Advances in Applied Mathematics 应用数学进展, 2026, 15(5), 544-557
Published Online May 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.155250

Hans汉斯

多元零膨胀几何分布的参数估计

连禹晴，卢飞龙

辽宁科技大学理学院，辽宁 鞍山

收稿日期：2026年4月27日；录用日期：2026年5月22日；发布日期：2026年5月28日

摘　要

本文探讨了多元计数数据中常见的零膨胀现象，并提出了一个针对多元零膨胀几何(ZIG)分布的理论与回归框架。该模型能够有效区分结构性零和额外零，适用于多元建模场景。文章系统推导了该分布的联合概率函数、累积分布函数、矩特征以及条件分布性质，基于期望最大化(EM)算法和Fisher评分算法建立了参数估计方法，并讨论了假设检验和置信区间构建问题。模拟研究表明，所提出的估计方法具有良好的有限样本性质。

关键词

多元零膨胀几何分布，EM算法，似然比检验，零膨胀

Parameter Estimation for the Multivariate Zero-Inflated Geometric Distribution

Yuqing Lian, Feilong Lu

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

Received: April 27, 2026; accepted: May 22, 2026; published: May 28, 2026

Abstract

This paper addresses the common zero inflation phenomenon in multivariate count data and proposes a theoretical and regression framework for a multivariate zero-inflated geometric (ZIG) distribution. This model can effectively distinguish between structural zeros and extra zeros and is applicable to multivariate modeling scenarios. The article systematically derived the joint probability function, cumulative distribution function, moment characteristics, and conditional distribution properties of this distribution, established parameter estimation methods based on the EM algorithm and Fisher scoring algorithm, and discussed hypothesis testing and confidence interval construction issues. Simulation studies show that the proposed estimation method has good finite sample properties.

文章引用: 连禹晴, 卢飞龙. 多元零膨胀几何分布的参数估计[J]. 应用数学进展, 2026, 15(5): 544-557.
DOI: 10.12677/aam.2026.155250

Keywords

Multivariate Zero-Inflated Geometric Distribution, EM Algorithm, Likelihood Ratio Test, Zero Inflation

1. 引言

计数数据时间序列分析是统计方法学的基础，在不同科学领域有着广泛的应用。从流行病学到生态学，从金融学到社会科学，研究人员在日常工作中常常会遇到记录固定观测单位内事件发生频率的数据[1]。这类数据的基本特征为非负整数，这些挑战在近几十年里推动了统计学的方法创新。

现实的计数数据时间序列中一个普遍的特征是零膨胀现象，即观测到的零计数比例超过标准离散分布的预期[2]。这一模式在众多领域均有体现：Böhning [3]等发现在医疗保健研究中，大部分患者在研究期间可能没有医疗索赔记录；Welsh [4]等发现在生态学研究中，许多物种在大多数采样点未被检测到；Crépon 和 Duguet [5]发现在经济学中，企业可能多年报告零创新支出。这种现象的普遍性凸显了构建专门统计框架的迫切需求，该框架需能够区分结构性零和额外零。

传统的计数数据模型，如泊松回归和负二项回归，在处理零膨胀数据时表现不佳，因为它们无法适应零生成机制中的这种基本异质性。将这些传统方法应用于零膨胀数据集时，会导致参数估计有偏差、标准误差被低估以及对推断结果产生误导[6]。Min 和 Agresti [7]探讨了针对此类响应变量的重复测量随机效应模型。其中，具有随机效应的障碍模型是一种实用模型，它能分别处理零观测值和正计数。

统计学界通过开发零膨胀模型应对这些挑战，Greene [8]针对计数数据中零值数量超出泊松模型和负二项模型常规预测的情况，提出了两种改进模型。Lambert [9]的开创性工作引入了零膨胀泊松(ZIP)模型，为处理过多零值提供了理论框架和实际估计策略。随后，Ridout [10]将这一方法扩展到零膨胀负二项(ZINB)模型，以同时处理零膨胀和过度离散问题。Bakouch 和 Ristić [11]提出了一种具有零截断泊松边缘分布的一阶平稳整数值自回归过程。Jazi [12]等提出了一种具有零膨胀泊松创新项的新型平稳一阶整数值自回归过程。在这些方法中，几何分布具有独特地位。作为负二项分布在离散参数固定为 1 时的特殊情况，几何分布在理论和实践上都更为简洁。

