Advances in Applied Mathematics 应用数学进展, 2025, 14(6), 135-147
Published Online June 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.146307

Hans汉斯

基于经验似然INAR(1)模型离群点的检测与估计

王芊一一，卢飞龙*

辽宁科技大学理学院，辽宁 鞍山

收稿日期：2025年5月17日；录用日期：2025年6月11日；发布日期：2025年6月18日

摘　要

整值时间序列可出现在教育，金融，医疗，交通等诸多领域，本文旨在研究利用经验似然方法对整值时间序列中的加性离群点与新息离群点进行检测与估计，并针对凸包问题进行了详细讨论，最后通过数值模拟充分验证了经验似然方法检测离群点的有效性。仿真实验结果表明，经验似然方法可以有效检测与估计出不同新息分布下一阶整值时间序列模型中的离群点。

关键词

离群值，经验似然，INAR(1)模型

Detection and Estimation of Outliers in the Empirical Likelihood INAR(1) Model

Qianyiyi Wang, Feilong Lu*

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

Received: May 17th, 2025; accepted: Jun. 11th, 2025; published: Jun. 18th, 2025

Abstract

Integer-valued time series can appear in various fields such as education, finance, healthcare, and transportation. This paper aims to investigate the detection and estimation of additive outliers and innovation outliers in integer-valued time series based on the empirical likelihood method. Additionally, the convex hull problem is discussed in detail. Finally, numerical simulations are conducted to fully verify the effectiveness of the empirical likelihood method in detecting outliers. The simulation results show that the empirical likelihood method can effectively detect and estimate outliers in first-order integer-valued time series models with different innovation distributions.

*通讯作者。

文章引用: 王芊一一, 卢飞龙. 基于经验似然 INAR(1)模型离群点的检测与估计[J]. 应用数学进展, 2025, 14(6): 135-147.
DOI: 10.12677/aam.2025.146307

Keywords

Outlier, Empirical Likelihood, INAR(1) Model

1. 引言

经验似然是一种众所周知的方法，通过经验似然方法获得的置信区间没有对称性约束，因此可以更好地描述分布的真实形状。它有类似于 Bootstrap 的抽样特性，它允许在不指定数据概率分布的情况下对观察样本进行统计推断。它基于样本数据，通过估计其概率分布的最大似然参数来推断总体参数的值。具体来说，经验似然方法将样本数据看作样本的概率密度函数，并通过最大化该密度函数来估计总体参数。经验似然方法常用于建立概率模型、参数优化、假设检验等问题中。它是一种基本的统计方法，具有广泛的应用和重要的理论意义。

自从 Owen [1] [2]的开创性成果以来，经验似然方法受到学术界广泛的关注。在理论研究中，Owen [3]首先提出经验似然方法并应用到线性模型中，Qin 和 Lawless [4]基于无偏估计函数推进发展了经验似然方法，并将经验似然方法应用于广义估计方程。Tang 和 Leng [5]提出了一种利用经验似然的方法来同时进行参数估计和变量选择。这种方法通过优化经验似然函数来达到目的，并引入了惩罚项来约束参数的数量和复杂度。Shi 和 Lau [6]、Wei [7]等人以及 Liu 和 Yuan [8]等学者在各自的研究中将经验似然方法应用于不同的领域，展示了其在实际应用中的有效性。与此同时，Kitamura [9]和 Monti [10]等人也利用经验似然方法对相依数据进行了分析和预测工作。他们的研究表明，经验似然方法在数据科学领域的广泛应用，为研究人员提供了一种有效的工具来处理复杂的数据模型，帮助他们更好地理解并预测数据的行为。

在时间序列领域中，Chuang 和 Chan [11]以及 Chan 和 Ling [12]在其研究中采用经验似然方法，针对不同的假设条件下自回归模型进行了建模工作。其中，他们建立了在平稳和非平稳情况下的对数经验似然比统计量的极限分布。在研究中，Zhao 和 Wang [13]应用了经验似然方法，提供了自回归模型中回归参数的置信区间估计。他们的研究结果表明，对数经验似然比统计量在一定条件下收敛于卡方分布，为在自回归模型中对参数估计提供了理论依据。Nordman 和 Lahiri [14]给出了全面系统性的回顾和总结。

检测数据集中离群值的位置并估计其大小是统计学中比较常见的主题之一。Fox [15]首先在时间序列模型中提出加性离群点(Additive Outlier, AO)以及新息离群点(innovation Outlier, IO)，在此之后的时间里，对于离群点检测的方法研究受到了学术界广泛关注。如 Tsay [16]提出了一个迭代方法来识别离群点，以及消除它们对时间序列的影响。近几年研究人员的注意力转移到整值计数时间序列中离群值的检测上。Fokianos 和 Fried [17]提出了整数值广义条件异方差(INGARCH)模型框架下两类离群点的检测和估计方法。Fried [18]等人提出了在存在离群点和干预效应的情况下，采用广义线性模型对计数时间序列进行稳健拟合。INAR(1)是现有文献中提出的最成功的整值时间序列模型之一，Barczy [19] [20]分别考虑了在已知时间段内受创新离群点(IO)和加性离群点(AO)污染的 INAR(1)模型参数的条件经验似然估计。Silva [21]研究了 INAR(1)模型中含有加性离群点时的检测问题，并对加性离群点(AO)进行检测和定位，并同时对离群值大小进行估计。Baragona 等人[22]提出了使用经验似然方法检验 AR(p)模型中的离群点 Hua Shang

[23]提出了吉布斯抽样算法来估计在泊松 INAR(1)模型下的参数和离群点的大小。本研究首次将经验似然方法系统性地引入一阶整值自回归(INAR(1))模型的离群点分析领域，以及首次提出了将平衡经验似然用于 INAR(1)模型中进行优化，经验似然方法创新性地构建了基于约束优化与非参数似然比统计量的离群点联合检测框架，为后续离群点检测研究提供了理论支持。

