Advances in Applied Mathematics 应用数学进展, 2026, 15(5), 511-522
Published Online May 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.155247

Hans汉斯

基于NGINAR(p)模型的贝叶斯非参数预测

姜　姝，卢飞龙

辽宁科技大学理学院，辽宁　鞍山

收稿日期：2026年4月21日；录用日期：2026年5月15日；发布日期：2026年5月23日

摘　要

本文将Bisaglia和Canale (2016)提出的贝叶斯非参数预测框架从INAR模型拓展至基于负二项稀疏算子的广义p阶整数值自回归模型(NGINAR(p))。该拓展利用负二项稀疏算子刻画高阶依赖结构与过度离散特征，并采用舍入高斯混合先验对创新项分布进行非参数建模，从而灵活捕捉计数数据中普遍存在的多峰、偏态等复杂分布形态。针对高阶模型的结构复杂性，构建了基于数据增强的Gibbs采样算法，实现自回归系数与创新项分布的联合后验推断，并直接获得整数值的*h*步向前预测分布。模拟研究与中国某地区月度总云量数据的实证分析表明，该方法在不同样本量、自回归系数及创新项分布设定下均表现出良好的预测精度与稳定性，在处理高阶复杂计数时间序列时具有显著的适应性与有效性。

关键词

整数值时间序列，NGINAR(p)模型，贝叶斯非参数，Dirichlet过程混合，Gibbs采样

Bayesian Nonparametric Forecasting for NGINAR(p) Models

Shu Jiang, Feilong Lu

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

Received: April 21, 2026; accepted: May 15, 2026; published: May 23, 2026

Abstract

This paper extends the Bayesian nonparametric prediction framework proposed by Bisaglia and Canale (2016) from the INAR model to the generalized p-order integer-valued autoregressive model based on the negative binomial thinning operator (NGINAR(p)). This extension utilizes the negative binomial thinning operator to describe the high-order dependency structure and overdispersion characteristics, and employs a rounded Gaussian mixture prior for nonparametric modeling of the innovation term distribution, thereby flexibly capturing complex distributional forms such as multimodality

文章引用: 姜姝, 卢飞龙. 基于 NGINAR(p)模型的贝叶斯非参数预测[J]. 应用数学进展, 2026, 15(5): 511-522.
DOI: 10.12677/aam.2026.155247

and skewness commonly found in count data. To address the structural complexity of high-order models, a data augmentation-based Gibbs sampling algorithm is constructed to jointly infer the posterior distribution of the autoregressive coefficients and the innovation term distribution, and directly obtain the *h*-step ahead predictive distribution of integer values. Simulation studies and empirical analysis of monthly total cloud cover data from a certain region in China demonstrate that this method exhibits excellent prediction accuracy and stability under different sample sizes, autoregressive coefficient settings, and innovation term distribution assumptions, and shows significant adaptability and effectiveness in handling high-order complex count time series.

Keywords

Integer-Valued Time Series, NGINAR(p) Model, Bayesian Nonparametric, Dirichlet Process Mixture, Gibbs Sampling

1. 引言

整数值时间序列广泛存在于保险精算、流行病学、网络通信等诸多领域，记录了特定时段内事件的发生次数，如月度交通事故数、传染病日新增病例数及网站日访问量等。这类数据具有非负性、整数性和离散性的本质特征，使得传统的连续值时间序列模型在直接应用时面临根本性困难，不仅可能产生无意义的非整数预测值，其连续分布假设也与数据的离散结构不匹配。因此，针对计数数据发展专门的建模与预测方法，具有重要的理论意义与应用价值。

为应对这一挑战，稀疏算子的引入成为关键。Steutel 和 Van Harn [1]提出的二项稀疏算子，为整数值时间序列建模奠定了数学基础。Al-Osh 和 Alzaid [2]与 McKenzie [3]在此基础上正式提出基于二项稀疏算子的一阶整数值自回归 INAR(1)模型，成功将传统 AR(1)模型的结构移植至整数域，开创了整数值时间序列建模的新方向。此后，研究者开展了系统性拓展：Du 和 Li [4]建立了高阶 INAR(p)模型的理论框架；Al-Osh 和 Alzaid [5]研究了整数值移动平均模型，逐步丰富了模型体系。另一方面，Ferland 等[6]提出的 INGARCH 模型及其后续发展，通过对条件均值和方差进行联合建模，有效捕捉了计数数据的过度离散性和波动集聚性；Fokianos 等[7]进一步在广义线性模型框架下构建了泊松自回归模型，实现了对协变量的灵活处理，极大拓展了模型的应用范围。然而，正如 Weiß [8]在其专著中指出的，尽管上述模型构成了该领域的主流参数化框架，但如何灵活处理创新项复杂分布的问题，仍是尚未解决的核心挑战。

随着研究的深入，传统基于二项稀疏算子的 INAR 类模型在应对过度离散等复杂数据特征时显现出一定的理论局限。为克服这一不足，Ristić [9]等提出了基于负二项稀疏算子的 NGINAR(1)模型。该算子的独特之处在于，允许一个历史个体在下一期产生多个“后代”，从而天然契合了过度离散数据的生成机制。在此基础上，Nastić [10]等将其扩展至高阶 CGINAR(p)模型，进一步增强了对复杂依赖结构的刻画能力。尽管 NGINAR 类模型在结构上不断演进，但其创新项的建模仍普遍依赖于参数化假设(如负二项、几何分布等)，当实际数据呈现多峰、零膨胀或偏态等更为复杂的分布形态时，其预测精度与适应性仍面临严峻挑战。