零膨胀框架与几何分布相结合的研究方法正受到越来越多的关注。Dietz 和 Böhning [13]的早期研究奠定了零膨胀几何(ZIG)分布的理论基础，近期的研究则侧重于参数估计和模型拓展。Srisuradetchai [14]等为 ZIG 模型的参数开发了新的区间估计方法，显示出相较于传统 Wald 区间的优势。Mallick 和 Krishnamoorthy [15]提出了广义零膨胀几何分布并应用于保险数据，Pandya 和 Dey [16]探索了 ZIG 建模的贝叶斯方法。尽管取得了这些进展，但零膨胀几何模型的研究仍存在一些不足。理论发展主要集中在单变量 ZIG 分布，对多变量回归框架的关注不足。Liu 和 Tian [17]基于单变量零膨胀泊松的表示方法提出了一种多元零膨胀泊松(MZIP)分布，用于建模具有额外零值的相关多元计数数据。Zhang 等[18]构建了多元 INAR(1)模型，其创新项可采用多元零膨胀泊松分布或多元零膨胀障碍泊松分布进行建模。虽然关于多元 ZIP 分布已有大量研究，但多元零膨胀几何分布的发展仍然有限。本文构建了一种多元零膨胀几何(MZIG)分布的方法以适用于多元建模场景。

本文的后续内容安排如下：第 2 节建立零膨胀几何分布的理论基础并构建多变量回归框架；第 3 节对所提出的分布进行基于似然的推断和参数估计；第 4 节通过模拟研究评估模型在可控条件下的性能。

2. 多元零膨胀几何分布及其性质

2.1. ZIG 的定义

为了给出多元零膨胀几何分布的定义，下面将对零膨胀几何分布进行简要介绍。若随机变量 Y 服从参数为 π 和 π 的零膨胀几何(ZIG)分布，记为 $Y\sim \mathrm{ZIG}(\pi,\mu)$，则其概率质量函数(pmf)定义为

$$P(Y=y)=\begin{cases}\pi+(1-\pi)\dfrac{1}{1+\mu}, & y=0,\\ (1-\pi)\dfrac{\mu^{y}}{(1+\mu)^{y+1}}, & y=1,2,\cdots\end{cases}$$

其中 $\mu>0$，$0\le\pi<1$。随机变量 Y 的概率生成函数(pgf)为 $\varphi_Y(s)=\dfrac{1+\pi\mu(1-s)}{1+\mu(1-s)}$，且 $|s|<\dfrac{1+\mu}{\mu}$。

Y 的均值与方差为 $E(Y)=\mu(1-\pi)$，$\mathrm{Var}(Y)=\mu(1-\pi)\left[1+\mu(1+\pi)\right]$，$Y$ 的变异系数为 $CV(Y)=\mathrm{Var}(Y)/E(Y)=1+\mu(1+\pi)$，由此可知，$Y\sim\mathrm{ZIG}(\pi,\mu)$ 是过度离散的。关于 ZIG 的定义可参见文献[13]。

定义 1 一个 m 维离散随机向量 $\boldsymbol{Y}=(Y_1,\cdots,Y_m)^{\mathrm{T}}$ 遵循具有特定参数 $\pi\in[0,1)$ 和 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_m)^{\mathrm{T}}\in\mathbb{R}_+^m$ 的多元零膨胀几何分布形式，如果

$$\boldsymbol{Y}\overset{d}{=}Z\boldsymbol{x}=\begin{cases}\boldsymbol{0}, & \text{概率为 }\pi\\ \boldsymbol{x}, & \text{概率为}1-\pi\end{cases}\tag{1}$$

其中 Z 服从参数为 $1-\pi$ 的伯努利分布，且 $\boldsymbol{x}=(X_1,\cdots,X_m)^{\mathrm{T}}$，$X_i$ 服从参数为 μ_i 的几何分布，同时 $(Z,X_1,\cdots,X_m)$ 是相互独立的。我们记作 $\boldsymbol{Y}\sim\mathrm{ZIG}(\pi;\mu_1,\cdots,\mu_m)$ 或者 $\boldsymbol{Y}\sim\mathrm{ZIG}_m(\pi,\boldsymbol{\mu})$ 并且称 $\boldsymbol{x}$ 为 $\boldsymbol{Y}$ 的基础向量。

给定 $Z=1$ 时，公式(1)表明 $\boldsymbol{Y}$ 和 $\boldsymbol{x}$ 具有相同的分布，即 $\boldsymbol{Y}\overset{d}{=}\boldsymbol{x}\,|\,(Z=1)$。这说明 $Z=1$ 时，Y 的每个观测值的生成过程就等价于独立的几何分布，因此它们之间条件独立。

2.2. 联合概率质量函数和累积分布函数

已知 $\boldsymbol{Y}\sim\mathrm{ZIG}_m(\pi,\boldsymbol{\mu})$，如果 $\boldsymbol{y}=\boldsymbol{0}_m$，则 $\boldsymbol{y}$ 的联合概率质量函数(pmf)为

$$\begin{aligned}f(\boldsymbol{y}\,|\,\pi,\boldsymbol{\mu})&=\Pr(ZX_1=0,\cdots,ZX_m=0)\\&=\Pr(Z=0)+\Pr(Z=1,X_1=0,\cdots,X_m=0)\\&=\pi+(1-\pi)\prod_{i=1}^{m}\frac{1}{1+\mu_i}.\end{aligned}\tag{2}$$