本文的结构是，在第二节中，详细介绍了经验似然方法，第三节给出了含有加性离群点与新息离群点的 INAR(1)模型，并给出了估计统计量的渐近结果以及经验似然的改进方法——平衡经验似然。在第四节中通过模拟实验及其分析验证了本文所提经验似然方法的有效性，第五节介绍了经验似然方法检测离群点在实际生活中的应用，第六节为结论与展望，对全文进行总结，并提出未来可研究的方向。

2. 经验似然

经验似然(EL)方法是由 Owen [2]最早提出的一种非参数统计推断方法，与传统的参数统计方法不同，它无需事先假设数据服从某个特定的分布族。这种特性使得 EL 方法在处理实际问题时更具灵活性和广泛的适用性。通过直接计算总体分布的各阶矩，EL 方法避免了对参数渐近方差的估计，从而能够提供更为准确的统计推断结果。

设随机变量序列 $X_1, X_2, \cdots, X_n$ 具有独立同分布(i.i.d.)的特性，其具有共同且未知的概率分布 F_θ，感兴趣的未知参数记为 $\boldsymbol{\theta} = \left(\theta_1, \cdots, \theta_p\right)^{\mathrm{T}}$，那么此时经验似然定义为：

$$L\left(F_\theta\right) = \prod_{i=1}^{n} dF_\theta\left(X_i\right) = \prod_{i=1}^{n} p_i, \tag{1}$$

其中，$p_i = dF_\theta\left(X_i\right) = \Pr\left(X = X_i\right)$，$0 \le p_i \le 1$，$\sum_{i=1}^{n} p_i = 1$。

而经验累积分布函数 $F_n\left(x\right) = n^{-1}\sum_{i=1}^{n} I\left(X_i < x\right)$ 是 F_θ 的非参数极大似然估计，经验似然比(ELR)函数定义为：

$$R\left(F_\theta\right) = \frac{L\left(F_\theta\right)}{L\left(F_n\right)} = \prod_{i=1}^{n} np_i. \tag{2}$$

若对参数 θ 进行估计推断，通常需构造 r 个估计方程满足：$E_F\left\{g_j\left(\boldsymbol{X}, \boldsymbol{\theta}\right)\right\} = 0$，$g_j\left(\boldsymbol{X}, \boldsymbol{\theta}\right)$，$j = 1, 2, \cdots, r$，$r \ge p$，将向量表达式设为：

$$g\left(\boldsymbol{X}, \boldsymbol{\theta}\right) = \left(g_1\left(\boldsymbol{X}, \boldsymbol{\theta}\right), \cdots, g_r\left(\boldsymbol{X}, \boldsymbol{\theta}\right)\right)^{\mathrm{T}},$$

此时经验似然比函数为：

$$R\left(\boldsymbol{\theta}\right) = \sup\left\{\prod_{i=1}^{n} np_i \mid p_i \ge 0, \sum_{i}^{n} p_i = 1, \sum_{i=1}^{n} p_i g\left(\boldsymbol{X}, \boldsymbol{\theta}\right) = 0\right\}. \tag{3}$$

由拉格朗日乘子法，可作辅助函数：

$$H = \sum_{i}^{n} \log p_i + \lambda\left(1 - \sum_{i}^{n} p_i\right) - n\boldsymbol{t}^{\mathrm{T}} \sum_{i}^{n} p_i g\left(X_i, \boldsymbol{\theta}\right), \tag{4}$$

其中，λ 和 $\boldsymbol{t} = \left(t_1, t_2, \cdots, t_r\right)^{\mathrm{T}}$ 均为拉格朗日乘子。将辅助函数(4)对其 p_i 求偏导，得到 $p_i = \frac{1}{n} \cdot \frac{1}{1 + \boldsymbol{t}^{\mathrm{T}} g\left(X_i, \boldsymbol{\theta}\right)}$，并且 $\boldsymbol{t}$ 满足 $\sum_{i=1}^{n} \frac{1}{n} \cdot \frac{g\left(X_i, \boldsymbol{\theta}\right)}{1 + \boldsymbol{t}^{\mathrm{T}} g\left(X_i, \boldsymbol{\theta}\right)} = 0$。

由此，可得到对数经验似然比为：

$$l_E(\boldsymbol{\theta}) = -2\log R_E(\boldsymbol{\theta}) = 2\sum_{i=1}^{n}\log\left(1+\boldsymbol{t}^{\mathrm{T}} g(X_i,\boldsymbol{\theta})\right). \tag{5}$$

Owen (1991)证明了，当 n 趋于无穷大时，若 $\theta=\theta_0$，上述对数经验似然比统计量 $l_E(\boldsymbol{\theta})$ 渐近服从自由度为 p 的 χ^2 分布。

3. 基于经验似然 INAR(1)模型的离群点检测

本章针对 INAR(1)模型的离群点检测问题。INAR(1)模型作为离散时间序列的一种特殊形式，广泛应用于描述计数数据(如事件发生次数、事故发生频率等)。在该模型中，离群点的存在可能会影响自回归参数的估计，进而导致不准确的预测和决策。因此，本章提出了一种基于经验似然方法对加性离群点(AO)与新息离群点(IO)的检测方案。

3.1. 含有离群点的 INAR(1)模型及经验似然

首先介绍传统 INAR 模型，INAR 模型是一种常见的离散时间模型，广泛应用于金融、生物医学、通信等领域。其研究方法是通过稀疏算子构建模型，其中基于二项稀疏算子的整值一阶自回归 INAR(1)模型是最为经典的模型之一。其定义如下：

$$Z_t = \alpha\circ Z_{t-1} + \varepsilon_t,\quad t=0,\pm1,\pm2,\cdots \tag{6}$$

其中，Z 是非负的整数值随机变量，$\alpha\in[0,1)$，$\{\varepsilon_t\}$ 为独立同分布且均值为 μ，方差为 σ^2 的取值为整数值的随机变量序列，“$\circ$”为二项稀疏算子，其被定义为 $\alpha\circ Z_{t-1}=\sum_{i=1}^{Z_{t-1}} Y_i$，这里 $\{Y_i\}$ 是一列独立同分布的以 α 为参数的伯努利随机变量，并独立于 Z，其分布律为 $P\{Y_i=1\}=1-P\{Y_i=0\}=\alpha$。