为突破这一局限，将贝叶斯非参数方法引入 NGINAR 模型体系成为自然的选择。基于上述发展脉络，本文将 Bisaglia 和 Canale [11]提出的贝叶斯非参数预测框架从一阶模型拓展至基于负二项稀疏算子的广

义 p 阶整数值自回归模型(NGINAR(p))。这一拓展融合了 NGINAR 模型通过负二项稀疏算子刻画过度离散与高阶依赖的结构优势，以及 Canale 和 Dunson [12]提出的舍入高斯混合先验对创新项分布进行非参数建模的灵活性，从而突破了传统参数化假设的局限。通过构建基于数据增强的 Gibbs 采样算法，实现自回归系数与创新项分布的联合后验推断，并直接获得整数值的 h 步向前预测分布，为高阶复杂计数时间序列的精确预测提供了一种兼具理论严谨性与实践适应性的新方法。

本文后续部分组织如下：第 2 节详细阐述模型设定，包括 NGINAR(p)过程及创新项的舍入高斯混合先验；第 3 节介绍贝叶斯采样算法；第 4 节通过数值模拟评估方法的有限样本表现；第 5 节基于实际计数时间序列数据进行实证分析；第 6 节对全文进行总结与展望。

2. 模型设定

2.1. NGINAR(p)模型

NGINAR(p)模型是 NGINAR(1)的多阶滞后扩展，用于建模整数值的时间序列数据。该模型的形式为：

$$Y_t = \sum_{i=1}^{p} \alpha_i * Y_{t-i} + \varepsilon_t \tag{1}$$

其中，“$*$”表示负二项稀疏(Negative Binomial Thinning)算子，该算子的定义依赖于几何分布，对于任意非负整数随机变量 Y 和参数 $\alpha_i \in [0,1)$，有 $\alpha_i * Y_{t-i} = \sum_{j=1}^{Y_{t-i}} W_j^{(i)}$，其中 $\{W_j^{(i)}\}$ 为独立同分布的几何随机变量，其成功概率为 $1/(1+\alpha_i)$。

该算子是整数自回归模型的核心，通过一种特定的随机过程将当前值与滞后项结合，从而生成新的值。模型中，$\alpha_i \in [0,1), (i=1,2,\cdots,p)$ 自回归系数，反映各滞后项对当前值的影响强度；ε_t 为非负整数值创新项，独立同分布且与历史观测独立。与经典 NGINAR(1)不同，本文不预设 ε_t 的具体分布形式，允许其根据实际数据特征选择，如广义泊松分布、负二项分布或混合分布等，从而更灵活地适应过度离散、零膨胀等不同类型的计数数据。该稀疏机制的直观含义与二项稀疏算子存在本质区别：每个历史个体不仅可能“存活”，还可能产生多个“后代”，且每个个体产生的后代数量服从几何分布。这一增殖机制，结合多阶滞后项的引入，使模型能够同时捕捉计数数据中的过度离散特征和复杂的高阶依赖结构。传统的 NGINAR(p)模型通常对创新项分布施加参数化假设，当实际数据呈现多峰、零膨胀或偏态等复杂分布时，这种设定缺乏灵活性。为此，本文引入贝叶斯非参数方法，借鉴 Bisaglia 和 Canale [9]的策略，采用含入高斯混合先验对创新项分布进行灵活建模[13] [14]。

尽管在该设定下创新项分布无法获得显式表达式，但由于观测值可表示为各滞后保留项与创新项之和，其条件分布仍可通过卷积形式表达[15] [16]。对于 NGINAR(p)模型，其条件概率结构可表示为：

$$\Pr(Y_t = y_t \mid y_{t-1}, \alpha, p) = \sum_{s=0}^{\min(y_t, y_{t-1})} \Pr(B_t = s \mid y_{t-1}, \alpha) \cdot p(y_t - s) \tag{2}$$

其中 $B_t \sim NB(r = y_{t-1}, p = 1/1+\alpha)$，$\Pr(B_t = s \mid y_{t-1}, \alpha)$ 表示基于负二项稀疏算子的个体保留概率，$p(y_t - s)$ 为创新项在新增个体数为 $y_t - s$ 条件下的概率质量函数。

给定样本观测序列 $\boldsymbol{y} = (y_1, \cdots, y_T)$，模型参数为参 $\alpha \in [0,1)$ 数 $\theta = (\alpha, p)$，其中为稀疏强度参数，p 是创新项 ε_t 的概率质量函数，在不对 p 进行具体分布假设的前提下，仍可通过边际化的方法表达观测值的条件概率。基于上述条件概率结构，可以进一步构建观测数据的似然函数形式如下：

$$L(\theta \mid \boldsymbol{y}) = \prod_{t=2}^{T} \sum_{s=0}^{\min(y_t, y_{t-1})} \Pr(B_t = s \mid y_{t-1}, \alpha) \cdot p(y_t - s) \tag{3}$$

由于创新项分布 p，通过贝叶斯非参数方法建模得到，其形式并不显式，从而导致预测分布难以解析表达。为此，本文在贝叶斯框架下采用后验预测分布对未来 h 步进行预测，其数学表达为：

$$\Pr(Y_{T+h} = j \mid \boldsymbol{y}) = \int \Pr(Y_{T+h} = j \mid \boldsymbol{y}, \theta) \mathrm{d}\pi(\theta \mid \boldsymbol{y}) \tag{4}$$

2.2. 创新项的非参数建模

针对 NGINAR(1)模型中创新项 ε_t 的分布建模问题，本文借鉴了 Canale 与 Dunson [12]提出的截断高斯混合(Rounded Mixture of Gaussians, RMG)方法。该方法通过 Dirichlet 过程高斯混合(DPM)对潜变量进行连续密度估计，先对潜在变量建模为连续分布，再通过区间截断实现离散化，从而构造出灵活的创新项概率质量函数，兼具理论严谨性与计算可行性。