如果 $\boldsymbol{y}\ne\boldsymbol{0}_m$，可以得到

$$\begin{aligned}f(\boldsymbol{y}\,|\,\pi,\boldsymbol{\mu})&=\Pr(ZX_1=y_1,\cdots,ZX_m=y_m)\\&=\Pr(Z=1,X_1=y_1,\cdots,X_m=y_m)\\&=(1-\pi)\prod_{i=1}^{m}\frac{\mu_i^{y_i}}{(1+\mu_i)^{y_i+1}}.\end{aligned}\tag{3}$$

将公式(2)和公式(3)结合起来，可以得到 $\boldsymbol{y}$ 的联合概率质量函数为

$$\begin{aligned} f\left(\boldsymbol{y}\mid\pi,\boldsymbol{\mu}\right) &= \left[\pi+\left(1-\pi\right)\prod_{i=1}^{m}\frac{1}{1+\mu_i}\right]I\left(\boldsymbol{y}=\boldsymbol{0}\right) \\ &\quad+\left[\left(1-\pi\right)\prod_{i=1}^{m}\frac{\mu_i^{y_i}}{\left(1+\mu_i\right)^{y_i+1}}\right]I\left(\boldsymbol{y}\neq\boldsymbol{0}\right) \\ &= \pi\Pr\left(\boldsymbol{\xi}=\boldsymbol{y}\right)+\left(1-\pi\right)\Pr\left(\boldsymbol{x}=\boldsymbol{y}\right). \end{aligned} \tag{4}$$

其中 $\boldsymbol{\xi}=\left(\xi_1,\cdots,\xi_m\right)^{\mathrm{T}}$，$\left\{\xi_i\right\}_{i=1}^{m}\overset{\text{i.i.d.}}{\sim}\text{Degenerate}\left(0\right)$。对于任意非负实向量 $\boldsymbol{y}=\left(y_1,\cdots,y_m\right)^{\mathrm{T}}$，$\boldsymbol{y}$ 的联合累积分布函数由下式给出

$$\begin{aligned} \Pr\left(\boldsymbol{Y}\le\boldsymbol{y}\right) &= \pi\Pr\left(\boldsymbol{\xi}=\boldsymbol{0}\right)+\left(1-\pi\right)\Pr\left(\boldsymbol{x}\le\boldsymbol{y}\right) \\ &= \pi\cdot 1+\left(1-\pi\right)\prod_{i=1}^{m}\Pr\left(X_i\le y_i\right) \\ &= \pi+\left(1-\pi\right)\prod_{i=1}^{m}\left[1-\left(\frac{\mu_i}{1+\mu_i}\right)^{\lfloor y_i\rfloor+1}\right]. \end{aligned}$$

其中 $\lfloor y_i\rfloor$ 表示不大于 y_i 的最大整数。

2.3. 混合矩和矩母函数

由公式(1)，可以得到 $E\left(\boldsymbol{y}\right)=\left(1-\pi\right)\boldsymbol{\mu}$，$E\left(\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\right)=\left(1-\pi\right)\left[\text{diag}\left(\mu_i\left(1+\mu_i\right)\right)+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right]$，$\text{Var}\left(\boldsymbol{y}\right)=\left(1-\pi\right)\text{diag}\left(\mu_i\left(1+\mu_i\right)\right)+\pi\left(1-\pi\right)\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$。由此可知

$$\text{Corr}\left(Y_i,Y_j\right)=\frac{\pi\left(1-\pi\right)\mu_i\mu_j}{\sqrt{\text{Var}\left(Y_i\right)\text{Var}\left(Y_j\right)}}\quad\left(i\neq j\right),$$

其中 $\text{Var}\left(Y_i\right)=\mu_i\left(1-\pi\right)\left[1+\mu_i\left(1+\pi\right)\right]$。由于 $X_i\sim\text{Geometric}\left(\mu_i\right)$，其 n 阶矩为 $E\left(X_n\right)=\sum_{k=1}^{n}\left\{{n\atop k}\right\}\mu^k$，其中 $\left\{{n\atop k}\right\}=\frac{1}{k!}\sum_{j=0}^{k}\left(-1\right)^{k-j}\binom{k}{j}j^n$ 是第二类 Stirling 数。因此，对于任何 $r_1,\cdots,r_m\ge 0$，$\boldsymbol{Y}$ 的混合矩由下式给出

$$E\left(\prod_{i=1}^{m}Y_i^{r_i}\right)=\left(1-\pi\right)\prod_{i=1}^{m}\sum_{k=1}^{r_i}\left\{{r_i\atop k}\right\}\mu_i^k.$$