在时间序列中，离群点的种类繁多，一般分为四类：加性离群点(AO 类)、新息离群点(IO 类)、水平移位离群点(LS 类)和暂时变更离群点(TC 类)。

其中，加性离群点(AO)是指在某个特定时间点发生的外部误差或外生变化，这种误差或变化只影响到该时间点的观测值，而不对其他时间点的观测值产生影响，这种离群点通常是由于观测或记录过程中的错误或误差所导致。新息离群点(IO)是由内部变化或噪声过程中的内源性因素引起的，其影响会传播到所有后续的观测值，并且其影响方式与系统的动态模型密切相关。水平移位离群点(LS)的出现会导致序列均值的偏移，从而影响后续所有观测值。这种离群点的产生通常是由于干扰造成的，进而改变了系统的整体结构，进而对时间序列产生持久的影响。暂时变更离群点(TC)的出现不仅会影响当前的观测值，还会对后续所有观测值产生影响，不过这种影响的程度会随着时间的推移而呈指数衰减。本文主要研究 AO 与 IO 两种离群点形式的检测。

定义 1 随机过程 $(Y_n)_{n\in\mathbb{Z}_+}$ 称为含有有限多个加性离群点(AO)的 INAR(1)时间序列模型，若满足：

$$Y_k = Z_k + \sum_{i=1}^{I} c_{k,s_i}\omega_i,\quad k\in\mathbb{Z}_+ \tag{7}$$

其中，$(Z_n)_{n\in\mathbb{Z}_+}$ 为满足 INAR(1)模型的序列，ω_i 为离群点的大小；c_{k,s_i} 为一个脉冲，当 $k=s_i$ 时取值为 1，否则为 0。此模型假设只有在特定的时间点 s_i 发生离群扰动，并且扰动的大小由 ω_i 决定

若单个 AO 位置任意，假设($I=1$，$s_1:=s$)，由公式(7)可以得出：

$$\begin{aligned}\mathrm{E}\left(Y_k \mid \mathcal{F}_{k-1}^{Y}\right) &= \alpha X_{k-1} + \mu_\varepsilon + c_{k,s}\omega \\ &= \alpha\left(Y_{k-1} - c_{k-1,s}\omega\right) + c_{k,s}\omega \\ &= \alpha Y_{k-1} + \mu_\varepsilon + \left(-\alpha c_{k-1,s} + c_{k,s}\right)\omega \\ &= \begin{cases} \alpha Y_{k-1} + \mu_\varepsilon & \text{if } k = 1,\cdots,s-1 \\ \alpha Y_{k-1} + \mu_\varepsilon + \omega & \text{if } k = s \\ \alpha Y_{k-1} + \mu_\varepsilon - \alpha\omega & \text{if } k = s+1 \\ \alpha Y_{k-1} + \mu_\varepsilon & \text{if } k \geq s+2 \end{cases}. \end{aligned} \tag{8}$$

(8)公式表明，在时间点 s 发生离群扰动时，模型的期望值会在其前后位置发生变化。

因此，计算其残差平方和得到：

$$\begin{aligned}\sum_{k=1}^{n}\left(Y_k - \mathrm{E}\left(Y_k \mid \mathcal{F}_{k-1}^{Y}\right)\right)^2 = \sum_{k=1}^{n}{}^{(s,s+1)}\left(Y_k - \alpha Y_{k-1} - \mu_\varepsilon\right)^2 + \left(Y_s - \alpha Y_{s-1} - \mu_\varepsilon - \omega\right)^2 \\ + \left(Y_{s+1} - \alpha Y_s + \mu_\varepsilon + \alpha\omega\right)^2.\end{aligned} \tag{9}$$

为了对模型中的参数 α， μ， ω 进行最优化估计，可以使用最小二乘法(LS)来获得。针对模型(7)，可以通过一组估计方程来获得参数的最小二乘估计。

最小二乘估计可以写成如下形式：

$$\begin{aligned} g_1\left(Y_k,\alpha,\mu,\omega\right) &= -2\sum_{k=1}^{n} Y_{k-1}\left(Y_k - \alpha Y_{k-1} - \mu_\epsilon\right) - 2\sum_{k=s+1}^{n} Y_{k-1}\left(Y_k - \alpha Y_{k-1} - \mu_\epsilon\right) \\ &\quad + 2\left(Y_{s+1} - \alpha Y_s - \mu_\epsilon + \alpha\omega\right)\left(-Y_s + \omega\right) \\ g_2\left(Y_k,\alpha,\mu,\omega\right) &= -2\sum_{k=1}^{n}\left(Y_k - \alpha Y_{k-1} - \mu_\epsilon\right) - 2\sum_{k=s+1}^{n}\left(Y_k - \alpha Y_{k-1} - \mu_\epsilon\right) \\ &\quad - 2\left(Y_{s+1} - \alpha Y_s - \mu_\epsilon + \alpha\omega\right) \\ g_3\left(Y_k,\alpha,\mu,\omega\right) &= -2\alpha\left(Y_{s+1} - \alpha Y_s - \mu_\epsilon + \alpha\omega\right). \end{aligned} \tag{10}$$

定义 2 令 $\left(\varepsilon_k\right)_{k\in\mathbb{N}}$ 为独立同分布(i.i.d)非负整值随机变量序列，随机过程 $\left(Y_n\right)_{n\in\mathbb{Z}_+}$ 称为含有有限多个新息离群点(IO)的 INAR(1)时间序列模型，若满足

$$Y_k = \sum_{j=1}^{Y_{k-1}} \xi_{k,j} + \eta_k, \quad k \in \mathbb{N}. \tag{11}$$

其中，$\left(\xi_{k,j}\right)_{j\in\mathbb{N}}$ 为独立同分布的均值 $\alpha\in(0,1)$ 的伯努利随机变量，且与 $\left(\varepsilon_l\right)_{l\in\mathbb{N}}$ 相互独立，Y_0 为独立于 $\left(\xi_{k,j}\right)_{j\in\mathbb{N}}, k\in\mathbb{N}$ 以及 $\left(\varepsilon_l\right)_{l\in\mathbb{N}}$ 的非负整值随机变量，且

$$\eta_k = \varepsilon_k + \sum_{i=1}^{I} c_{k,s_i}\omega_i, \quad k \in \mathbb{Z}_+.$$

