

GPIGINAR(1)模型的贝叶斯估计

李 平, 卢飞龙*

辽宁科技大学理学院, 辽宁 鞍山

收稿日期: 2025年3月25日; 录用日期: 2025年4月16日; 发布日期: 2025年4月27日

摘 要

当前研究表明, 整数值时间序列通常呈现出重尾特征。然而, 能够处理这类具有重尾特征数据的模型相对较少。本文聚焦于广义泊松逆高斯INAR(1) (GPIGINAR(1))模型, 该模型能够拟合重尾特征。文中采用自适应马尔可夫链蒙特卡洛(MCMC)算法, 对该模型中所关注的参数进行贝叶斯估计, 并将其结果与极大似然估计结果进行对比, 以此证明所提方法的有效性。最后, 通过具体实例进一步说明该模型在处理此类数据时的有效性, 以及该算法在解决这类问题时的可行性。

关键词

GPIGINAR(1)模型, 贝叶斯估计, MCMC算法

Bayesian Estimation of the GPIGINAR(1) Model

Ping Li, Feilong Lu*

College of Science, University of Science and Technology Liaoning, Anshan Liaoning

Received: Mar. 25th, 2025; accepted: Apr. 16th, 2025; published: Apr. 27th, 2025

Abstract

Current research shows that integer-valued time series usually exhibit heavy-tailed characteristics. However, relatively few models are capable of handling data with such heavy-tailed characteristics. This paper focuses on the Generalized Poisson Inverse Gaussian INAR(1) (GPIGINAR(1)) model, which is able to fit heavy-tailed features. In this paper, the Adaptive Markov Chain Monte Carlo (MCMC) algorithm is adopted to perform Bayesian estimation on the parameters of interest in this model. The results are then compared with those of maximum likelihood estimation to prove the effectiveness of the proposed method. Finally, specific examples are used to further illustrate the

*通讯作者。

effectiveness of this model in processing such data and the feasibility of the algorithm in solving such problems.

Keywords

GPIGINAR(1) Model, Bayesian Estimation, MCMC Algorithm

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

对于重尾型的整数值时间序列, 还很少被考虑, 在时间序列中, 经常可以观察到数据呈现重尾的特征, 这也就意味着尾部概率是不可忽略的。如果忽略尾部, 那么可能会导致有用信息的丢失。对于具有重尾特征的数据, Maiti 等(2018) [1]提出了一类新的 INAR(1)过程, 用于模拟存在大量零和一的低计数的重尾时间序列。泊松逆高斯(PIG)分布, 可以建模比负二项分布和泊松分布尾部要厚的数据。该分布已被许多研究者研究[2]-[5], Barreto-Souza (2019) [6]介绍了边际分布为 PIG 分布的 INAR 模型。钱连勇(2021) [7]研究了若干的重尾整数值时间序列模型, 其中基于 GPIG 分布为新息项的 INAR 模型为本文所讨论的模型。

随着贝叶斯统计影响的扩大, 学者们对贝叶斯方法进行了深入研究。Bolstad (2007) [8]指出, 使用先验信息的贝叶斯分析通常会比忽略先验信息的频率论者提供更好的推断。通常情况下, 后验分布的计算需要进行积分, 但大部分情况是该积分没有显式解, 因此可以采用 MCMC 的方法从复杂的后验分布中获取样本, 从而实现对参数的贝叶斯推断。Hay 和 Pettitt (2001) [9]率先对广义线性混合模型的参数展开了全面的贝叶斯分析, 并深入探究其在传染病控制方面的应用, 为后续贝叶斯方法在各类模型中的运用奠定了基础。从自适应 MCMC 算法的历程来看, Chen 和 Lee (1995) [10]创新性地提出自适应 MCMC 算法, 用于估计阈值自回归(TAR)模型和双线性模型, 开启了自适应 MCMC 算法在模型参数估计领域的应用篇章。进入 21 世纪, 相关研究不断拓展深化。Chen 和 So (2006) [11]运用 MCMC 自适应算法对阈值异方差模型进行参数估计, 进一步拓展 MCMC 自适应算法的应用范围。Chen 等(2019) [12]针对气象变量存在的滞后依赖、过分散、连续零、非线性动力学及时变系数等复杂问题, 在贝叶斯框架下成功构建马尔可夫切换泊松 INGARCH 模型, 有效解决了气象数据处理中的难题。与此同时, 整数值时间序列模型的研究也取得显著进展。Ausín 和 Galeano (2007) [13]研究高斯混合广义自回归条件异方差(GARCH)模型的贝叶斯估计问题, Bu 和 McCabe (2008) [14]运用 MCMC 方法深入探讨 INAR (p)模型的模型选择、估计及预测问题, 朱复康和李琦(2009) [15]聚焦 INGARCH 模型的贝叶斯估计问题, Neal 和 Rao (2007) [16]探究整数值 ARMA 过程的贝叶斯估计, 张哲等(2010) [17]借助贝叶斯方法对 INAR(1)模型的参数估计展开研究, Drovandi 等(2016) [18]则考虑了低维整数值时间序列模型的贝叶斯推理。这些研究逐步完善了整数值时间序列模型的贝叶斯分析体系。在特定模型的参数估计与推断方面, 众多学者成果丰硕。Truong 等 (2017) [19]提出在 MCMC 采样方法下对泊松 INGARCH 模型进行参数估计, 并采用贝叶斯信息准则进行模型比较, Chen 和 Lee (2017) [20]提出基于泊松、负二项和对数线性泊松 INGARCH 模型的贝叶斯因果检验方法, 并应用于气候和犯罪数据, Chu 和 Yu (2023) [21]对对数 BNB-INGARCH 模型应用 MCMC 自适应算法进行贝叶斯推断。此外, 王佩(2019) [22]对门限 ARMA 模型的贝叶斯估计展开研究, Li 等(2019)

[23]提出带有解释变量的随机波动模型, 并思考该模型的贝叶斯估计问题, Garay 等(2020) [24]通过设置潜在变量的形式, 利用 M-H 算法对 ZINAR(p)模型进行贝叶斯推断, 黄菊(2021) [25]研究空间自回归条件异方差模型的贝叶斯估计, Yang 等(2022) [26]采用 MCMC 算法对阈值一阶自回归模型进行贝叶斯推断, 于鑫洋(2022) [27]研究自激励整数值门限自回归模型的贝叶斯估计及应用。Kang Y 等(2024) [28]对改进的混合二项自回归模型进行了贝叶斯分析, 并将其应用于雨天数和空气质量水平数据研究。因此本文将基于 MCMC 算法讨论 GPIGINAR(1)模型的贝叶斯估计问题。

2. 模型定义及其性质

2.1. GPIG 分布

在本节中介绍 GPIG 分布的定义及其基本性质, Zhu (2009) [29]中给出了定义:

定义 2.1. 若随机变量 Y 有如下所示的概率生成函数的形式:

$$G_y(h; a, b, c) = \exp\left\{b\left[(1-c)^a - (1-ch)^a\right]\right\} \quad (1)$$

其中 $0 < a \leq 1$, $b > 0$, $0 \leq c \leq 1$ 。则称该随机变量 Y 服从 GPIG 分布, 记为 $GPIG(a; b; c)$ 。