通过使用公式 $E\left(\xi\right)=E\left[E\left(\xi\mid\eta\right)\right]$，则 $\boldsymbol{Y}\sim\text{ZIG}\left(\pi;\mu_1,\cdots,\mu_m\right)$ 的矩母函数(mgf)由以下公式给出

$$\begin{aligned} M_Y\left(t_1,\cdots,t_m\right) &= \mathbb{E}\left(\exp\left(t_1Y_1+\cdots+t_mY_m\right)\right) \\ &= \pi+\left(1-\pi\right)\prod_{i=1}^{m}\frac{1}{1+\mu_i}+\left(1-\pi\right)\sum_{y\neq 0}\exp\left(t_1y_1+\cdots+t_my_m\right)\prod_{i=1}^{m}\frac{\mu_i^{y_i}}{\left(1+\mu_i\right)^{y_i+1}} \\ &= \pi+\left(1-\pi\right)\prod_{i=1}^{m}\frac{1}{1+\mu_i}+\left(1-\pi\right)\left[\prod_{i=1}^{m}\frac{1}{1+\mu_i-\mu_i\mathrm{e}^{t_i}}-\prod_{i=1}^{m}\frac{1}{1+\mu_i}\right] \\ &= \pi+\left(1-\pi\right)\prod_{i=1}^{m}\frac{1}{1+\mu_i-\mu_i\mathrm{e}^{t_i}}. \end{aligned}$$

2.4. 边际分布

已知$\boldsymbol{Y}\sim \mathrm{ZIG}\left(\pi;\mu_1,\cdots,\mu_m\right)$。将$\boldsymbol{Y}$分成两部分$y=\begin{pmatrix}y^{(1)}\\ y^{(2)}\end{pmatrix}$，其中$\boldsymbol{y}^{(1)}=\begin{pmatrix}Y_1\\ \vdots\\ Y_r\end{pmatrix}$，$\boldsymbol{y}^{(2)}=\begin{pmatrix}Y_{r+1}\\ \vdots\\ Y_m\end{pmatrix}$。用同样的方式划分$x$。根据定义 1 可知：

$$\begin{cases}\boldsymbol{y}^{(1)}\overset{d}{=}Z\boldsymbol{x}^{(1)}\sim \mathrm{ZIG}\left(\pi;\mu_1,\cdots,\mu_r\right)\\ \boldsymbol{y}^{(2)}\overset{d}{=}Z\boldsymbol{x}^{(2)}\sim \mathrm{ZIG}\left(\pi;\mu_{r+1},\cdots,\mu_m\right)\end{cases}.$$

事实上，对于任何正整数$i_1,\cdots,i_r$满足$1\le i_1<\cdots<i_r\le m$，可以得到

$$\begin{pmatrix}Y_{i_1}\\ \vdots\\ Y_{i_r}\end{pmatrix}\overset{d}{=}Z\begin{pmatrix}X_{i_1}\\ \vdots\\ X_{i_r}\end{pmatrix}\sim \mathrm{ZIG}\left(\pi;\mu_{i_1},\cdots,\mu_{i_r}\right).$$

因此可以证明多元 ZIG 分布的每个边际分布都是一元 ZIG 分布。

2.5. 条件分布

2.5.1. $y^{(1)}\mid y^{(2)}$的条件分布

根据条件概率公式：

$$\begin{aligned}f\left(y^{(1)}\mid y^{(2)}\right)&=\frac{f\left(y\mid\pi,\mu\right)}{\Pr\left\{y^{(2)}=y^{(2)}\right\}}\\&=\frac{\left[\pi+\left(1-\pi\right)\prod_{i=1}^{m}\frac{1}{1+\mu_i}\right]I\left(\boldsymbol{y}=0\right)+\left[\left(1-\pi\right)\prod_{i=1}^{m}\frac{\mu_i^{y_i}}{\left(1+\mu_i\right)^{y_i+1}}\right]I\left(\boldsymbol{y}\ne 0\right)}{\left[\pi+\left(1-\pi\right)\prod_{i=r+1}^{m}\frac{1}{1+\mu_i}\right]I\left(y^{(2)}=0\right)+\left[\left(1-\pi\right)\prod_{i=r+1}^{m}\frac{\mu_i^{y_i}}{\left(1+\mu_i\right)^{y_i+1}}\right]I\left(y^{(2)}\ne 0\right)}.\end{aligned}$$

在两种不同的情况下讨论这个问题。

情况一：$y^{(2)}=0$。此时$y^{(1)}$可能为 0 或不为 0，当$y^{(1)}=0$时：

$$f\left(y^{(1)}=0\mid y^{(2)}=0\right)=\frac{\pi+\left(1-\pi\right)\prod_{i=1}^{r}\frac{1}{1+\mu_i}\cdot\prod_{i=r+1}^{m}\frac{1}{1+\mu_i}}{\pi+\left(1-\pi\right)\prod_{i=r+1}^{m}\frac{1}{1+\mu_i}}=\pi^*+\left(1-\pi^*\right)\prod_{i=1}^{r}\frac{1}{1+\mu_i}.$$