其中，$I\in\mathbb{N}, s_i, \omega_i \in \mathbb{N}, i = 1,\cdots,I$；$\omega_i$ 为离群点大小；c_{k,s_i} 为一个脉冲，当 $k = s_i$ 时取值为 1，否则为 0。

由此可见，若单个 IO 位置任意，假设($I=1$，$s_1 := s$)，由公式(11)，可以得出

$$\mathrm{E}\left(Y_k \mid \mathcal{F}_{k-1}^{Y}\right) = \alpha Y_{k-1} + \mathrm{E}\,\eta_k = \alpha Y_{k-1} + \mu_\varepsilon + c_{k,s}\omega, k \in \mathbb{N}.$$

因此，对于所有，$n \geq \max(3, s+1)$，都有

$$\sum_{k=1}^{n}\left(Y_k - \mathrm{E}\left(Y_k \mid \mathcal{F}_{k-1}^{Y}\right)\right)^2 = \sum_{k=1}^{n}\left(Y_k - \alpha Y_{k-1} - \mu_\varepsilon\right)^2 + \left(Y_s - \alpha Y_{s-1} - \mu_\varepsilon - \omega\right)^2. \tag{12}$$

通过分别对其 α， μ， ω 求偏导得出最小二乘估计方程为：

$$
\begin{aligned}
g_1(Y_k,\alpha,\mu,\omega) &= -2\sum_{k=1}^{n} Y_{k-1}\left(Y_k-\alpha Y_{k-1}-\mu_\epsilon\right)-2\sum_{s=1}^{n} Y_{s-1}\left(Y_s-\alpha Y_{s-1}-\mu_\epsilon-\omega\right)\\
g_2(Y_k,\alpha,\mu,\omega) &= -2\sum_{k=1}^{n}\left(Y_k-\alpha Y_{k-1}-\mu_\epsilon\right)-2\sum_{s=1}^{n}\left(Y_s-\alpha Y_{s-1}-\mu_\epsilon-\omega\right)\\
g_3(Y_k,\alpha,\mu,\omega) &= -2\sum_{s=1}^{n}\left(Y_s-\alpha Y_{s-1}-\mu_\epsilon-\omega\right).
\end{aligned}
\tag{13}
$$

此时，针对两种不同类型的离群点构造，经验似然比(ELR)为：

$$
\ell(\alpha,\mu,\omega) = -2\sup_k\left\{\sum_{k=2}^{N}\log(Nw_k):\sum_{k=2}^{N}w_k=1,\sum_{k=2}^{N}w_k g_j(Y_k,\alpha,\mu,\omega)=0,\ j=1,2,3\right\}. \tag{14}
$$

定理 1 若$(\alpha_0,\mu_0,\omega_0)$为模型参数的真值，那么当样本量趋于无穷时，统计量$\ell(\alpha_0,\mu_0,\omega_0)$渐近服从以自由度为 3 的中心$\chi^2$分布。

通过定理 1 对$\ell(\alpha_0,\mu_0,\omega_0)$的渐近分布，很容易导出其置信区域，此外，还可以根据这个分布推导出参数检验，其定理的证明可以仿造 Baragona [22]中的证明步骤，在此不给出具体的证明。若有多个离群值，可以仿造上述思路构建经验似然比统计量。

3.2. 平衡经验似然(BEL)

在经验似然方法中，解决凸包问题一直是个重要的研究方向，本文参考了文献中提出的几种解决方案(Chen 等人，Emerson 和 Owen 以及 Tsao) [24]-[26]，研究表明，调节常数的有限样本校准是影响方法性能的关键因素。基于此，可以采用平衡经验似然(BEL)的方法，其优势在于可通过解析准则选择调节常数，在保持经验似然渐近性质的同时，减少了计算复杂性，并提高了对小样本数据的适应性。

在检测离群点问题中，估计方程约束系统的不可行性通常仅发生于最后一个矩条件。基于此特性，可以提出假设：$\bar{g}_j(\alpha,\mu,\omega)\simeq 0$，$j=1,2,3$。此时，BEL 方法通过加入两个人工观测值来简化约束系统，从而避免无解的情况：

$$
\begin{aligned}
g(y_{N+1},\alpha,\mu,\omega) &= \left[0,0,\cdots,0,-\delta\bar{g}_2(\alpha,\mu,\omega)\right]^{\mathrm{T}}\\
g(y_{N+2},\alpha,\mu,\omega) &= \left[0,0,\cdots,0,(2+\delta)\bar{g}_2(\alpha,\mu,\omega)\right]^{\mathrm{T}}.
\end{aligned}
$$

在该方法中，引入调节参数$\delta>0$的目的是确保所添加的人工数据不会对最终的求解结果造成显著的干扰。借助这种处理方式，原本因约束条件无法满足而无解的问题得以成功解决。

此外，考虑到$\bar{g}_2(\alpha,\mu,\omega)$的量级相对于调节参数$\delta$来说可以忽略，因此，可以假设：

$$
\delta\bar{g}_2(\alpha,\mu,\omega)=(2+\delta)\bar{g}_2(\alpha,\mu,\omega)=\Delta,
$$

基于上述，BEL 方法中加入的两个人工观测值可以写成：

$$
\begin{aligned}
g(y_{N+1},\alpha,\mu,\omega) &= \left[0,0,\cdots,0,-\Delta\right]^{\mathrm{T}}\\
g(y_{N+2},\alpha,\mu,\omega) &= \left[0,0,\cdots,0,\Delta\right]^{\mathrm{T}},
\end{aligned}
$$

那么，平衡经验似然比(BELR)为：

$$
\begin{aligned}
\ell_\Delta^*(\alpha,\mu,\omega) = -2\sup_w\Bigg\{&\sum_{k=2}^{N+2}\log\left[(N+1)w_k\right]:\sum_{k=2}^{N+2}w_k=1;\\
&\sum_{k=2}^{N}w_k g_j(y_k,\alpha,\mu,\omega)=0,\ j=1,2,3;\\
&\sum_{k=2}^{N}w_k g_2(y_k,\alpha,\mu,\omega)-w_{N+1}\Delta+w_{N+2}\Delta=0\Bigg\}.
\end{aligned}
\tag{15}
$$