注 2.1. 接下来讨论边界与特殊情况:

(1) 当 $a=1$, $c \neq 0$ 时, 其概率生成函数为 $G_y(h; a, b, c) = \exp\{bc(h-1)\}$, 是均值为 bc 的泊松分布。

(2) 当 $c=1$ 且 $0 < a \leq 1$ 时, 其概率生成函数为 $G_y(h; a, b, c) = \exp\{-b(1-h)^a\}$, 是稳定的离散分布。

(3) 当 $a=0$ 或 $c=0$ 时, 其概率生成函数为 $G_y(h; a, b, c) = 1$, 是零膨胀分布。

(4) 当 $a = \frac{1}{2}$, $b = \lambda\theta^{-1}\sqrt{1+2\theta^2/\lambda}$ 且 $c = 2\theta^2\lambda^{-1}/(1+2\theta^2/\lambda)$ 时, 其概率生成函数为

$$G_y(h; a, b, c) = \exp\left\{\lambda\theta^{-1}\left(1 - \left[1 + 2\theta^2\lambda^{-1}(1-h)\right]^{\frac{1}{2}}\right)\right\},$$

其中 $\lambda > 0$, $\theta > 0$ 时, 为 PIG 分布。

注 2.2. 假设 $Y \sim GPIG(a, b, c)$, 有

$$(1-ch)^{1-a} G'(h) = abcG(h) \quad (2)$$

当 $0 \leq c < 1$ 时, 其均值 μ , 方差 σ^2 , 离散指数 I 以及偏度 γ 分别为:

$$\mu = abc/(1-c)^{1-a}, \quad \sigma^2 = \frac{abc(1-ac)}{(1-c)^{2-a}}, \quad I = \frac{\sigma^2}{\mu} = \frac{1-ac}{1-c}, \quad \gamma = \frac{(1-c)^{a-3} abc[1+c-3ac+a^2c^2]}{\sigma^3}.$$

概率密度函数 $\{m_k\}$ 的具体形式通过递推算法得到, 对于 GPIG 分布来说, m_0 与 m_1 很容易被计算, 因为它们能够通过计算概率生成函数在 $h=0$ 处的一阶导数和二阶导数的值。对式(2)进行泰勒展开, 存在实数序列 $\{l_i : i=1, 2, \dots\}$, 满足以下等式:

$$\begin{aligned} (1-ch)^{1-a} &= 1 - (1-ah)ch - \frac{(1-a)a}{2!}c^2h^2 - \frac{(1-a)a(1+a)}{3!}c^3h^3 - \dots \\ &= 1 - l_1h - l_2h^2 - l_3h^3 - \dots \end{aligned}$$

$$l_1 = (1-a)c, \quad l_{i+1} = \left(\frac{i-1+a}{i+1}\right)cl_i, \quad i=1, 2, \dots.$$

如果 $G_Y(h; a, b, c) = \sum m_i h^i$ 且有 $m_i = m_i(a, b, c)$, 对式(2)进行泰勒展开并进行再一次整理, 可以得到:

$$\begin{aligned} & abc[m_0 + m_1 h + m_2 h^2 + m_3 h^3 + \dots] \\ &= [1 - l_1 h - l_2 h^2 - l_3 h^3 - \dots] \times [m_1 + 2m_2 h + 3m_3 h^2 + 4m_4 h^3 + \dots] \\ &= m_1 + (2m_2 - m_1 l_1)h + (3m_3 - 2m_2 l_1 - m_1 l_2)h^2 + (4m_4 - 3m_3 l_1 - m_2 l_2 - m_1 l_3)h^3 \end{aligned}$$

从而可以得到概率密度函数 $\{m_k\}$ 的递归关系有:

$$m_0 = e^{b[(1-c)^a - 1]}, \quad m_1 = abcm_0, \quad m_{r+1} = \frac{1}{1+r} \left(abcm_r + \sum_{i=1}^r i l_{r+1-i} m_i \right), \quad r = 1, 2, \dots.$$

2.2. GPIGINAR(1)模型

定义 2.2. 基于二项稀疏算子“ $\circ$ ”, 以 GPIG 分布为新息项的一阶整值自回归 INAR(1)过程, 即有如下表示形式:

$$Y_t = \alpha \circ Y_{t-1} + \varepsilon_t, \quad t \in Z \quad (3)$$

其中 $0 \leq \alpha < 1$, ε_t 是 i.i.d 的 $GPIG(a, b, c)$ 随机变量, 与 $X_s (s < t)$ 独立。这个过程简称为 GPIGINAR(1)过程。

由模型的定义及概率论的基础知识可以容易得到该模型如下的一些基本性质:

- (1) X_t 的条件期望为: $E(X_t | X_{t-1}) = \alpha X_{t-1} + E(\varepsilon_t)$,
 - (2) X_t 的条件方差为: $Var(X_t | X_{t-1}) = \alpha(1-\alpha)X_{t-1} + Var(\varepsilon_t)$,
 - (3) X_t 的期望为: $E(X_t) = \frac{\mu_\varepsilon}{1-\alpha} = \frac{abc}{(1-\alpha)(1-c)^{1-a}}$,
 - (4) X_t 的方差为: $Var(X_t) = \frac{a\mu_\varepsilon + \sigma_\varepsilon^2}{1-\alpha^2} = \frac{\alpha abc}{(1-\alpha)^2(1-c)^{1-a}} + \frac{abc(1-ac)}{(1-\alpha)^2(1-c)^{2-a}}$,
 - (5) 自协方差函数与自相关函数为: $\gamma_k = Cov(X_t, X_{t-k}) = \alpha^k \gamma_0$, $\rho_k = \alpha^k$ 。
- 该模型的转移概率为:

$$p_{mj} = p(X_t = j | X_{t-1} = m) = \sum_{l=0}^{\min(j,m)} \binom{m}{l} \alpha^l (1-\alpha)^{m-l} p(\varepsilon_t = j-l) \quad (4)$$

其中:

$$p(\varepsilon_t = j-l) = \begin{cases} p_0 = e^{b[(1-c)^a - 1]}, & j-l=0, \\ p_1 = abcp_0, & j-l=1, \\ p_{k+1} \frac{1}{1+k} \left(abcp_k + \sum_{i=1}^k i r_{k+1-i} p_i \right), & j-l=k+1, \quad k \geq 1. \end{cases}$$

3. 参数估计

设序列 $X = (X_1, X_2, \dots, X_n)^T$ 是来自 GPIGINAR(1)模型的一组观测数据, 其中参数 $\theta = (a, b, c, \alpha)^T$ 为未知参数, 要进行贝叶斯估计, 首先写出似然函数为:

$$L(X|\theta) = \sum_{t=2}^n l(X|\theta) = \sum_{t=2}^n \log P(X_t = x_t | X_{t-1} = x_{t-1}) \quad (5)$$