具体而言，对于每个整数 $j \in \mathbb{N}$，创新项的概率质量函数定义为：

$$p(j) = \Pr(\varepsilon_t = j) = g(f)[j] = \int_{a_j}^{a_{j+1}} f(\varepsilon^*) \mathrm{d}\varepsilon^* \tag{5}$$

其中，潜在连续密度 $f(\varepsilon^*)$ 为创新项的潜在连续密度函数，$g(\cdot)$ 表示对其进行离散化(即区间舍入)的映射操作。为了充分捕捉复杂分布形态，$f(\varepsilon^*)$ 采用非参数高斯混合模型建构：

$$f(\varepsilon^*; P) = \int \phi(\varepsilon^*; \mu, \sigma^2) \mathrm{d}P(\mu, \sigma^2), \quad P \sim DP(\eta, P_0) \tag{6}$$

其中是 $\phi(\cdot; \mu, \sigma^2)$ 表示均值为 μ，方差为 σ^2 的高斯密度函数；混合测度 P 服从以浓度参数 η 和基分布 P_0 为参数的过程。为实现 Dirichlet 过程式，采用 Sethuraman [17]提出的 stick-breaking 构造，该构造为 DPM 提供了可计算的序列形式表示：

$$P = \sum_{l=1}^{\infty} \pi_l \delta_{\theta_l}, \quad \theta_l \overset{\text{i.i.d.}}{\sim} P_0$$

其中，$\pi_1 = V_1$，而 $\pi_l = V_l \prod (1 - V_r)$，每个 $V_l \sim \mathrm{Beta}(1, \eta)$。这种构造可被解释为将单位长度“棒”依次按比例打断，生成一组加和为 1 的无穷权重序列 $\{\pi_l\}$ 赋予每个混合成分以相应的概率权重。

每个混合组分对应的参数 $\theta_l = (\mu_l, \sigma_l^2)$ 从 Normal-Gamma 基分布 P_0 中独立抽取，为反映数据的均值，方差不确定性，本文设置基分布为 Normal-Gamma 联合分布，即：

$$\sigma_l^{-2} \overset{\text{i.i.d.}}{\sim} Ga(a, b), \quad \mu_l \overset{\text{i.i.d.}}{\sim} N(\mu_0, \kappa\sigma_l^2)$$

其中，$Ga(a, b)$ 表示形状参数为 a，尺度参数为 b 的 Gamma 分布，κ 为控制均值与方差关系的比例系数。本文设置 $\mu_0 = \bar{y}$ 和样本方差 $\kappa = s^2$，并设置 $a = b = 0.5$。完成连续潜变量建模后，通过舍入函数 $r: \mathbb{R} \to \mathbb{N}$，将潜变量 $\varepsilon^* \in \mathbb{R}$ 映射为离散值，满足：$r(\varepsilon^*) = j$ 且 $a_j < \varepsilon^* \le a_{j+1}$。其中，$\{a_j\}$ 表示区间边界设置为 $a_0 = -\infty$，$a_j = j\ (j = 1, 2, \cdots)$。公式(5)和公式(6)在整数集上的概率质量函数空间中引入了先验分布该构造不仅提供了灵活的创新项分布建模能力，还具有直观的概率解释与良好的后验一致性，适合复杂计数数据建模。

3. 贝叶斯采样算法

在基于 Dirichlet 过程高斯混合模型(DPM)对 NGINAR(1)模型中创新项进行建模的框架下，由于创新项分布 p 无封闭形式表达，传统的参数估计方法难以直接应用。为解决这一问题，本文采用基于数据增强的 Gibbs 采样方法，该方法借鉴了 Canale 和 Dunson [12]提出的计算策略。通过引入潜在变量构建完整数据集，并在条件后验分布可解析的基础上，迭代实现模型参数与创新项分布的联合后验推断。具体的

采样过程包括以下几个步骤：

(1) 数据增强步骤(给定当前创新项分布 p 和 α)

a. 隐变量 B_t 的生成

对于每个时间点 t 和滞后阶 $k=1,\cdots,p$ ，定义隐变 $B_{t,k}$ 表示由历史观测 y_{t-k} 中保留至时间 t 的个体数。根据负二项稀疏算子 Ristić 等，其条件先验分布为：

$$B_{t,k} \mid y_{t-k},\alpha_k \sim \mathrm{NB}\left(y_{t-k}, p=\frac{1}{1+\alpha_k}\right)$$

即：

$$P\left(B_{t,k}=j \middle| y_{t-k},\alpha_k\right)=\binom{y_{t-k}+j-1}{j}\left(\frac{1}{1+\alpha_k}\right)^{y_{t-k}}\left(\frac{\alpha_k}{1+\alpha_k}\right)^{j}$$

给定观测值 y_t 和所有隐变量组成的向量 $\boldsymbol{B}_t=\left(B_{t,1},\cdots,B_{t,p}\right)$ ，隐变量 $\boldsymbol{B}_t$ 的联合后验分布与创新项分布 $p(\varepsilon)$ 相关，其条件后验概率为：

$$P\left(\boldsymbol{B}_t=\boldsymbol{j} \middle| y_t,y_{t-1},\cdots,y_{t-p},\boldsymbol{\alpha},p\right)\propto\left[\prod_{k=1}^{p}\binom{y_{t-k}+j_k-1}{j_k}\left(\frac{\alpha_k}{1+\alpha_k}\right)^{j_k}\right]\cdot p\left(y_t-\sum_{k=1}^{p}j_k\right)$$

b. 潜变量 ε_t^* 的生成

给定观测值 y_t 和隐变量 B_t ，定义创新项为 $\varepsilon_t=y_t-\sum_{k=1}^{p}B_{t,k}$ ，并从连续密度 f 中采样潜在变量 ε_t^* ，并约束其满足区间条件： $a_{\varepsilon_t}\le\varepsilon_t^*\le a_{\varepsilon_t+1}$ 。

