *Quantitative Finance*, **18**, 193–198.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process, *Journal of Applied Probability*, **11**, 493–503.

Hilbe, J. M. (2011). *Negative Binomial Regression*, 2nd ed., Cambridge University Press, Cambridge.

Kalair, K., Connaughton, C. and Loro, P. A. D. (2021). A non-parametric Hawkes process model of primary and secondary accidents on a UK smart motorway, *Journal of the Royal Statistical Society Series. C* (*Applied Statistics*), **70**, 80–97.

Kirchner, M. (2016). Hawkes and INAR(∞) processes, *Stochastic Processes and Their Applications*, **126**, 2494–2525.

Kirchner, M. (2017). An estimation procedure for the Hawkes process, *Quantitative Finance*, **17**, 571–595.

Kitagawa, G. (2010). *Introduction to Time Series Modeling*, Chapman & Hall/CRC, Boca Raton.

Kobayashi, R. and Lambiotte, R. (2016). TiDeH: Time-dependent Hawkes process for predicting retweet dynamics, *ICWSM 2016*, 191–200.

Koyama, S. and Fujiwara, Y. (2019). Modeling event cascades using networks of additive count sequences, *Journal of Statistical Mechanics*: *Theory and Experiment*, **2019**, 023402.

Koyama, S. and Shinomoto, S. (2020). Statistical physics of discovering exogenous and endogenous factors in a chain of events, *Physical Review Research*, **2**, 043358.

Koyama, S., Horie, T. and Shinomoto, S. (2021). Estimating the time-varying reproduction number of COVID-19 with a state-space method, *PLOS Computational Biology*, **17**, e1008679.

Lewis, E., Mohler, G., Brantingham, P. J. and Bertozzi, A. L. (2012). Self-exciting point process models of civilian deaths in Iraq, *Security Journal*, **25**, 244–264.

Meyer, S., Elias, J. and Hohle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence, *Biometrics*, **68**, 607–616.

Mohler, G., Short, M. B., Brantingham, P. J., Schoenberg, F. P. and Tita, G. E. (2011). Self-exciting point process modeling of crime, *Journal of the American Statistical Association*, **106**, 100–108.

Moller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes, *Advances in Applied Probability*, **37**, 629–646.

中嶋雅彦, 酒折文武, 川崎能典 (2017). 整数値自己回帰モデルの最近の発展, 統計数理, **65**, 323–339.

Nishiura, H. and Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends, *Mathematical and Statistical Estimation Approaches in Epidemiology*, 103–121, Springer, Dordrecht.

Nishiura, H., Lintona, N. M. and Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections, *International Journal of Infectious Diseases*, **93**, 284–286.

近江崇宏, 野村俊一 (2019).『点過程の時系列解析』, 共立出版, 東京.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *Journal of the American Statistical Association*, **83**(401), 9–27.

Pernice, V., Staude, B., Cardanobile, S. and Rotter, S. (2011). How structure determines correlations in neuronal networks, *PLoS Computational Biology*, **7**(5), e1002059.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm, *Journal of the American Statistical Association*, **103**, 614–624.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures, *American Journal of Epidemiology*, **160**, 509–516.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed., Springer-Verlag, New York.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal Hawkes — Type point process

model with periodic background for crime data, *Journal of the Royal Statistical Society. Series A* (*Statistics in Society*), **182**, 919–942.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences, *Journal of the American Statistical Association*, **97**, 369–380.

Hawkes-type Count Time Series Models

Shinsuke Koyama

The Institute of Statistical Mathematics

We propose a "Hawkes-type" count time series model. By emphasizing the branching process representation of the Hawkes process on a real number line, our model has a similar representation. The applicability of the proposed model is demonstrated using an epidemiological example where the effective reproduction number proposed by Wallinga and Teunis (2004) is generalized for over-dispersion.

Key words: Hawkes process, branching process, count time series model, negative binomial distribution, effective reproduction number.