令 $S_t = \alpha \circ X_{t-1}$, $t = 1, \dots, n$, 那么

$$\varepsilon_t = X_t - S_t \quad (6)$$

在给定 $(\alpha, a, b, c, X_{t-1})$ 的情况下, (S_t, ε_t) 是相互独立的。因此, 对于 $t \geq 1$ 有

$$P_{S_t, \varepsilon_t}(S_t, \varepsilon_t | X_{t-1}, \alpha, a, b, c) = \sum_{i=1}^n \binom{X_{t-1}}{S_t} \alpha^{S_t} (1-\alpha)^{X_{t-1}-S_t} P(\varepsilon_t = X_t - S_t) \quad (7)$$

其中 $S_t \leq X_{t-1}$, 新的观测值由 $X_t = S_t + \varepsilon_t$ 给出。令 $X = (X_0, X_1, \dots, X_n)$, $S_t = (S_1, S_2, \dots, S_n)$, $\varepsilon_t = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, 那么, 模型(3)的对数似然函数可以改写为

$$L_{S_t, \varepsilon_t}(S_t, \varepsilon_t | \alpha, a, b, c) = \log P_{S_t, \varepsilon_t}(S_t, \varepsilon_t | X_{t-1}, \alpha, a, b, c) \quad (8)$$

3.1. 先验

为了完成 GPIGNAR(1)过程的贝叶斯分析, 需要定义模型中未知参数的先验分布, 假设参数 θ 是相互独立的, 有

$$P(\theta) = P(a)P(b)P(c)P(\alpha) \quad (9)$$

本文为参数选取先验分布, 考虑以下先验设置:

$$P(a) = U(0, 1), \quad P(\alpha) = U(0, 1), \quad P(c) = U(0, 1), \quad P(b) = \text{Gamma}(m, n) \propto b^{m-1} e^{-nb}.$$

即为参数 a, c, α 选取无信息先验, 为参数 b 选取共轭伽马先验, 其中参数 m 与 n 都为超参数, 且它们都大于 0。结合似然和先验, 得到如下的联合后验密度函数:

$$\begin{aligned} f(S, \varepsilon, \alpha, a, b, c | X) &= \sum_{t=2}^n \log P_{S_t, \varepsilon_t}(S_t, \varepsilon_t | X_{t-1}, \alpha, a, b, c) P(a) P(b) P(c) P(\alpha) \\ &\propto \sum_{S_t=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{S_t} \alpha^{S_t} (1-\alpha)^{x_{t-1}-S_t} p(\varepsilon_t = x_t - S_t) b^{m-1} e^{-nb} \end{aligned} \quad (10)$$

在理论上, 可以对式(10)中的 θ 进行积分, 以获得参数的边际后验密度。然而这一过程复杂没有显示解。因此, 本文采用 Gibbs 抽样、随机游走链与 M-H 算法相结合的抽样方法来进行贝叶斯估计。为此推导出参数 α 与 S_t 的满条件分布, 其形式如下:

(1) 参数 α 的满条件分布

$$\begin{aligned} P(\alpha | S, X) &\propto \prod_{t=1}^n \alpha^{S_t} (1-\alpha)^{X_{t-1}-S_t} \\ &\propto \alpha^{\sum_{t=1}^n S_t} (1-\alpha)^{\sum_{t=1}^n X_{t-1}-S_t} \end{aligned} \quad (11)$$

由式(11)可知, α 的满条件分布为 $\text{Beta}\left(1 + \sum_{t=1}^n S_t, 1 + \sum_{t=1}^n X_{t-1} - S_t\right)$ 。

(2) 潜变量 S_t 的满条件分布, $t = 1, \dots, n$

$$P(S_t | \alpha, X) \propto \binom{X_{t-1}}{S_t} \alpha^{S_t} (1-\alpha)^{X_{t-1}-S_t} \quad (12)$$

由式(12)可知, S_t 的满条件分布为 $\text{Bin}(X_{t-1}, \alpha)$, $t = 1, \dots, n$ 。

3.2. 抽样方案

(1) 更新参数 α : 从 $\text{Beta}\left(1 + \sum_{t=1}^n S_t, 1 + \sum_{t=1}^n X_{t-1} - S_t\right)$ 中抽取 α 。

(2) 更新 (S, ε) , 采用 M-H 算法:

步骤 1: 从 $\text{Bin}(X_{t-1}, \alpha)$ 中抽取 S' ,

步骤 2: 若 $X_t < \varepsilon'_t$, 则重复步骤 1,

步骤 3: 令 $\varepsilon'_t = X_t - S_t$,

步骤 4: 计算接受概率: $P = \min \left\{ 1, \frac{f(\varepsilon'_t)}{f(\varepsilon_t)} \right\}$, 其中 ε'_t 是 ε_t 的后一次抽样值。

(3) 由于 ε_t 的后验分布没有明确的表达式, 故不能直接得出参数 (a, b, c) 的估计值, 本文采用了 Chen 和 So (2006) [11] 提出的自适应 MCMC 算法。该过程一共进行 N 次迭代, 其中包括在前 M 次迭代中采用随机游走 M-H 算法来抽取样本, 为燃烧期。之后的 $N-M$ 次迭代中采用独立核 M-H 算法来抽取参数样本。具体步骤如下:

步骤 1: 设置初始值 $\theta^{(0)} = (a^{(0)}, b^{(0)}, c^{(0)})^T$ 。

步骤 2: 当 $1 \leq h \leq M$ 时, 对于参数 $\theta^{(j)}$ 采用随机游走 M-H 算法, 具体过程如下所示:

步骤 2.1: 选取候选样本 $\theta^* = \{\theta_j^*\}$, $j=1, 2, \dots, 4$ 。其中 $\theta_j^* = \theta_j^{(h-1)} + m\varepsilon$, $\varepsilon \sim N(0, \sigma_j^2)$ 。其中 $\theta_j^{(h-1)}$ 表示第 $h-1$ 次迭代, m 表示步长。

步骤 2.2: 选取步骤 2.1 中 θ^* 满足 $0 < \theta_1 < 1$, $\theta_2 > 0$, $0 < \theta_3 < 1$ 条件的候选值。否则返回步骤 2.1。

步骤 2.3: 计算接受率

$$P(\theta^*, \theta_j^{(h-1)}) = \min \left\{ 1, \frac{f(\theta^*)}{f(\theta_j^{(h-1)})} \right\} \quad (13)$$

其中 $f(\cdot)$ 表示目标后验分布。

步骤 2.4: 从均匀分布中随机生成 U , $U \in [0, 1]$ 。若 $U < P(\theta^*, \theta_j^{(h-1)})$, 则接受新的候选值, 即 $\theta_j^* = \theta_j^{(h)}$ 。否则, 设置 $\theta_j^{(h)} = \theta_j^{(h-1)}$ 。

步骤 3: 当 $h \geq M+1$, 对于样本 $\theta_j^{(h)}$ 采用独立核 M-H 算法:

步骤 3.1: 从 $\theta_j^* = \mu_\theta + \varepsilon$ 中生成候选值 θ_j^* , 其中 $\varepsilon \sim N(0, \Omega_\theta)$ 。 μ_θ 和 Ω_θ 分别为样本均值和样本协方差矩阵, 其计算采用的是燃烧期前 M 次迭代的样本。