常数 Δ 的选取对 BELR 计算结果具有重要影响，如果选择 $\Delta = o\left(N^{1/2}\right)$ 那么根据定理 2 的结果，$\ell_{\Delta}^{*}\left(\alpha_0, \mu_0, \omega_0\right)$ 在渐近意义下将与经验似然比(ELR)(14)相等。

定理 2 当 $\Delta = o\left(N^{1/2}\right)$ 时，ELR(14)与 BELR(15)的差别随着样本量\$N\$的增大依概率收敛为 0。

最后注意到，

$$\sum_{k=2}^{N+2} g_2\left(y_k, \alpha, \mu, \omega\right) = \sum_{k=2}^{N} g_2\left(y_k, \alpha, \mu, \omega\right).$$

这表明在计算平衡经验似然比(BELR)时，加入的人工观测值对整体结果没有影响。由此可知，参数的最大平衡经验似然估计量(MBELE)与最大经验似然估计量(MELE)以及联合最小二乘估计是等价的。

当 BELR 在与真实参数不同的点上计算时，它往往会变得任意大，如定理 3 所示。

定理 3 当 $\omega \neq \omega_0$，随着 N 逐渐增大，BELR 具有发散特性：

$$N^{-1/3}\ell_{\Delta}^{*}\left(\alpha_0, \mu_0, \omega\right) \to \infty$$

定理 2 与 3 的证明可以仿造 Baragona [22]中的证明步骤，在此不给出具体的证明。

这一特性揭示了在计算平衡经验似然比(BELR)时，若所采用的参数与真实值存在偏差，哪怕这种偏差微乎其微，BELR 的值也会显著上升。由此可知，该方法对于参数估计的准确性极为敏感，一旦参数估计偏离真实值，BELR 的显著变化即可作为模型误设检测的有力理论依据。

4. 模拟实验及其分析

在本节中，进行了一项模拟研究，生成 1000 组长度为 $N = 100$ 的模拟时间序列。所使用的一阶整值自回归模型如下：

模型 1：一阶整值自回归模型 $y_k = 0.2 \circ y_{k-1} + \varepsilon_t$。

模型 2：一阶整值自回归模型 $y_k = 0.5 \circ y_{k-1} + \varepsilon_t$。

模型 3：一阶整值自回归模型 $y_k = 0.8 \circ y_{k-1} + \varepsilon_t$。

其中，$\{\varepsilon_t\}$ 是独立同分布(i.i.d.)的取值为整数值的随机变量序列，且根据以下三种不同的概率分布生成：泊松分布(Poisson)，几何分布(Geometric)，二项分布(Binomial)。

本模拟采用了三种不同的方法来进行离群点检测，每种方法如下：

方法 1：使用公式(14)中的 ELR，其检验统计量为 $\inf_{\alpha,\mu} \ell\left(\alpha, \mu, 0\right)$。

方法 2：使用公式(15)中的 BELR，其检验统计量为 $\inf_{\alpha,\mu} \ell_{\Delta}^{*}\left(\alpha, \mu, 0\right)$。

方法 3：使用贝叶斯检测法，参考 Silva 等人[21]的计算步骤。

在本研究中，采用一种顺序检测的程序，该程序在每个步骤中识别最可疑的观测值，并去除其对剩余数据的影响。具体程序如下：

对每个时间点 s $(2 \le s \le N)$ 计算了相应的统计量，并计算所有时间点统计量的最大值。根据卡方分布的 $1-\alpha/(N-p)$ 分位数进行假设检验，如果统计量的最大值超过了该分位数，则拒绝原假设，即认为数据中存在离群点，离群点发生的时间被假定为统计量最大值对应的时间点，并在此时间点处估计离群值的大小再进行调整去除，在调整后的数据序列 $\{y_k'\}$ 上重复上述检测程序，直到没有显著的统计量为止。

首先，考虑泊松分布生成的模拟时间序列。在 1000 次重复实验中，对于检验无离群点的假设与存在加性离群点的假设之间的对比，理论显著性水平 $\alpha = 0.05$，每种模型和方法的观察频率如表 1 所示。可以看出，观察到的大小总是非常接近名义值，贝叶斯方法相较于其他两种方法略差一些，对于方法 2 (BELR)，其结果稍微更准确。

Table 1. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series generated according to Poisson innovations
表 1. 基于泊松创新生成的 1000 组序列中，拒绝无离群点假设的观测相对频率

模型	方法		
	1	2	3
1	0.078	0.064	0.073
2	0.084	0.067	0.092
3	0.072	0.078	0.081

随后，在每个模拟系列的时间点 $q=50$ 插入一个加性离群点，大小为 $\omega_0=3.5$，并重复检验，得到了表 2 所示的观察功效。并且区分了检测到的离群点是否正确定位在时间 50。

Table 2. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series generated according to Poisson innovations ($q=50$, $\omega_0=3.5$)
表 2. 基于泊松序列创新生成 1000 组包含加性离群点的序列中，无离群点的假设被拒绝的观测相对频率，其中 $q=50$ ，$\omega_0=3.5$

模型	时间位置	方法		
		1	2	3
1	准确	0.643	0.818	0.532
	错误	0.034	0.020	0.044
2	准确	0.382	0.671	0.431
	错误	0.043	0.029	0.072
3	准确	0.552	0.842	0.623
	错误	0.048	0.035	0.423

首先，我们注意到，使用方法 1 与方法 2 (即 ELR 与 BELR)的结果差距较为明显，并且方法 2 (BELR)错误定位的频率相对较低，效果更为准确。而方法 3 的检测功效相对其他两种方法相对较低。

接下来，考虑非泊松创新的情况。在模拟实验中，使用了模型 1，并通过几何分布(Geometric)和二项分布(Binomial)生成创新项。我们应用上述三种方法进行检验，并将结果报告在表 3 中。