其中$\pi^*=\dfrac{\pi}{\pi+\left(1-\pi\right)\dfrac{1}{1+\mu_2}}$，当$y^{(1)}\ne 0$时：

$$f\left(y^{(1)}\ne 0\mid y^{(2)}=0\right)=\frac{\left(1-\pi\right)\prod_{i=r+1}^{m}\frac{1}{1+\mu_i}\cdot\prod_{i=1}^{r}\frac{\mu_i^{y_i}}{\left(1+\mu_i\right)^{y_i+1}}}{\pi+\left(1-\pi\right)\prod_{i=r+1}^{m}\frac{1}{1+\mu_i}}=\left(1-\pi^*\right)\prod_{i=1}^{r}\frac{\mu_i^{y_i}}{\left(1+\mu_i\right)^{y_i+1}}.$$

将上面两个式子结合可知条件分布仍然是一个零膨胀参数为π^*的多元 ZIG 分布。

情况二： $y^{(2)} \neq 0$ 。可以得到： $\Pr\left\{y^{(1)} = y^{(1)} \mid y^{(2)} = y^{(2)}\right\} = \prod_{i=1}^{r} \frac{\mu_i^{y_i}}{(1+\mu_i)^{y_i+1}}$ 。这表示 $y^{(1)} \mid y^{(2)} = x^{(1)}$ 不依赖于 Z 。

2.5.2. $Z \mid y$ 的条件分布

由于 $Z \sim \text{Bernoulli}(1-\pi)$， Z 仅取值 0 或 1。

当 $\boldsymbol{y} = \mathbf{0}$ 时，即所有分量 $y_i = 0$ 。因此：

$$P(Z=1 \mid \boldsymbol{y}=\mathbf{0}) = \frac{\pi \cdot 1}{\pi + (1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}},\ P(Z=0 \mid \boldsymbol{y}=\mathbf{0}) = \frac{(1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}}{\pi + (1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}}.$$

由此可知当所有观测值为零时， $Z \mid y$ 的条件概率取决于参数 π 和 μ_i 。

当 $\boldsymbol{y} \neq \mathbf{0}$ (即至少一个分量 $y_i > 0$)时，可以得到

$$P(Z=1 \mid \boldsymbol{y} \neq \mathbf{0}) = 0;\quad P(Z=0 \mid \boldsymbol{y} \neq \mathbf{0}) = 1.$$

由此可知当观测值 y 不全为零时，那么它一定来自几何分布部分(即 $Z=0$)，因此 Z 的条件分布是确定的。

2.5.3. $x \mid y$ 的条件分布

给定观测值 y ，潜变量 x 的条件分布为：

情况一：当 $y \neq 0$ 时，如果观测值不全为零，那么它一定来自几何分布部分(即 $Z=0$)，因此：

$$P(\boldsymbol{x} = \boldsymbol{y} \mid \boldsymbol{y} \neq \mathbf{0}) = 1;\quad P(\boldsymbol{x} \neq \boldsymbol{y} \mid \boldsymbol{y} \neq \mathbf{0}) = 0.$$

这是一个退化分布，集中在观测值 y 上。

情况二：当 $y = 0$ 时，则全零观测值可能来自零膨胀部分或几何分布部分，若 $x = 0$ 则条件分布为：

$$P(\boldsymbol{x}=0 \mid \boldsymbol{y}=\mathbf{0}) = \frac{(1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}}{\pi + (1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}}$$

若 $x \neq 0$ 则条件分布为： $P(x \neq 0 \mid y = 0) = 0$ 。这也是一个退化分布，但只集中在 0 上，概率反映了观测到的全零值来自几何分布部分的可能性。

3. 基于似然的推断

若 $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ 是从 m 维零膨胀几何分布 $\text{ZIG}(\pi; \mu_1, \cdots, \mu_m)$ 中抽取出的 n 个样本，其中 $\boldsymbol{y}_j = (Y_{1j}, \cdots, Y_{mj})^{\mathsf{T}}$ ，且 $j = 1, \cdots, n$ 。并且记 $\mathcal{Y}_{\text{obs}} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ 为观测数据。其中 $\mathcal{J} = \{j \mid \boldsymbol{y}_j = \mathbf{0}, j = 1, \cdots, n\}$ 表示集合 $\mathcal{J}$ 中的元素数量，且 $m_0 = \sum_{j=1}^{n} I(\boldsymbol{y}_j = \mathbf{0})$ 表示零的数量。可得观测数据似然函数为

$$L(\pi, \boldsymbol{\mu} \mid \mathcal{Y}_{\text{obs}}) \propto \left[\pi + (1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}\right]^{m_0} \times (1-\pi)^{n-m_0} \prod_{i=1}^{m} \frac{\mu^{N_i}}{(1+\mu)^{N_i+(n-m_0)}}.$$