步骤 3.2: 选取步骤 3.1 中 θ_j^* 满足 $0 < \theta_1^* \leq 1$, $\theta_2^* > 0$, $0 \leq \theta_3^* < 1$ 条件的候选值。否则返回步骤 3.1。

步骤 3.3: 计算接受率

$$P(\theta_j^*, \theta_j^{(h-1)}) = \min \left\{ 1, \frac{f(\theta_j^*)k(\theta_j^{(h-1)})}{f(\theta_j^{(h-1)})k(\theta_j^*)} \right\} \quad (14)$$

其中 $\theta_j^{(h)}$ 表示 θ_j 的第 h 次迭代, $k(\theta)$ 表示是多元正态分布, 样本均值 μ_θ 和样本协方差 Ω_θ 采用前 M 次迭代 θ_j 的样本。

步骤 3.4: 从均匀分布中随机生成 U , $U \in [0, 1]$ 。若 $U < P(\theta_j^*, \theta_j^{(h-1)})$, 则接受新的候选值, 即 $\theta_j^* = \theta_j^{(h)}$ 。否则, 设置 $\theta_j^{(h)} = \theta_j^{(h-1)}$ 。

步骤 4: 如果链没有达到收敛, 则重复进行直至链收敛。

4. 数值模拟

为了检验 MCMC 算法对 GPIGINAR(1)模型的性能, 从 GPIGINAR(1)模型中生成不同样本量大小的样本, 在此数据下使用提出的 MCMC 算法对每个数据集进行拟合, 获得后验样本, 从而估计模型的参数。

先验分布为 3.1 节中所提到的, 计数时间序列 X_t 从 GPIGNAR(1) 模型中生成, 在接下来的数值模拟中, 对不同的样本量的数据执行 30000 次迭代, 前 15000 次迭代为燃烧期, 重复模拟 1000 次, 获得结果。

在本节中, 考虑以下方案设置:

方案 1: $(a, b, c, \alpha) = (0.5, 3, 0.55, 0.1)$;

方案 2: $(a, b, c, \alpha) = (0.55, 2, 0.6, 0.4)$ 。

如图 1 所示, 展示了方案 1 在样本量为 300 时序列的样本路径图、直方图、ACF 图及 PACF 图。可以看出, 序列为平稳序列并且存在短期自相关。可以看出, 序列为平稳序列并且存在短期自相关。

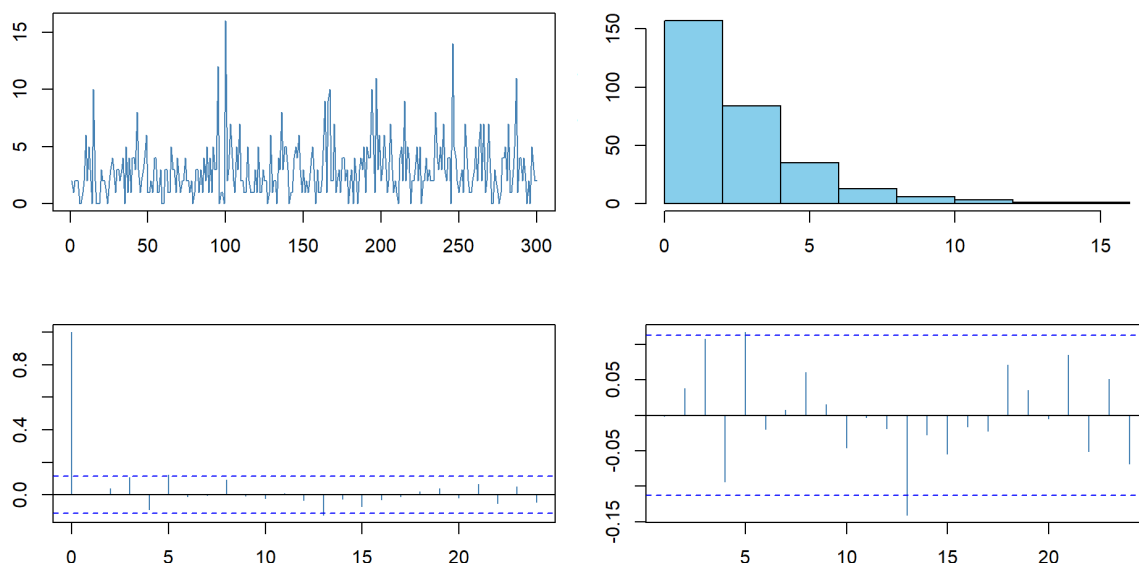


Figure 1. Sequence, histogram, ACF and PACF plots with $n = 300$

图 1. 样本量 $n = 300$ 时序列的样本路径图、直方图、ACF 图及 PACF 图

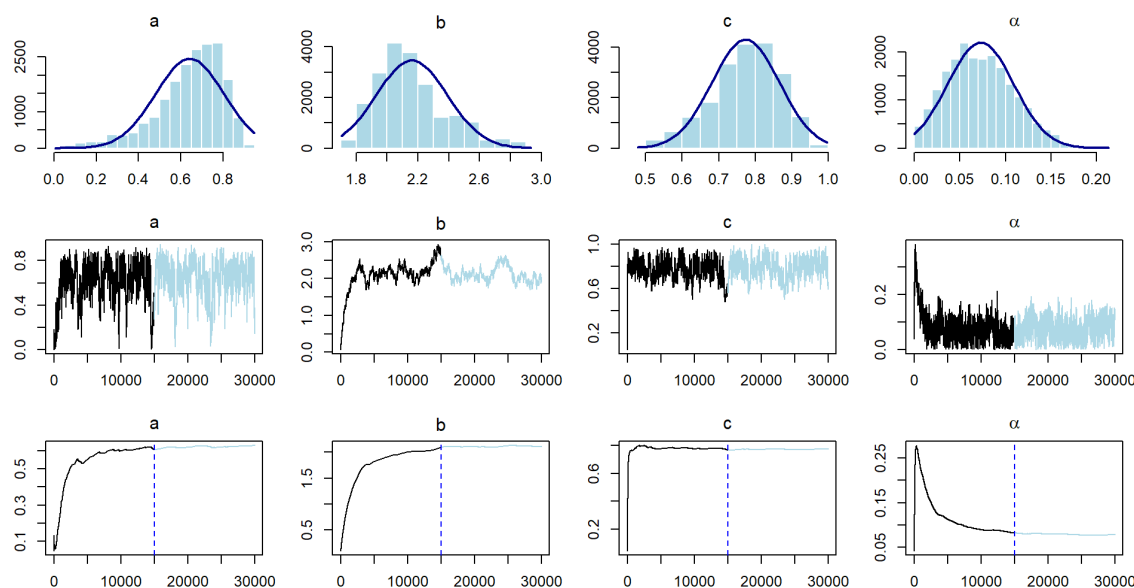


Figure 2. Convergence plots, histogram plots and ergodic mean plots for each parameter with sample size of 300