可以观察到，对于 Geometric 分布生成的创新项，方法 2 (BELR)观测到的检验大小方与泊松创新的情况没有显著变化。然而，方法 1 (ELR)与方法 3 观察到的检验大小明显增加。对于 Binomial 分布生成的创新项，结果则显示出更加严重的问题：除了方法 2 的观测大小依然膨胀外，方法 1 观察到的检验大小是原来的六倍，方法 3 观测到的检验大小也出现了显著膨胀，这表明这些方法在面对 Binomial 分布的创新项时，均存在显著的偏差。

若考虑在非泊松创新的情况下，依旧向每个模拟序列的时间点 $q=50$ 插入大小为 $\omega_0=3.5$ 的加性离群点，并重复进行检测，所得到的观察功效如表 4 所示，可以得知，对于 Geometric 分布生成的创新项，方法 1，方法 2 与方法 3 的观察功效与泊松创新的情形相似，增加离群值的大小并未导致功效有显著提升，对其产生的影响是有限的，但方法 3 的功效略差于其他两种方法，方法 2 错误定位的频率要明显低于方

法 1 与方法 3，在检测到离群点时能更准确地定位到正确的时间点。对于 Binomial 分布生成的创新项，三种方法均存在显著偏差，但方法 1 和方法 2 的效果更优于方法 3。

Table 3. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series generated according to model 1 and non-Poisson innovations
表 3. 基于模型 1 和非泊松创新生成的 1000 组序列中，拒绝无离群点假设的观测相对频率

创新	方法		
	1	2	3
Geometric	0.254	0.078	0.332
Binomial	0.478	0.247	0.572

Table 4. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series with an additive outlier at $q=50$ and size $\omega_0=3.5$ generated according to model 1 and non-Poisson innovations
表 4. 基于模型 1 和非泊松创新生成的 1000 组序列中，假设无离群点的假设被拒绝的观测相对频率，其中 $q=50$，$\omega_0=3.5$

创新	时间位置	方法		
		1	2	3
Geometric	准确	0.684	0.859	0.578
	错误	0.041	0.032	0.052
Binomial	准确	0.287	0.543	0.256
	错误	0.032	0.029	0.042

在接下来的模拟实验中，继续考虑非泊松分布情况。具体而言，我们采用了模型 2，并利用几何分布(Geometric)和二项分布(Binomial)来生成创新项。随后，我们运用上述三种方法对这些数据进行了检验，并将检验结果汇总在表 5 中。

Table 5. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series generated according to model 2 and non-Poisson innovations
表 5. 基于模型 2 和非泊松创新生成的 1000 组序列中，拒绝无离群点假设的观测相对频率

创新	方法		
	1	2	3
Geometric	0.342	0.087	0.432
Binomial	0.562	0.278	0.572

我们观察到，当创新项由几何分布(Geometric)生成时，方法 2 (BELR)的检验结果与泊松分布创新项的情况基本一致，检验大小仅稍微增大。而方法 1 (ELR)与方法 3 的检验大小却明显增大。

当创新项由二项分布(Binomial)生成时，问题更为严重。方法 2 的检验大小仍然存在膨胀现象，而方法 1 与方法 3 的检验大小更是可以观察到显著增加。这说明在处理二项分布(Binomial)创新项时，这三种方法都存在明显的偏差。

在非泊松创新的场景下，我们对每个模拟序列的时间点 $q=50$ 插入大小为 $\omega_0=3.5$ 的加性离群值，并重复进行检测。检测结果如表 6 所示。分析发现：对于由几何分布(Geometric)生成的创新项，方法 1，方法 2 与方法 3 的检测功效与泊松分布创新项的情况相近，方法 3 的功效略低于其他两种方法，可以表明尽管增加离群值的大小，但检测功效并未显著提升。此外，方法 2 在检测到离群点时，错误定位的频率明显低于方法 1 与方法 3，能够更准确地定位到离群值的正确时间点。对于由二项分布(Binomial)生成的创新项，三种方法的检测结果都存在显著偏差。

Table 6. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series with an additive outlier at $q=50$ and size $\omega_0=3.5$ generated according to model 2 and non-Poisson innovations

表 6. 基于模型 2 和非泊松创新生成的 1000 组序列中，假设无离群点的假设被拒绝的观测相对频率，其中 $q=50$，$\omega_0=3.5$

创新	时间位置	方法		
		1	2	3
Geometric	准确	0.375	0.685	0.411
	错误	0.052	0.025	0.068
Binomial	准确	0.266	0.382	0.213
	错误	0.043	0.033	0.051

对于模型 3，我们依然考虑非泊松分布情况，并利用几何分布(Geometric)和二项分布(Binomial)来生成创新项，继续运用上述三种方法对这些数据进行了检验，并将检验结果汇总在表 7 中。我们观察到，当创新项由几何分布(Geometric)生成时，方法 2 (BELR)的检验结果与泊松分布创新项的情况基本一致，而方法 1 (ELR)与方法 3 的检验大小显著增大。当创新项由二项分布(Binomial)生成时，问题依旧更为严重。方法 1，方法 2 与方法 3 均显著膨胀。

Table 7. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series generated according to model 3 and non-Poisson innovations

表 7. 基于模型 3 和非泊松创新生成的 1000 组序列中，拒绝无离群点假设的观测相对频率

创新	方法		
	1	2	3
Geometric	0.314	0.091	0.472
Binomial	0.563	0.315	0.682

在非泊松创新的场景下，我们对每个模拟序列的时间点 $q=50$ 插入大小为 $\omega_0=3.5$ 的加性离群点，并重复进行检测。检测结果如表 8 所示。分析可观察到，对于由几何分布(Geometric)生成的创新项，方法 1 和方法 2 的检测功效与泊松分布创新项的情况相近，方法 3 略差一些。此外，方法 2 在检测到离群点时，错误定位的频率显著低于方法 1 与方法 3。对于由二项分布(Binomial)生成的创新项，三种方法的检测结果都存在显著偏差。