其中 $N_i \triangleq \sum_{j \notin \mathcal{J}} y_{ij} = \sum_{j=1}^{n} y_{ij}$ ，所以对数似然函数为

$$\ell(\pi, \boldsymbol{\mu} \mid \mathcal{Y}_{\text{obs}}) = m_0 \log\left[\pi + (1-\pi)\prod_{i=1}^{m} \frac{1}{1+\mu_i}\right] + (n-m_0)\log(1-\pi)$$
$$+ \sum_{i=1}^{m}\left(N_i \log \mu_i - (N_i + n - m_0)\log(1+\mu_i)\right).$$

3.1. 基于 Fisher 评分算法的 MLE

为了使用 Fisher 评分算法估计参数 π 和 μ，首先需要计算出对数似然函数关于参数 π 和 μ_i 的导数，即计算出 Score 向量 $\nabla\ell(\pi,\mu|Y_{\text{obs}})$ 和 Hessian 矩阵 $\nabla^2\ell(\pi,\mu|Y_{\text{obs}})$。其中 Score 向量 $\nabla\ell(\pi,\mu|Y_{\text{obs}})$ 中的元素有 $\frac{\partial\ell}{\partial\pi}$ 和 $\frac{\partial\ell}{\partial\mu_k}$，Hessian 矩阵 $\nabla^2\ell(\pi,\mu|Y_{\text{obs}})$ 中的元素有 $\frac{\partial^2\ell}{\partial\pi^2}$、$\frac{\partial^2\ell}{\partial\mu_k^2}$、$\frac{\partial^2\ell}{\partial\pi\partial\mu_k}$ 和 $\frac{\partial^2\ell}{\partial\mu_k\partial\mu_l}$。

对于 $i,k=1,\cdots,m$ 和 $i\neq k$，通过将上述二阶偏导数中的 m_0 和 N_i 分别替换为它们的期望值 $n\left(\pi+(1-\pi)\prod_{i=1}^{m}\frac{1}{1+\mu_i}\right)$ 和 $n(1-\pi)\mu_i$，我们可以计算出 Fisher 信息矩 $J(\pi,\mu)=\mathbb{E}\left\{-\nabla^2\ell(\pi,\mu|Y_{\text{obs}})|\pi,\mu\right\}$。设 $\left(\pi^{(0)},\mu^{(0)}\right)$ 为 MLE 的初始值。如果 $\left(\pi^{(t)},\mu^{(t)}\right)$ 表示 $(\hat{\pi},\hat{\mu})$ 的第 t 次近似，则通过 Fisher 评分算法可以得到它们的第 $t+1$ 次近似值为

$$\begin{pmatrix}\pi^{(t+1)}\\ \mu^{(t+1)}\end{pmatrix}=\begin{pmatrix}\pi^{(t)}\\ \mu^{(t)}\end{pmatrix}+J^{-1}\left(\pi^{(t)},\mu^{(t)}\right)\nabla\ell\left(\pi^{(t)},\mu^{(t)}|Y_{\text{obs}}\right).$$

其中 Fisher 信息矩阵是估计量协方差矩阵的逆矩阵。设参数向量为 $\theta=(\pi,\mu_1,\mu_2,\cdots,\mu_m)$，则协方差矩阵为 $\text{Cov}\left(\hat{\theta}\right)=J\left(\hat{\theta}\right)^{-1}$。

这里，$\hat{\theta}$ 是通过最大似然估计得到的参数向量，$J\left(\hat{\theta}\right)$ 是估计的 Fisher 信息矩阵。参数估计 $\hat{\theta}$ 的标准误差是协方差矩阵 $J\left(\hat{\theta}\right)^{-1}$ 的对角线元素 j^{kk} 的平方根：$SE\left(\hat{\theta}\right)=\sqrt{\text{Cov}\left(\hat{\theta}\right)}$。因此，$\pi$ 和 $\{\mu_i\}_{i=1}^{m}$ 的 $(1-\alpha)100\%$ 渐近 Wald 置信区间(CI)由下式给出

$$\left[\hat{\pi}-z_{\alpha/2}\sqrt{\frac{1}{n}},\hat{\pi}+z_{\alpha/2}\sqrt{\frac{1}{n}}\right],\left[\hat{\mu}_i-z_{\alpha/2}\sqrt{\frac{1}{n_i+1}},\hat{\mu}_i+z_{\alpha/2}\sqrt{\frac{1}{n_i+1}}\right].$$