图 2. 每个参数在样本量为 300 时的链轨迹图、后验分布直方图以及遍历均值图

在接下来的数值模拟中通过对不同样本量的样本分别进行 30,000 次迭代, 丢弃前 15,000 次, 重复模拟进行 1000 次得到其结果。为了展示 3.2 节中抽样方案的收敛情况, 画出每个参数在样本量为 300 时的链轨迹图、后验分布直方图以及遍历均值图, 如图 2 所示, 可以很明显地看到其后验样本的路径没有明显的波动, 后验样本的遍历均值轨迹在 15,000 次迭代后趋于稳定, 表明后验样本的平稳性, 后验样本的直方图清楚地说明各参数后验样本的分布情况。总而言之, 图 2 提供证据表明 MCMC 算法的收敛性, 从而也进一步表明所提出的贝叶斯估计的可行性。

为了展现两种估计效果的好坏, 本节用 Bias 和 MSE 两种度量指标来衡量。为了与此方法所得出的估计结果进行比较, 采用极大似然估计方法进行模拟, 其结果如表 1 与表 2 所示, 表 1 与表 2 中分别为方案 1 与方案 2 中重复模拟不同样本量的样本 1000 次后, 得到的贝叶斯方法与极大似然方法的比较结果, 其中包括贝叶斯方法所得到的后验均值(MC Mean)、后验中位数(MC Median)、后验标准差(MC SD)、95%置信区间、Bias、MSE, 极大似然方法所得到的 Bias、MSE。

根据表 1 及表 2 的结果可知:

- (1) 随着样本量 n 从 80 增大到 700, 贝叶斯估计中多数参数的均值、中位数与真实值的 Bias 逐渐减小, 标准差也在降低, 表明估计值更集中于真实值。极大似然估计下部分参数表现类似, 但参数 b 的变化规律较复杂, 其估计值与真实值的偏差并非随样本量单调变化。
- (2) 贝叶斯估计下, 多数参数的 Bias 绝对值在不同样本量时相对较小; 极大似然估计中, 部分参数如 b 的 Bias 绝对值较大。多数情况下, 贝叶斯估计的 MSE 小于极大似然估计, 说明贝叶斯估计的准确性更高。例如参数 a , c , α 在各样本量下, 贝叶斯估计的 MSE 普遍更低。贝叶斯估计的 $Q_{0.025}$ 和 $Q_{0.975}$ 所表示的置信区间, 以及标准差, 在多数参数和样本量下相对更优, 显示出估计值的离散程度和不确定性控制得更好。
- (3) 综合来看, 采用贝叶斯方法估计该模型更精确一点。

Table 1. Simulation results
表 1. 模拟结果

参数	真实值	贝叶斯估计						极大似然估计		
		MC Mean	MC Me- dian	MC SD	$\mathcal{Q}_{0.025}$	$\mathcal{Q}_{0.975}$	Bias	MSE	Bias	MSE
$n = 80$										
a	0.5	0.4723	0.4742	0.2458	0.0465	0.9006	-0.0277	0.0092	-0.1252	0.1645
b	3	3.0130	3.0178	0.5926	1.9375	4.1455	0.0130	1.3423	1.2168	12.2101
c	0.55	0.5575	0.5452	0.1124	0.3797	0.7983	0.0075	0.0112	-0.0261	0.0488
α	0.1	0.1383	0.1326	0.0700	0.0235	0.2860	0.0383	0.0051	0.0008	0.0037
$n = 150$										
a	0.5	0.4768	0.4794	0.2192	0.0921	0.8589	-0.0232	0.0220	-0.1264	0.1432
b	3	2.6291	2.6290	0.4246	1.8974	3.4110	-0.3709	0.6300	0.6017	2.7939
c	0.55	0.5128	0.5003	0.0982	0.3614	0.7222	-0.0372	0.0150	-0.0247	0.0397
α	0.1	0.1160	0.1121	0.0556	0.0226	0.2327	0.0160	0.0026	-0.0052	0.0045
$n = 300$										
a	0.5	0.4804	0.4929	0.2214	0.0595	0.8574	-0.0196	0.0069	-0.0335	0.1101

续表

<i>b</i>	3	3.2591	3.2464	0.3390	2.6545	3.8989	0.2591	0.6603	0.4759	2.6815
<i>c</i>	0.55	0.5734	0.5632	0.0868	0.4293	0.7587	0.0234	0.0158	−0.0225	0.0287
α	0.1	0.1044	0.1024	0.0438	0.0267	0.1942	0.0044	0.0016	−0.0040	0.0025
<i>n</i> = 700										
<i>a</i>	0.5	0.4877	0.4965	0.1890	0.1062	0.8184	−0.0123	0.0255	−0.0882	0.0820
<i>b</i>	3	3.1331	3.1470	0.4017	2.4321	3.8330	0.1331	0.2805	0.2661	0.6620
<i>c</i>	0.55	0.5169	0.5075	0.0682	0.4112	0.6718	−0.0331	0.0035	−0.0173	0.0155
α	0.1	0.0988	0.0984	0.0323	0.0379	0.1626	−0.0012	0.0008	−0.0021	0.0010

Table 2. Simulation results
表 2. 模拟结果

参数	真实值	贝叶斯估计					极大似然估计			
		MC Mean	MC Me- dian	MC SD	$\mathcal{Q}_{0.025}$	$\mathcal{Q}_{0.975}$	Bias	MSE	Bias	MSE
<i>n</i> = 80										
<i>a</i>	0.55	0.4459	0.4458	0.2084	0.0879	0.8133	−0.1041	0.0288	−0.1749	0.1457
<i>b</i>	2	2.6114	2.5424	0.6506	1.5788	3.9098	0.6114	0.8749	0.6475	2.5182
<i>c</i>	0.6	0.5098	0.4919	0.1341	0.3075	0.7940	−0.0902	0.0191	−0.0709	0.0533
α	0.4	0.3992	0.3990	0.0552	0.2937	0.5075	−0.0008	0.0037	−0.0079	0.0036
<i>n</i> = 150										
<i>a</i>	0.55	0.5057	0.5140	0.2357	0.0881	0.9013	−0.0443	0.0348	−0.1602	0.1297
<i>b</i>	2	2.5319	2.4617	0.6610	1.5058	3.9007	0.5319	0.6970	0.5649	0.5069
<i>c</i>	0.6	0.5195	0.5049	0.1110	0.3509	0.7512	−0.0805	0.0149	−0.0606	0.0389
α	0.4	0.3952	0.3942	0.0947	0.2157	0.5833	−0.0048	0.0110	−0.0240	0.0117
<i>n</i> = 300										
<i>a</i>	0.55	0.5231	0.5342	0.2298	0.1078	0.8998	−0.0269	0.0293	−0.1473	0.1140
<i>b</i>	2	2.5127	2.4588	0.5293	1.6697	3.6001	0.5127	0.6531	0.4822	1.1736
<i>c</i>	0.6	0.5350	0.5238	0.1015	0.3789	0.7410	−0.0650	0.0128	−0.0349	0.0274
α	0.4	0.3996	0.3994	0.0364	0.3298	0.4706	−0.0004	0.0014	−0.0041	0.0020
<i>n</i> = 700										
<i>a</i>	0.55	0.5352	0.5061	0.0949	0.3861	0.7249	−0.0148	0.0184	−0.1197	0.1031
<i>b</i>	2	2.4024	2.3495	0.4620	1.6729	3.3589	0.4024	0.3889	0.4058	0.9847
<i>c</i>	0.6	0.5353	0.5261	0.0949	0.3861	0.7249	−0.0647	0.0117	−0.0350	0.0231
α	0.4	0.3999	0.3996	0.4291	0.3176	0.4837	−0.0004	0.0020	−0.0034	0.0014