(2) 更新创新分布 p

a. 分配变量 S_t 的更新：

从混合分布中采样每个观测点的分量标识 S_t ，其单元格概率为：

$$P\left(S_t=l \mid \varepsilon_t^*\right)\propto\pi_l\cdot\phi\left(\varepsilon_t^*;\mu_l,\sigma_l^2\right)$$

其中， π_l 是第 l 个混合成分的先验权重， $\phi\left(\varepsilon_t^*;\mu_l,\sigma_l^2\right)$ 是第 l 个成分的正态概率密度函数，表示在给定潜在变量 ε_t^* 下，观测值属于该混合成分的概率。

b. stick-breaking 权重更新

$$V_l\sim\mathrm{Be}\left(1+n_l,\eta+\sum_{r>l+1}n_r\right)$$

其中， n_l 是分配给第 l 个混合成分的观测数量。

c. 更新混合核参数 $\left(\mu_l,\sigma_l^2\right)$ ：

$$N\left(\hat{\mu}_l,\hat{\kappa}_l\sigma_l^2\right)InvGam\left(a+\frac{n_l}{2}+1,b+\hat{b}_l\right)$$

其中

$$\hat{\mu}_l=\hat{\kappa}_l\left(\kappa\mu_0+n_l\bar{\varepsilon}_l^*\right),\quad\hat{\kappa}_l=\left(\kappa^{-1}+n_l\right)^{-1}$$

$$\hat{b}_l=\frac{1}{2}\left\{\sum_{S_t=l}\left(\varepsilon_t^*-\bar{\varepsilon}_l^*\right)^2+\frac{n_l}{1+\kappa n_l}\left(\bar{\varepsilon}_l^*-\mu_0\right)^2\right\}$$

并且 $\bar{\varepsilon}_l^*$ 是第 l 类的样本均值。

(3) 更新参数：Metropolis-Hastings 步骤

在 NGINAR(p)模型中，自回归参数为向量 $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_p)$。给定隐变量 $\boldsymbol{B}$ 与观测序列 $\boldsymbol{y}$，每个 α_k 的完全条件后验分布为：$\pi(\alpha\mid \boldsymbol{y},\boldsymbol{B},p)\propto \pi_\alpha(\alpha)\times L(\alpha\mid \boldsymbol{y},\boldsymbol{B},p)$

$$P(\alpha_k\mid \boldsymbol{B},\boldsymbol{y})\propto \prod_t \binom{y_{t-k}+B_{t,k}-1}{B_{t,k}}\left(\frac{\alpha_k}{1+\alpha_k}\right)^{B_{t,k}}\left(\frac{1}{1+\alpha_k}\right)^{y_{t-k}}\cdot\pi(\alpha_k)$$

其中，$\pi(\alpha_k)$ 为先验分布，采用逐维 Metropolis-Hastings (MH) [17] [18]采样进行更新。具体步骤如下：对于每个 $k=1,\cdots,p$ 依次执行：

a. 生成候选值：从以当前值为中心的正态分布中抽样：

$$\alpha_k^*\sim \mathrm{N}\left(\alpha_k^{(m)},\sigma_k^2\right)$$

b. 计算接受概率：

$$r=\min\left\{1,\frac{\prod_t P\left(B_{t,k}\middle|y_{t-k},\alpha_k^*\right)\cdot\pi\left(\alpha_k^*\right)}{\prod_t P\left(B_{t,k}\middle|y_{t-k},\alpha_k\right)\cdot\pi\left(\alpha_k\right)}\right\}$$

c. 接受/拒绝：若随机接受，则令 α_k 的新值为 α_k^*；否则保留作为下一状态。

4. 数值模拟与结果分析

本研究通过蒙特卡洛模拟实验评估 NGINAR(p)模型在多步超前预测中的表现，这里取 $p=2$ 时，即采用 NGINAR(2)模型进行数值模拟。实验设计包含 1000 次独立重复试验，样本量分别设置为 50、100 和 250，同时引入 200 次预热迭代以消除初始值的影响。模型 A：$\alpha_1=0.2$，$\alpha_2=0.3$；模型 B：$\alpha_1=0.4$，$\alpha_2=0.3$；模型 C：$\alpha_1=0.6$，$\alpha_2=0.2$。整个模拟过程在相同数据生成机制下独立重复 1000 次。对于每个数据集，进行 17,000 次迭代采样，其中前 2000 次作为预热期(burn-in)以确保链收敛至平稳分布，保留后 15,000 次迭代结果用于后续分析。为充分检验贝叶斯非参数方法对不同类型计数数据的适应性，本文设计四种创新项分布：

i. 泊松分布[19]：$\lambda=3$；

ii. 负二项分布[20]：$k=8,p=0.5$；

iii. 二项分布[21]：$k=12,p=0.7$；

iv. CMP 分布[22]：$\lambda=25,v=4$。

为全面评估预测分布的准确性，本文采用多种评价指标，包括严格适当评分规则中的二次得分，以及 KL 散度和 Bhattacharyya 系数两种分布差异度量。二次得分(Quadratic Score)是一种严格适当的评分规则，由 Brier [23]首次提出用于概率预报的验证，是 Brier 得分在多类别情形下的推广形式。对于离散型整数值预测，记 $\hat{p}_h$ 为 h 步向前的预测概率质量函数，y_{t+h} 为实际观测值，则二次得分计算如下：

$$S(\hat{p}_h,y_{T+h})=2\hat{p}_h(y_{t+h})-\sum_j \hat{p}_h(j)^2-1$$

其中，求和中的 j 遍历了预测分布 $\hat{p}_h$ 所定义的所有可能取值。为了覆盖全部重要概率质量，这一范围通常设定为观测数据的最小值与最大值之间，并在此基础上添加一定的缓冲区±20，以确保计算时包含了预测分布的主要概率质量部分。Gneiting 和 Raftery [24]系统阐述了严格适当评分规则的理论框架，证明了二次得分等评分规则在概率预报评估中的优越性。