其中 z_α 表示标准正态分布的第 α 个上分位数。

3.2. 基于 EM 算法的 MLE

多元零膨胀几何分布的观测零向量可分为两类：一类是由于种群变异性而在零点处退化分布所产生的额外零向量；另一类是来自独立的普通几何分布的结构零向量。因此，可以将 $\mathcal{J}$ 划分为 $\mathcal{J}_{\text{extra}}$ 和 $\mathcal{J}_{\text{structural}}$。然后，我们用表示 $\mathcal{J}_{\text{extra}}$ 的数目的潜变量 W 来增加 Y_{obs}，以将 m_0 拆分成 W 和 m_0-W 两部分。在给定 Y_{obs} 和 (π,μ) 的情况下，得到的 W 的条件预测分布是

$$W|(Y_{\text{obs}},\pi,\mu)\sim\text{Binomial}\left(m_0,\frac{\pi}{\pi+(1-\pi)\prod_{i=1}^{m}\frac{1}{1+\mu_i}}\right).$$

另一方面，完整数据似然函数为

$$L(\pi,\mu|Y_{\text{com}},W)=\pi^w(1-\pi)^{n-w}\times\prod_{i=1}^{m}\frac{\mu_i^{N_i}}{(1+\mu_i)^{N_i+1}}.$$

其中 $N_i\triangleq\sum_{j\notin\mathcal{J}}y_{ij}=\sum_{j=1}^{n}y_{ij}$，基于潜变量 W 的完整数据对数似然函数为

$$\ell_c(\pi,\mu|Y_{\text{com}},W)=w\log\pi+(n-w)\log(1-\pi)+\sum_{i=1}^{m}\left[N_i\log\mu_i-(N_i+1)\log(1+\mu_i)\right].$$

这里，W 是来自零膨胀部分的全零观测值的数量，$n-W$ 是来自几何分布部分的观测值的数量(包括全零和非全零观测值)。因此，E 步就是将上面表达式中的 w 替换为其条件期望

$$E\left(W \mid Y_{\text{obs}}, \pi, \mu\right)=\frac{m_0 \pi}{\pi+(1-\pi)\prod_{i=1}^{m}\frac{1}{1+\mu_i}}.$$

M 步骤是找到完整数据的极大似然估计值：

$$\hat{\pi}=\frac{w}{n}\text{；}\quad \hat{\mu}_i=\frac{N_i}{n-w}, i=1,\cdots,m\text{。}$$

3.3. 小样本量的 Bootstrap 置信区间

对于小样本量，Bootstrap 方法是一个有用的工具，可以找到 π 和 $\{\mu_i\}_{i=1}^m$ 的任意函数的置信区间，例如 $\vartheta=h\left(\pi,\mu_1,\cdots,\mu_m\right)$。设 $\hat{\vartheta}=h\left(\hat{\pi},\hat{\mu}_1,\cdots,\hat{\mu}_m\right)$ 表示 ϑ 的 MLE，其中 $\hat{\pi}$ 和 $\{\hat{\mu}_i\}_{i=1}^m$ 分别表示 π 和 $\{\mu_i\}_{i=1}^m$ 通过 EM 算法计算得到的 MLE。基于获得的 π 和 $\{\hat{\mu}_i\}_{i=1}^m$，可以生成 $y_1^*,\cdots,y_n^* \overset{\text{i.i.d.}}{\sim} \text{ZIG}^{(1)}\left(\hat{\pi},\hat{\mu}_1,\cdots,\hat{\mu}_m\right)$，得到 $Y_{\text{obs}}^*=\left\{y_1^*,\cdots,y_n^*\right\}$，然后计算 Bootstrap 重复 $\hat{\pi}^*$ 和 $\{\hat{\mu}_i^*\}_{i=1}^m$，得到 $\hat{\vartheta}^*=h\left(\hat{\pi}^*,\hat{\mu}_1^*,\cdots,\hat{\mu}_m^*\right)$。独立地重复这个过程 G 次，我们得到 G 个 Bootstrap 复制 $\{\hat{\vartheta}_g^*\}_{g=1}^G$。因此，$se\left(\hat{\vartheta}\right)$ 的标准误差 $\hat{\vartheta}$ 可以通过 G 次重复的样本标准差来估计，即

$$\widehat{se}\left(\hat{\vartheta}\right)=\left\{\frac{1}{G-1}\sum_{g=1}^{G}\left[\hat{\vartheta}_g^*-\frac{\hat{\vartheta}_1^*+\cdots+\hat{\vartheta}_G^*}{G}\right]^2\right\}^{1/2}$$

如果 $\{\hat{\vartheta}_g\}_{g=1}^G$ 是近似正态分布，则 ϑ 的第一个 $(1-\alpha)100\%$ Bootstrap 置信区间为

$$\left[\hat{\vartheta}-Z_{\alpha/2}\cdot\widehat{se}\left(\hat{\vartheta}\right),\hat{\vartheta}+Z_{\alpha/2}\cdot\widehat{se}\left(\hat{\vartheta}\right)\right]$$