5. 实例分析

通过对 1877 年 12 月 1 日至 1879 年 4 月 15 日期间每周太阳黑子群的数量数据应用说明了所提出的

方法, 采用雅典的 SCHMIDT 独立观测的数据, 其太阳黑子群的数量数据由 91 个观测值组成。样本均值和样本方差分别为 1.154 和 4.651, 表现出相当大的过分散。图 3 中描述太阳黑子群的数量数据的序列图、直方图、ACF 图和 PACF 图。其中 ACF 图与 PACF 图表明数据存在显著的自相关。且从计数序列图与序列直方图中可以看出数据具有拖尾性质。数据可由美国国家地理数据中心 (<http://www.ncei.noaa.gov/>) 获得。

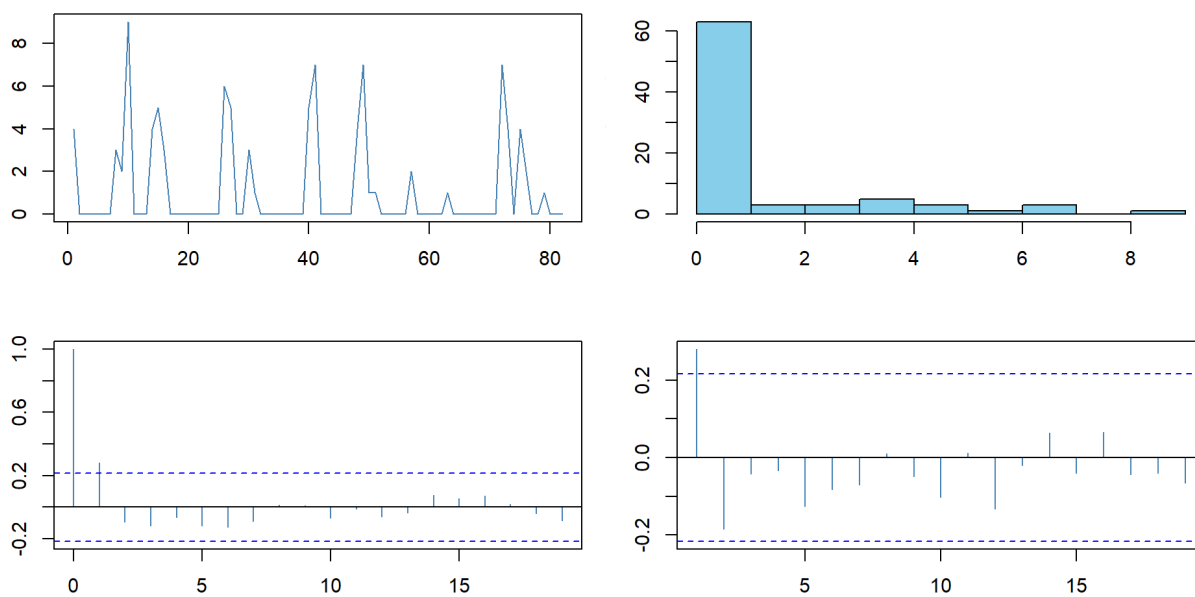


Figure 3. Sequence, histogram, ACF and PACF plots of the sunspot subgroups data
图 3. 太阳黑子群的数量数据的序列图、直方图、ACF 图与 PACF 图

本文采用以下三个模型拟合该数据, 其拟合结果进行比较:

- (1) PINAR(1)模型[30];
- (2) PIGINAR(1)模型[6];
- (3) GPIGINAR(1)模型[31]。

本文运用 MCMC 算法对三个模型进行贝叶斯估计, 经 30,000 次迭代后, 舍弃前 15,000 次, 进而得到最终估计结果。表 3 中为 GPIGINAR(1)模型与其他模型比较的结果, 包括参数估计结果、期望、方差、 I_d 、AIC 与 BIC 的值。从表 3 结果可知, PINAR(1)模型表现欠佳, 离散度最低。PINAR(1)模型和 GPIGINAR(1)模型的均值与样本均值相近, 然而, 仅 GPIGINAR(1)模型的方差与样本方差接近, 且其离散度也最接近样本离散度。在所有模型中, GPIGINAR(1)模型的 AIC 和 BIC 值最小。综合各项指标, GPIGINAR(1)模型最适宜模拟该数据。

为了检验所讨论拟合的模型是否真的适合分析的数据。检验模型充分性最常见的方法是比较拟合模型的均值和离散特性, 采用标准的皮尔逊残差检验(Jung 等(2006) [32]):

$$Z_t = \frac{Y_t - E(Y_t | Y_{t-1})}{\sqrt{\text{Var}(Y_t | Y_{t-1})}} \quad (15)$$

根据 GPIGINAR(1)模型的条件期望与条件方差公式, 计算了在贝叶斯估计方法下残差的均值和方差, 认为拟合残差的均值和方差越接近 0 和 1, 模型的拟合效果越好。在本节的实际数据中, 所拟合的贝叶斯

方法 GPIGINAR(1)模型残差的均值和方差分别为-0.0013 和 1.0111,该结果也再次说明应用 GPIGINAR(1)模型拟合该数据是合适的。

Table 3. Comparison results
表 3. 比较结果

模型	参数	估计值	期望	方差	I_d	AIC	BIC
PINAR (1)			1.0608	1.0608	1	308.8153	313.6287
	$\hat{\alpha}$	0.210					
	$\hat{\lambda}$	0.838					
PIGINAR (1)			0.951	11.8474	12.4578	299.0965	306.3167
	$\hat{\alpha}$	0.109					
	$\hat{\mu}$	0.951					
	$\hat{\phi}$	0.083					
GPIGINAR (1)			1.0597	5.4371	5.1308	194.9537	203.609
	$\hat{\alpha}$	0.1064					
	$\hat{a}$	0.1841					
	$\hat{b}$	0.8633					
	$\hat{c}$	0.1122					

为了形象的展示残差的性质，基于拟合的 GPIGINAR (1)模型绘制了残差序列图、直方图、ACF 及 PACF 图，如图 4 所示。为了进一步验证残差序列是一个白噪声序列，采用 Ljung-Box (LB)检验的方法对残差 Z_t 进行检验，对其延迟从 1 到 12，对于每一个延迟，计算 LB 统计量和其 p 值，其结果在表 4 中，可以看到其 p 值均大于 0.05，从而说明该残差序列是一个白噪声序列。进一步说明 GPIGINAR(1)模型拟合这组数据是合适的。