综合来看，对于泊松数据，BELR 的检验方法更加有效，而在非泊松创新的情况下，BELR 的方法效果依旧比 ELR 与贝叶斯方法显著，此外，也充分证明了在 INAR(1)模型中对离群点的检测与估计使用经验似然方法是有效的。

Table 8. Observed relative frequency of rejection of the hypothesis of the absence of outliers on 1000 series with an additive outlier at $q=50$ and size $\omega_0=3.5$ generated according to model 3 and non-Poisson innovations

表 8. 基于模型 3 和非泊松创新生成的 1000 组序列中，假设无离群点的假设被拒绝的观测相对频率，其中 $q=50$，$\omega_0=3.5$

创新	时间位置	方法		
		1	2	3
Geometric	准确	0.674	0.836	0.534
	错误	0.051	0.031	0.055
Binomial	准确	0.316	0.427	0.312
	错误	0.047	0.031	0.062

5. 实例分析

曲杆菌病是由弯曲杆菌属病原体引发的急性细菌感染。该病原体主要攻击人体消化系统，是全球细菌性胃肠炎最主要的致病菌之一。临床研究显示，感染后潜伏期存在个体差异，多数患者在暴露后 2~5 日出现典型症状，但存在 1~10 日的波动区间，这与宿主免疫状态及感染剂量密切相关。

值得注意的是，该菌株具有血行播散潜能，可能导致菌血症等严重并发症。流行病学数据显示，婴幼儿、老年群体及免疫功能受损患者的致死风险显著升高，病死率可达普通人群的 3~5 倍。其病例数据常表现为整数值时间序列，适合使用一阶整值自回归模型(INAR(1))进行建模。本节针对加拿大魁北克省北部 1990~2000 年弯曲杆菌病例数据[27] (每 28 天记录一次，取 120 个观测值)，其样本均值为 $\bar{X}=10.1833$，样本方差为 $s^2=39.6933$，绘制时间序列图，如图 1 所示，红色三角形标记检测到的离群点，蓝色虚线标识真实异常。

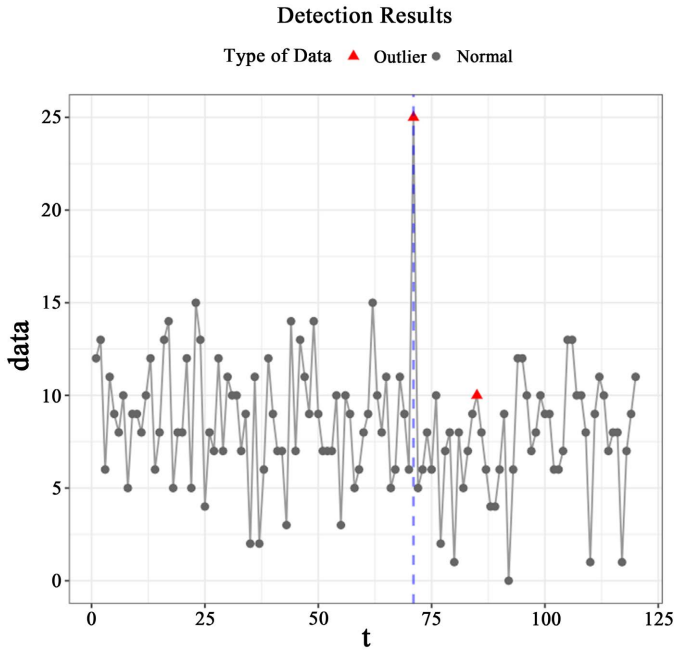

Figure 1. *Campylobacter* data

图 1. 弯曲杆菌数据

由图 1 可以看出在时间 $t=70$ 前病例数围绕某一基线水平(约 8~12 例/周期)波动，表现出弱平稳性，标准差约为 2.5 例，表明感染传播处于相对可控状态。在时间 $t=70$ 后病例数出现显著爆发式增长，峰值达 25 例($t=73$)，较基线水平上升约 150%，且后续持续高位震荡(15~20 例/周期)，表明可能存在离群点，此现象可能由于外部系统性因素导致，如魁北克北部冬季寒冷，可能导致饮用水管道冻裂，引发水源性污染事件，短期内病例数激增，或宿主免疫波动，导致病例异常增加。离群点的存在会导致模型中参数估计存在偏差，以及预测性能下降。

接下来运用 INAR(1)与该数据进行了拟合，并使用经验似然方法检测离群点，展示了检测结果与病例数对比，可以观察到在 $t=70$ 处(对应真实病例激增位置)，经验似然方法成功检测到离群点，除真实离群点外，由于强离群点可能存在掩盖效应，仅在 $t=85$ 被误判为离群点，以上充分说明了经验似然方法对检测离群点的有效性，通过非参数框架、动态约束整合和稳健权重分配，实现了离群点检测的高精度与强解释性。

6. 结论与展望

本研究针对一阶整值自回归模型(INAR(1))的离群点检测与参数估计问题，提出了一套基于改进经验似然法的创新解决方案。针对传统经验似然方法在离散时间序列分析中存在的凸包问题，创新性地引入平衡经验似然(BEL)方法，通过构造对称约束条件有效解决了有限样本下矩方程无解的数值稳定性难题。

通过大量的模拟实验结果表明，经验似然方法可以有效地检测出 INAR(1)模型中的离群点，并且相对于传统的检测方法，经验似然方法对离群点的检测效果更加显著。

然而，当前研究仍存在一些局限性，未来的研究可以从以下几个方向进行拓展和深化：

- 本文主要关注一阶整值时间序列模型，但在实际应用中，许多时间序列数据可能具有更高阶的自相关性或更复杂的动态结构。因此，未来的研究可以考虑将经验似然方法应用于更高阶的整值时间序列模型，如二阶或更高阶的自回归模型(如 INAR(p)模型)，以提高离群值检测的适用性和准确性。
- 随着数据分析的复杂性增加，多变量时间序列数据在金融、气象、生物医学等领域中越来越常见。将经验似然方法扩展到多变量时间序列的离群值检测中，将是一个具有挑战性和应用价值的研究方向。这需要考虑变量之间的相关性以及离群值在多维空间中的表现形式。
- 尽管经验似然方法在离群值检测中表现良好，但其计算复杂度可能在处理大规模数据时成为瓶颈。未来的研究可以探索更高效的算法实现，例如通过并行计算、近似算法或利用机器学习技术来加速经验似然的计算过程，从而提高其在实际应用中的可行性。
- 离群值检测是一个多学科交叉的研究领域，经验似然方法可以与其他先进的统计或机器学习方法相结合，以进一步提高检测性能。例如，将经验似然与深度学习模型(如自编码器、神经网络)相结合，可能会在复杂时间序列数据中发现更隐蔽的离群值模式。