为评估预测效果，本文采用蒙特卡罗方法估计 Bhattacharyya 系数和 Kullback-Leibler 散度，这两种方法是广泛应用的分布差异度量技术[25] [26]。Hershey 和 Olsen [27]成功将该方法应用于高斯混合模型的信号处理任务中，用于估计 KL 散度和 BC 系数。这两种度量标准均可用于衡量预测分布与真实分布之间的差异，其计算公式如下：

$$\mathrm{BC}=\sum_{j}-\log\left(\sqrt{p_0(j)\hat{p}(j)}\right)\quad \mathrm{KL}=\sum_{j}p_0(j)\log\left(\frac{p_0(j)}{\hat{p}(j)}\right)$$

其中，$p_0(j)$表示真实的预测概率质量函数，$\hat{p}(j)$为通过模型估计的概率分布。计算时，j 表示预测变量的所有可能离散取值，并在此基础上加入±20 的缓冲区，以覆盖预测分布和真实分布的主要概率质量区域。

Table 1. Quadratic score of h step-ahead predictive distribution
表 1. h 步向前预测分布的二次得分

n	α	$p(\varepsilon)$	$h=1$	$h=2$	$h=3$	$h=4$
50	$\alpha_1=0.2$ $\alpha_2=0.3$	Pois	−0.9134	−0.8976	−0.9052	−0.9128
		NB	−0.8873	−0.8829	−0.8914	−0.8865
		Bin	−0.9278	−0.9258	−0.9258	−0.9269
		CMP	−0.8475	−0.8655	−0.8352	−0.8425
	$\alpha_1=0.4$ $\alpha_2=0.3$	Pois	−0.9215	−0.9332	−0.9287	−0.9410
		NB	−0.8642	−0.8715	−0.8698	−0.8733
		Bin	−0.9275	−0.9258	−0.9258	−0.9269
		CMP	−0.8205	−0.8344	−0.8351	−0.8282
	$\alpha_1=0.6$ $\alpha_2=0.2$	Pois	−0.9421	−0.9513	−0.9495	−0.9582
		NB	−0.9025	−0.9156	−0.9230	−0.9189
		Bin	−0.9593	−0.9411	−0.9652	−0.9674
		CMP	−0.9222	−0.9356	−0.9328	−0.9245
100	$\alpha_1=0.2$ $\alpha_2=0.3$	Pois	−0.9192	−0.9085	−0.8953	−0.9037
		NB	−0.8821	−0.8905	−0.8783	−0.8841
		Bin	−0.9245	−0.9277	−0.9207	−0.9236
		CMP	−0.8795	−0.8765	−0.8744	−0.8622
	$\alpha_1=0.4$ $\alpha_2=0.3$	Pois	−0.9258	−0.9384	−0.9176	−0.9269
		NB	−0.8573	−0.8625	−0.8528	−0.8592
		Bin	−0.9333	−0.9257	−0.9342	−0.9352
		CMP	−0.9142	−0.8778	−0.9201	−0.9111
	$\alpha_1=0.6$ $\alpha_2=0.2$	Pois	−0.9458	−0.9581	−0.9521	−0.9596
		NB	−0.8984	−0.9096	−0.9135	−0.9108
		Bin	−0.9543	−0.9642	−0.9624	−0.9675
		CMP	−0.9124	−0.9175	−0.9225	−0.9344

续表

250	$\alpha_1=0.2$ $\alpha_2=0.3$	Pois	−0.9235	−0.9315	−0.9091	−0.9183
		NB	−0.8786	−0.8892	−0.8828	−0.8789
		Bin	−0.8977	−0.8978	−0.8933	−0.9032
		CMP	−0.8642	−0.8534	−0.8564	−0.8563
	$\alpha_1=0.4$ $\alpha_2=0.3$	Pois	−0.9193	−0.9314	−0.9092	−0.9248
		NB	−0.8545	−0.8482	−0.8423	−0.8507
		Bin	−0.9212	−0.9233	−0.9242	−0.9345
		CMP	−0.8475	−0.8635	−0.8679	−0.8842
	$\alpha_1=0.6$ $\alpha_2=0.2$	Pois	−0.9482	−0.9557	−0.9519	−0.9605
		NB	−0.8942	−0.9097	−0.9118	−0.9086
		Bin	−0.9544	−0.9845	−0.9524	−0.9563
		CMP	−0.8852	−0.8997	−0.9075	−0.9012

表 1 展示了 NGINAR(2)模型在不同参数设定下 1 至 4 步向前预测的二次得分结果。结果表明，该模型在不同样本量、不同自回归系数及不同创新项分布下均表现出良好的预测能力，二次得分绝对值普遍接近理论最大值 0，预测分布较为集中且准确性较高。随着预测步长 h 从 1 增加至 4，各项得分保持稳定，未出现显著衰减，表明模型在短期至中期预测中具有较好的稳健性。