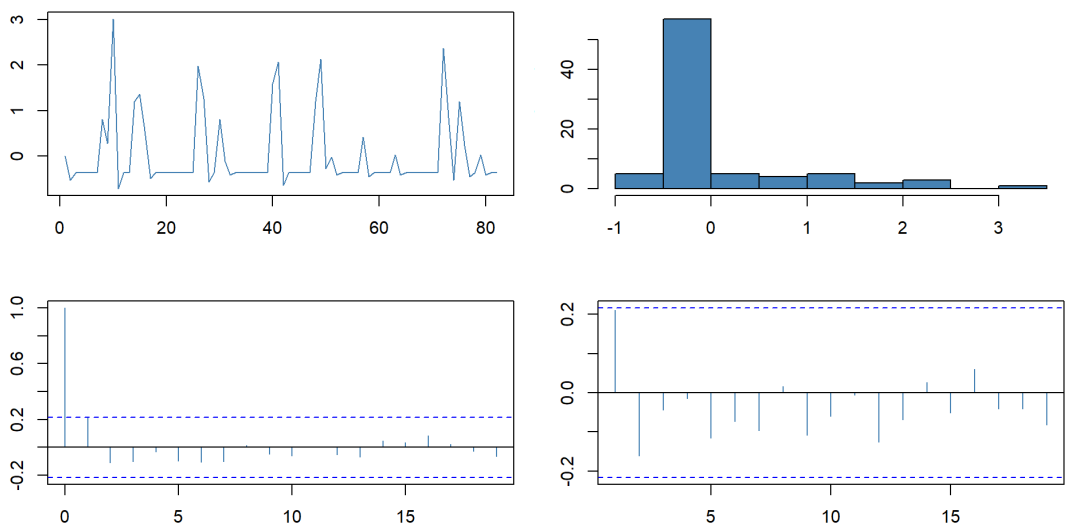


Figure 4. Residual sequence, histograms, ACF and PACF plots
图 4. 残差序列图、直方图、ACF 及 PACF 图

Table 4. LB test results
表 4. LB 检验结果

延迟阶数	$\{\hat{Z}_i\}$		$\{\hat{Z}_i^2\}$	
	LB 统计量	p 值	LB 统计量	p 值
1	0.5057	0.4769	3.8224	0.0506
2	0.9343	0.6268	4.8255	0.0896
3	1.2958	0.7301	5.7183	0.1262
4	1.4084	0.8427	5.7979	0.2148
5	1.4601	0.9176	6.6746	0.2460
6	2.4681	0.8720	7.6963	0.2612
7	3.0047	0.8846	8.6061	0.2822
8	3.0646	0.9302	8.6208	0.3753
9	3.4702	0.9427	8.8396	0.4522
10	4.2350	0.9361	9.1885	0.5143
11	4.2904	0.9606	9.1936	0.6040
12	4.3733	0.9757	9.4592	0.6633

为了说明 GPIGNAR(1)模型的马尔可夫链收敛, 如图 5 所示, 通过拟合模型的每个参数的轨迹图、直方图以及累积均值图, 研究了马尔可夫链的收敛诊断。通过观察轨迹图, 没有明显的上升或下降趋势, 也不存在周期性等规律变化, 从整体上可以看到每个参数随着时间的推移逐渐收敛到稳定状态, 认为马尔可夫链已经收敛。通过观察直方图, 可以知道每个参数都已经收敛到目标分布, 认为其马尔可夫链已经收敛。通过观察累积均值图, 可以看到曲线在某个值附近波动较小, 说明参数的均值已经稳定, 暗示 MCMC 算法已经收敛。

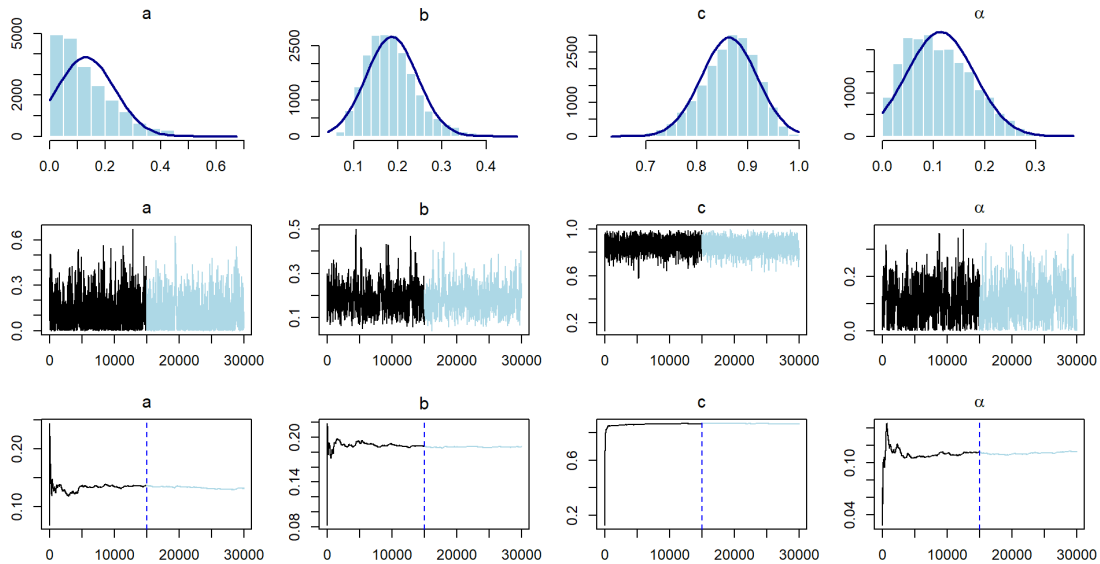


Figure 5. Fitting data convergence plots, histogram plots and ergodic mean plots
图 5. 拟合数据链收敛图、后验密度直方图、累计均值图

6. 结论

在对 GPIGNAR(1)模型进行研究时, 基于 MCMC 算法开展贝叶斯估计工作。具体而言, 通过在似然函数中巧妙引入潜变量, 成功得到完全数据似然, 进而在此基础上推导出参数的满条件分布。对于那些具有复杂后验分布的参数, 采用随机游走加独立核的方法进行采样操作, 以此获得参数的后验样本。随后, 通过数值模拟的方式, 将该方法与极大似然估计方法进行对比, 结果充分验证了贝叶斯方法的优越性。此外, 以雅典 SCHMIDT 观测的太阳黑子群数据作为实例, 详细介绍了该模型在实际中的应用过程。