基金项目

辽宁科技大学博士启动基金(6003000310)。

参考文献

[1] Owen, A. (1990) Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, **18**, 90-120. https://doi.org/10.1214/aos/1176347494

[2] Owen, A.B. (1988) Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, **75**, 237-249. https://doi.org/10.1093/biomet/75.2.237

[3] Owen, A. (1991) Empirical Likelihood for Linear Models. *The Annals of Statistics*, **19**, 1725-1747. https://doi.org/10.1214/aos/1176348368

[4] Qin, J. and Lawless, J. (1994) Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, **22**, 300-325. https://doi.org/10.1214/aos/1176325370

[5] Tang, C.Y. and Leng, C. (2010) Penalized High-Dimensional Empirical Likelihood. *Biometrika*, **97**, 905-920. https://doi.org/10.1093/biomet/asq057

[6] Shi, J. and Lau, T. (2000) Empirical Likelihood for Partially Linear Models. *Journal of Multivariate Analysis*, **72**, 132-148. https://doi.org/10.1006/jmva.1999.1866

[7] Wei, C., Luo, Y. and Wu, X. (2010) Empirical Likelihood for Partially Linear Additive Errors-in-Variables Models. *Statistical Papers*, **53**, 485-496. https://doi.org/10.1007/s00362-010-0354-1

[8] Liu, T. and Yuan, X. (2015) Weighted Quantile Regression with Missing Covariates Using Empirical Likelihood. *Statistics*, **50**, 89-113. https://doi.org/10.1080/02331888.2015.1033164

[9] Kitamura, Y. (1997) Empirical Likelihood Methods with Weakly Dependent Processes. *The Annals of Statistics*, **25**, 2084-2102. https://doi.org/10.1214/aos/1069362388

[10] Monti, A. (1997) Empirical Likelihood Confidence Regions in Time Series Models. *Biometrika*, **84**, 395-405. https://doi.org/10.1093/biomet/84.2.395

[11] Chuang, C.S. and Chan, N.H. (2002) Empirical Likelihood for Autoregressive Model, with Applications to Unstable Time Series. *Statistica Sinica*, **12**, 387-407.

[12] Chan, N.H. and Ling, S. (2006) Empirical Likelihood for GARCH Models. *Econometric Theory*, **22**, 403-428. https://doi.org/10.1017/s0266466606060208

[13] Zhao, Z. and Wang, D. (2011) Empirical Likelihood for an Autoregressive Model with Explanatory Variables. *Communications in Statistics—Theory and Methods*, **40**, 559-570. https://doi.org/10.1080/03610920903411267

[14] Nordman, D.J. and Lahiri, S.N. (2014) A Review of Empirical Likelihood Methods for Time Series. *Journal of Statistical Planning and Inference*, **155**, 1-18. https://doi.org/10.1016/j.jspi.2013.10.001

[15] Fox, A.J. (1972) Outliers in Time Series. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **34**, 350-363. https://doi.org/10.1111/j.2517-6161.1972.tb00912.x

[16] Tsay, R.S. (1988) Outliers, Level Shifts, and Variance Changes in Time Series. *Journal of Forecasting*, **7**, 1-20. https://doi.org/10.1002/for.3980070102

[17] Fokianos, K. and Fried, R. (2010) Interventions in INGARCH Processes. *Journal of Time Series Analysis*, **31**, 210-225. https://doi.org/10.1111/j.1467-9892.2010.00657.x

[18] Fried, R., Liboschik, T., Elsaied, H., Kitromilidou, S. and Fokianos, K. (2014) On Outliers and Interventions in Count Time Series Following GLMs. *Austrian Journal of Statistics*, **43**, 181-193. https://doi.org/10.17713/ajs.v43i3.30

[19] Barczy, M., Ispány, M., Pap, G., Scotto, M. and Eduarda Silva, M. (2010) Innovational Outliers in INAR(1) Models. *Communications in Statistics—Theory and Methods*, **39**, 3343-3362. https://doi.org/10.1080/03610920903259831

[20] Barczy, M., Ispány, M., Pap, G., Scotto, M. and Silva, M.E. (2011) Additive Outliers in INAR(1) Models. *Statistical Papers*, **53**, 935-949. https://doi.org/10.1007/s00362-011-0398-x

[21] Silva, M.E. and Pereira, I. (2012) Detection of Additive Outliers in Poisson Integer-Valued Autoregressive Time Series.

[22] Baragona, R., Battaglia, F. and Cucina, D. (2015) Empirical Likelihood for Outlier Detection and Estimation in Autoregressive Time Series. *Journal of Time Series Analysis*, **37**, 315-336. https://doi.org/10.1111/jtsa.12145

[23] Shang, H. and Zhang, B. (2018) Outliers Detection in INAR (1) Time Series. *Journal of Physics*: *Conference Series*, **1053**, Article 012094. https://doi.org/10.1088/1742-6596/1053/1/012094

[24] Chen, J., Variyath, A.M. and Abraham, B. (2008) Adjusted Empirical Likelihood and Its Properties. *Journal of Computational and Graphical Statistics*, **17**, 426-443. https://doi.org/10.1198/106186008x321068

[25] Emerson, S.C. and Owen, A.B. (2009) Calibration of the Empirical Likelihood Method for a Vector Mean. *Electronic Journal of Statistics*, **3**, 1161-1192. https://doi.org/10.1214/09-ejs518

[26] Tsao, M. (2013) Extending the Empirical Likelihood by Domain Expansion. *Canadian Journal of Statistics*, **41**, 257-274. https://doi.org/10.1002/cjs.11175

[27] 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