表 2 展示了不同创新项分布下 KL 散度与 Bhattacharyya 系数的估计结果。整体而言，随着样本量的增加，KL 散度与 BC 系数呈下降趋势，表明模型在分布拟合上具有良好的渐近一致性。当创新项真实分布为二项分布(Bin)时，KL 散度和 BC 系数在某些参数设定下显著高于其他分布。例如，在模型 C 且样本量为 250 时，二项分布对应的 KL 散度高达 2.08，BC 系数为 1.02，而同期泊松分布下的 KL 散度仅为 0.60，BC 系数为 0.19。这一差异可能与二项分布的“欠离散”特性有关。二项分布的方差小于其均值，而本文所采用的负二项稀疏算子基于几何分布，倾向于生成过度离散数据。当真实创新项呈现欠离散特征时，稀疏算子结构与数据生成机制之间存在内在冲突，导致模型拟合困难，表现为 KL 散度升高、BC 系数下降。这揭示了当前模型框架的一个潜在边界：贝叶斯非参数方法虽能灵活拟合各种分布形态，但核心的稀疏算子结构仍受限于特定的离散特征。

Table 2. Estimated KL divergence and Bhattacharyya coefficient
表 2. KL 散度与 BC 估计结果

参数	$p(\varepsilon)$	$n=50$		$n=100$		$n=250$	
		KL	BC	KL	BC	KL	BC
$\alpha_1=0.2$ $\alpha_2=0.3$	Pois	0.06	0.02	0.10	0.02	0.12	0.03
	NB	0.18	0.05	0.12	0.03	0.06	0.02
	Bin	0.22	0.07	0.15	0.31	0.14	0.22
	CMP	0.12	0.04	0.08	0.01	0.07	0.01
$\alpha_1=0.4$ $\alpha_2=0.3$	Pois	0.29	0.09	0.21	0.08	0.19	0.05
	NB	0.20	0.06	0.14	0.04	0.08	0.02
	Bin	1.26	0.78	1.18	0.54	0.87	0.42
	CMP	0.19	0.13	0.15	0.12	0.10	0.10

续表

	Pois	0.27	0.10	0.59	0.20	0.60	0.19
$\alpha_1 = 0.6$	NB	0.35	0.18	0.30	0.14	0.06	0.02
$\alpha_2 = 0.2$	Bin	2.32	1.11	2.22	1.08	2.08	1.02
	CMP	1.09	0.41	0.93	0.35	0.83	0.31

5. 实例分析

为了评估本章提出的基于贝叶斯非参数框架的 NGINAR(p)模型的预测性能，通过对 2018 年 1 月至 2022 年 12 月期间月度总云量数据的应用说明了所提出的方法。该数据来源于中国某地区气象观测站，其月度总云量数据由 60 个观测值组成。过离散指数为 2.95，表现出相当大的过离散特征。图 1 描述了月度总云量数据的直方图、序列图、ACF 图和 PACF 图。从 ACF 图和 PACF 图可以看出，数据存在显著的自相关结构，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。该数据可从中国气象数据网(http://data.cma.cn/)获得。

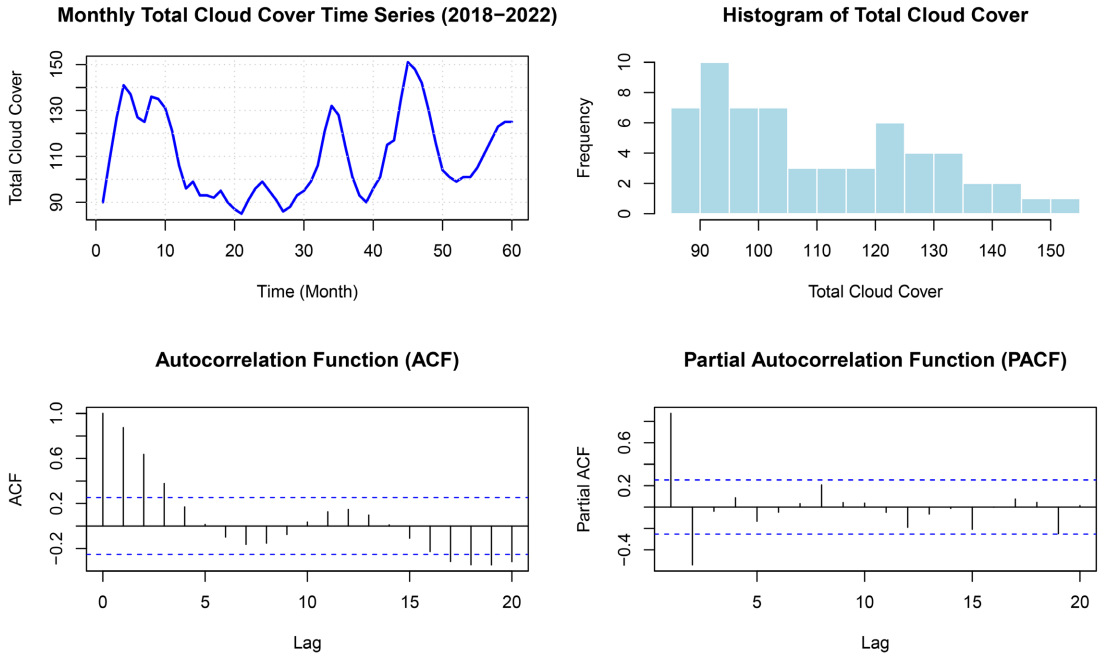

Figure 1. Time series plot, histogram, ACF and PACF of monthly total cloud cover data
图 1. 月度总云量数据的时序图、直方图、ACF 图与 PACF 图