基金项目

辽宁科技大学博士启动资金(6003000310)。

参考文献

- [1] Maiti, R., Biswas, A. and Chakraborty, B. (2017) Modelling of Low Count Heavy Tailed Time Series Data Consisting Large Number of Zeros and Ones. *Statistical Methods & Applications*, **27**, 407-435. <https://doi.org/10.1007/s10260-017-0413-z>
- [2] Holla, M.S. (1967) On a Poisson-Inverse Gaussian Distribution. *Metrika*, **11**, 115-121. <https://doi.org/10.1007/bf02613581>
- [3] Sichel, H.S. (1982) Repeat-Buying and the Generalized Inverse Gaussian-Poisson Distribution. *Applied Statistics*, **31**, 193-204. <https://doi.org/10.2307/2347993>
- [4] Tweedie, M.C., et al. (1984) An Index Which Distinguishes between Some Important Exponential Families. In: *Statistics: Applications and New Directions: Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, Indian Statistical Institute, 579-604.
- [5] Shoukri, M.M., Asyali, M.H., VanDorp, R. and Kelton, D. (2021) The Poisson Inverse Gaussian Regression Model in the Analysis of Clustered Counts Data. *Journal of Data Science*, **2**, 17-32. [https://doi.org/10.6339/jds.2004.02\(1\).135](https://doi.org/10.6339/jds.2004.02(1).135)
- [6] Bourguignon, M., Rodrigues, J. and Santos-Neto, M. (2018) Extended Poisson INAR(1) Processes with Equidispersion, Under-Dispersion and Overdispersion. *Journal of Applied Statistics*, **46**, 101-118. <https://doi.org/10.1080/02664763.2018.1458216>
- [7] 钱连勇. 若干重尾整数时间序列模型的研究[D]: [博士学位论文]. 长春: 吉林大学, 2021.
- [8] Bolstad, W.M. (2007) Introduction to Bayesian Statistics. Wiley. <https://doi.org/10.1002/9780470181188>
- [9] Hay, J.L. (2001) Bayesian Analysis of a Time Series of Counts with Covariates: An Application to the Control of an Infectious Disease. *Biostatistics*, **2**, 433-444. <https://doi.org/10.1093/biostatistics/2.4.433>
- [10] Chen, C.W.S. and Lee, J.C. (1995) Bayesian Inference of Threshold Autoregressive Models. *Journal of Time Series Analysis*, **16**, 483-492. <https://doi.org/10.1111/j.1467-9892.1995.tb00248.x>
- [11] Chen, C.W.S. and So, M.K.P. (2006) On a Threshold Heteroscedastic Model. *International Journal of Forecasting*, **22**, 73-89. <https://doi.org/10.1016/j.ijforecast.2005.08.001>
- [12] Chen, C.W.S., Khamthong, K. and Lee, S. (2019) Markov Switching Integer-Valued Generalized Auto-Regressive Conditional Heteroscedastic Models for Dengue Counts. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **68**, 963-983. <https://doi.org/10.1111/rssc.12344>
- [13] Ausin, M.C. and Galeano, P. (2007) Bayesian Estimation of the Gaussian Mixture GARCH Model. *Computational Statistics & Data Analysis*, **51**, 2636-2652. <https://doi.org/10.1016/j.csda.2006.01.006>
- [14] Bu, R. and McCabe, B. (2008) Model Selection, Estimation and Forecasting in Inar(p) Models: A Likelihood-Based Markov Chain Approach. *International Journal of Forecasting*, **24**, 151-162. <https://doi.org/10.1016/j.ijforecast.2007.11.002>
- [15] 朱复康, 李琦. INGARCH(1,1)模型参数的矩估计和 Bayes 估计[J]. 吉林大学学报(理学版), 2009, 47(5): 899-902.
- [16] Neal, P. and Subba Rao, T. (2007) MCMC for Integer-Valued ARMA Processes. *Journal of Time Series Analysis*, **28**, 92-110. <https://doi.org/10.1111/j.1467-9892.2006.00500.x>
- [17] 张哲, 张海祥, 张卓飞, 等. INAR(1)模型参数的 Bayes 估计[J]. 吉林大学学报(理学版), 2010, 48(6): 931-935.
- [18] Drovandi, C.C., Pettitt, A.N. and McCutchan, R.A. (2016) Exact and Approximate Bayesian Inference for Low Integer-

- Valued Time Series Models with Intractable Likelihoods. *Bayesian Analysis*, **11**, 325-352. <https://doi.org/10.1214/15-ba950>
- [19] Truong, B., Chen, C.W. and Sriboonchitta, S. (2017) Hysteretic Poisson INGARCH Model for Integer-Valued Time Series. *Statistical Modelling*, **17**, 401-422. <https://doi.org/10.1177/1471082x17703855>
- [20] Chen, C.W.S. and Lee, S. (2016) Bayesian Causality Test for Integer-Valued Time Series Models with Applications to Climate and Crime Data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **66**, 797-814. <https://doi.org/10.1111/rssc.12200>
- [21] Chu, Y. and Yu, K. (2023) Bayesian Log-Linear Beta-Negative Binomial Integer-Valued Garch Model. *Computational Statistics*, **39**, 1183-1202. <https://doi.org/10.1007/s00180-023-01386-w>
- [22] 王佩. 门限 ARMA 模型的贝叶斯估计[D]: [硕士学位论文]. 北京: 中国矿业大学, 2019.
- [23] Li, H., Yang, K. and Wang, D. (2018) A Threshold Stochastic Volatility Model with Explanatory Variables. *Statistica Neerlandica*, **73**, 118-138. <https://doi.org/10.1111/stan.12143>
- [24] Garay, A.M., Medina, F.L., Cabral, C.R.B. and Lin, T. (2020) Bayesian Analysis of the p-Order Integer-Valued AR Process with Zero-Inflated Poisson Innovations. *Journal of Statistical Computation and Simulation*, **90**, 1943-1964. <https://doi.org/10.1080/00949655.2020.1754819>
- [25] 黄菊. 空间自回归条件异方差模型的贝叶斯估计[D]: [硕士学位论文]. 长春: 吉林大学, 2021.
- [26] Yang, K., Yu, X., Zhang, Q. and Dong, X. (2022) On MCMC Sampling in Self-Exciting Integer-Valued Threshold Time Series Models. *Computational Statistics & Data Analysis*, **169**, Article 107410. <https://doi.org/10.1016/j.csda.2021.107410>
- [27] 于鑫洋. 自激励整数值门限自回归模型的贝叶斯估计及应用[D]: [硕士学位论文]. 长春: 长春工业大学, 2022.
- [28] Kang, Y., Lu, F. and Wang, S. (2023) Bayesian Analysis for an Improved Mixture Binomial Autoregressive Model with Applications to Rainy-Days and Air Quality Level Data. *Stochastic Environmental Research and Risk Assessment*, **38**, 1313-1333. <https://doi.org/10.1007/s00477-023-02633-8>
- [29] Zhu, R. and Joe, H. (2009) Modelling Heavy-Tailed Count Data Using a Generalised Poisson-Inverse Gaussian Family. *Statistics & Probability Letters*, **79**, 1695-1703. <https://doi.org/10.1016/j.spl.2009.04.011>
- [30] Al-Osh, M.A. and Alzaid, A.A. (1987) First-Order Integer-Valued Autoregressive (INAR(1)) Process. *Journal of Time Series Analysis*, **8**, 261-275. <https://doi.org/10.1111/j.1467-9892.1987.tb00438.x>
- [31] Qian, L., Li, Q. and Zhu, F. (2020) Modelling Heavy-Tailedness in Count Time Series. *Applied Mathematical Modelling*, **82**, 766-784. <https://doi.org/10.1016/j.apm.2020.02.001>
- [32] Jung, R.C., Kukuk, M. and Liesenfeld, R. (2006) Time Series of Count Data: Modeling, Estimation and Diagnostics. *Computational Statistics & Data Analysis*, **51**, 2350-2364. <https://doi.org/10.1016/j.csda.2006.08.001>