图 1 展示了该序列的时间序列图、自相关(ACF)图与偏自相关(PACF)图。从 ACF 与 PACF 图可以看出数据在多个滞后项上存在显著的自相关性，同时从时间序列图与直方图中可观察到数据具有偏态与尾重特性。用引入的高阶结构的 NGINAR(p)模型，能更准确地拟合此类实际计数型时间序列数据。

图 2 展示了步超前预测概率质量函数的后验均值 Pr(Y = j|y)，以及基于预热后 MCMC 样本 2.5 分位数和 97.5 分位数估计得到的 95%后验可信带。从图 2 可以看出，各步预测分布均呈现明显的右偏形态，概率质量主要集中于低值区域，这与原始数据的偏态特征一致。这一预测分布形态的源头可追溯至模型对创新项的后验估计——创新项分布同样呈现右偏态，且概率质量在 0 和 1 处存在局部聚集，这正是舍

入高斯混合先验所捕捉到的数据驱动特征，验证了非参数建模的必要性。图 3 展示了未来云量低于 125 的后验预测概率，即 $\Pr\left(Y_{t+h} \le 10\right)$ (实线)，以及对应预测步长 $h=1,\cdots,6$ 的 95%可信带(虚线)。这些概率清晰地反映了预测不确定性预测步长增加而演变的规律，同时保持了整数值预测的内在一致性。

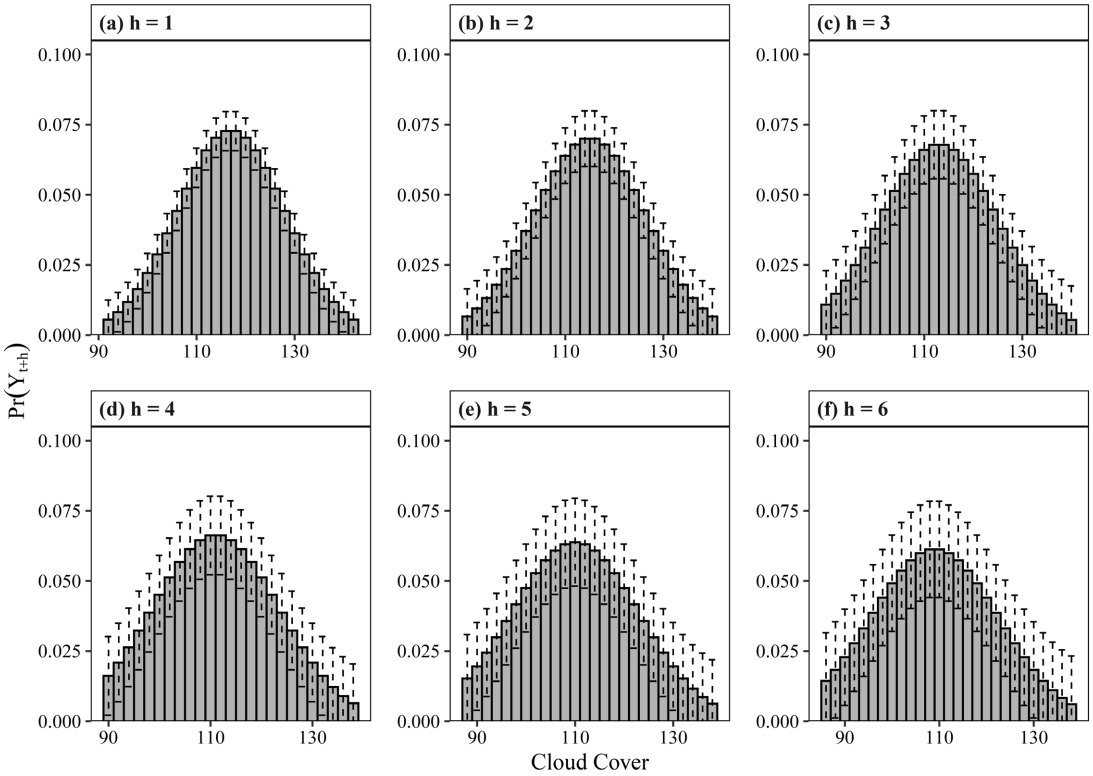

Figure 2. *h*-step-ahead predictive PMFs and 95% credible bands
图 2. *h* 步向前预测分布及其 95%可信区间

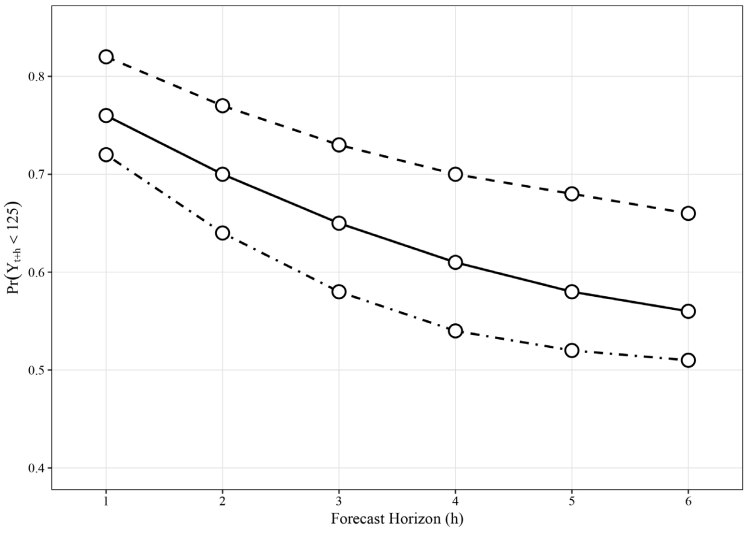

Figure 3. Posterior probabilities for less than 125 cases along with their 95% confidence intervals
图 3. 少于 125 的后验概率及其 95%可信区间

6. 结论与展望

本文提出了一种基于贝叶斯非参数框架的 NGINAR(p)模型预测方法。该方法通过将舍入高斯混合先验与负二项稀疏算子相结合，实现了对整数值时间序列中过度离散特征、高阶依赖结构及复杂分布形态的有效建模。在贝叶斯推断方面，构建了基于数据增强的 Gibbs 采样算法，实现了自回归系数与创新项分布的联合后验推断，并直接获得整数值的 h 步向前预测分布。

数值模拟研究表明，该方法在不同样本量、自回归系数及创新项分布设定下均表现出良好的预测精度与稳定性。在月度总云量数据的实证分析中，模型准确捕捉了序列的高阶自相关结构和过度离散特征，在多步向前预测中展现出优异的适应性。

综上所述，该方法为复杂计数时间序列的分析提供了一种兼具灵活性、稳健性与理论严谨性的新途径，在气象预测、环境监测、流行病预警及保险精算等领域具有良好的应用前景。后续研究可从以下方向进一步拓展：将贝叶斯非参数方法推广至多元整数值时间序列模型；优化 MCMC 采样算法以提高计算效率；以及探索模型在非平稳、季节性或带有协变量等更复杂场景下的应用。

基金项目

辽宁科技大学博士启动资金(6003000310)。

参考文献

[1] Steutel, F.W. and van Harn, K. (1979) Discrete Analogues of Self-Decomposability and Stability. *The Annals of Probability*, **7**, 893-899.

[2] 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

[3] Mckenzie, E. (1988) Some ARMA Models for Dependent Sequences of Poisson Counts. *Advances in Applied Probability*, **20**, 822-835. https://doi.org/10.2307/1427362

[4] Du, J.G. and Li, Y. (1991) The Integer-Valued Autoregressive (INAR(p)) Model. *Journal of Time Series Analysis*, **12**, 129-142. https://doi.org/10.1111/j.1467-9892.1991.tb00073.x

[5] Al-Osh, M. and Alzaid, A.A. (1988) Integer-Valued Moving Average (INMA) Process. *Statistical Papers*, **29**, 281-300. https://doi.org/10.1007/bf02924535

[6] Ferland, R., Latour, A. and Oraichi, D. (2006) Integer-Valued GARCH Process. *Journal of Time Series Analysis*, **27**, 923-942. https://doi.org/10.1111/j.1467-9892.2006.00496.x

[7] Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009) Poisson Autoregression. *Journal of the American Statistical Association*, **104**, 1430-1439. https://doi.org/10.1198/jasa.2009.tm08270

[8] Weiß, C.H. (2018) An Introduction to Discrete-Valued Time Series. Wiley. https://doi.org/10.1002/9781119097013

[9] Ristić, M.M., Bakouch, H.S. and Nastić, A.S. (2009) A New Geometric First-Order Integer-Valued Autoregressive (NGINAR(1)) Process. *Journal of Statistical Planning and Inference*, **139**, 2218-2226. https://doi.org/10.1016/j.jspi.2008.10.007

[10] Nastić, A.S., Ristić, M.M. and Bakouch, H.S. (2012) A Combined Geometric INAR(p) Model Based on Negative Binomial Thinning. *Mathematical and Computer Modelling*, **55**, 1665-1672. https://doi.org/10.1016/j.mcm.2011.10.080

[11] Bisaglia, L. and Canale, A. (2016) Bayesian Nonparametric Forecasting for INAR Models. *Computational Statistics & Data Analysis*, **100**, 70-78. https://doi.org/10.1016/j.csda.2014.12.011

[12] Canale, A. and Dunson, D.B. (2011) Bayesian Kernel Mixtures for Counts. *Journal of the American Statistical Association*, **106**, 1528-1539. https://doi.org/10.1198/jasa.2011.tm10552

[13] Lo, A.Y. (1984) On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, **12**, 351-357. https://doi.org/10.1214/aos/1176346412

[14] Escobar, M.D. and West, M. (1995) Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**, 577-588. https://doi.org/10.1080/01621459.1995.10476550

[15] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005) Univariate Discrete Distributions. Wiley. https://doi.org/10.1002/0471715816

[16] Fokianos, K. and Kedem, B. (2004) Partial Likelihood Inference for Time Series Following Generalized Linear Models. *Journal of Time Series Analysis*, **25**, 173-197. https://doi.org/10.1046/j.0143-9782.2003.00344.x

[17] Sethuraman, J. (1994) A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, **4**, 639-650.

[18] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**, 1087-1092. https://doi.org/10.1063/1.1699114

[19] Hastings, W.K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109.

[20] Poisson, S.D. (1837) Recherches sur la probabilité des jugements en matière criminelle et en matière civile: Précédées des règles générales du calcul des probabilités. Bachelier.

[21] Greenwood, M. and Yule, G.U. (1920) An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society*, **83**, 255-279. https://doi.org/10.2307/2341080

[22] Bernoulli, J. and Sheynin, O. (2005) On the Law of Large Numbers. NG Verlag.

[23] Brier, G.W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**, 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

[24] Gneiting, T. and Raftery, A.E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**, 359-378. https://doi.org/10.1198/016214506000001437

[25] Binder, K. and Heermann, D.W. (1992) Monte Carlo Simulation in Statistical Physics. Springer.

[26] Kroese, D.P., Taimre, T. and Botev, Z.I. (2013) Handbook of Monte Carlo Methods. Wiley.

[27] Hershey, J.R. and Olsen, P.A. (2007) Approximating the Kullback Leibler Divergence between Gaussian Mixture Models. 2007 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, 15-20 April 2007, 317-320. https://doi.org/10.1109/icassp.2007.366913