或者，如果 $\{\hat{\vartheta}_g\}_{g=1}^G$ 是非正态分布，则 ϑ 的第二个 $(1-\alpha)100\%$ Bootstrap 置信区间可以通过 $\left[\hat{\vartheta}_L,\hat{\vartheta}_U\right]$ 获得，其中 $\hat{\vartheta}_L$ 和 $\hat{\vartheta}_U$ 分别是 $\{\hat{\vartheta}_g^*\}_{g=1}^G$ 的 $100(\alpha/2)$ 和 $100(1-\alpha/2)$ 百分位数。

3.4. 大样本量的假设检验

3.4.1. 零膨胀的似然比检验

如果要检验原假设 $H_0:\pi=0$ 与备择假设 $H_1:\pi>0$，在 H_0 下，似然比检验统计量为

$$T_1=-2\left(\ell\left(0,\hat{\mu}_0 \mid Y_{\text{obs}}\right)-\ell\left(\pi,\hat{\mu} \mid Y_{\text{obs}}\right)\right).$$

其中 $\hat{\mu}_0=\left(\sum_{j=1}^{n}y_{1j}/n,\cdots,\sum_{j=1}^{n}y_{mj}/n\right)^{\mathrm{T}}$ 是 H_0 下 μ 的 MLE，$(\hat{\mu},\hat{\pi})$ 是 (μ,π) 的无约束 MLE。由于零假设对应于 π 位于参数空间的边界上，并且适合的参考分布是 χ^2 分布的混合[19] [20]。Jansakul 和 Hinde [21]指出 T_1 的参考分布是 χ_0^2 (零处的常数)和 χ_1^2 分布的均等混合，因此相应的 p 值为

$$p_{v_1}=\frac{1}{2}\Pr\left(T_1>t_1 \mid H_0\right)=\frac{1}{2}\Pr\left(\chi^2(1)>t_1\right)\text{。}$$

3.4.2. 零膨胀的得分检验

在本小节中，通过得分检验来重新参数化来测试多元 ZIG 模型中的零膨胀现象。让 $\theta=\frac{\pi}{1-\pi}$ 和 $\boldsymbol{\beta}=\left(\beta_1,\cdots,\beta_m\right)^{\top}=\left(\log\mu_1,\cdots,\log\mu_m\right)^{\top}$，那么测试 H_0 相当于测试 $H_0^*:\theta=0$。观测数据对数似然函数现在变为

$$\ell\left(\beta,\theta\,|\,\mathcal{Y}_{\text{obs}}\right)=-n\log\left(1+\theta\right)+\sum_{j=1}^{n}\left[I\left(y_j=0\right)\log\left(\theta+\prod_{i=1}^{m}\frac{1}{1+\mathrm{e}^{\beta_i}}\right)\right]$$
$$+\sum_{j=1}^{n}\left[I\left(y_j\neq 0\right)\sum_{i=1}^{m}\left(N_i\beta_i-\left(N_i+1\right)\log\left(1+\mathrm{e}^{\beta_i}\right)\right)\right].$$

得分向量是

$$\boldsymbol{U}\left(\beta,\theta\right)=\left(\frac{\partial\ell\left(\beta,\theta\,|\,\mathcal{Y}_{\text{obs}}\right)}{\partial\beta_1},\cdots,\frac{\partial\ell\left(\beta,\theta\,|\,\mathcal{Y}_{\text{obs}}\right)}{\partial\beta_m},\frac{\partial\ell\left(\beta,\theta\,|\,\mathcal{Y}_{\text{obs}}\right)}{\partial\theta}\right).$$

Fisher 信息矩阵为 $J\left(\beta,\theta\right)=\left(J_{ik}\right)=\mathbb{E}\left[I\left(\beta,\theta\,|\,Y_{\text{obs}}\right)\right]$。在 H_0^* 下，得分检验统计量为

$$T_2=U^{\top}\left(\hat{\beta}_0,\hat{\theta}_0\right)J^{-1}\left(\hat{\beta}_0,\hat{\theta}_0\right)U\left(\hat{\beta}_0,\hat{\theta}_0\right)\sim\chi^2\left(1\right).$$

其中 $\hat{\theta}_0=0$ 和 $\hat{\beta}_0=\left(\log\left(\frac{1}{n}\sum_{j=1}^{n}y_{1j}\right),\cdots,\log\left(\frac{1}{n}\sum_{j=1}^{n}y_{mj}\right)\right)^{\top}$ 表示 H_0^* 下 β 的极大似然估计值，且 $m_0=\sum_{j=1}^{n}I\left(y_j=0\right)$。相应的 p 值由下式给出

$$p_{v2}=\Pr\left(T_2>t_2\,|\,H_0\right)=\frac{1}{2}\Pr\left(\chi_1^2\left(1\right)>t_2\right)。$$

4. 模拟研究

4.1. 参数估计的模拟

在本节中，为了验证估计方法在有限样本下的效果，做了以下五组模拟：

(a) $\pi=0.1,\mu_1=0.1,\mu_2=0.2$；(b) $\pi=0.2,\mu_1=0.3,\mu_2=1$；(c) $\pi=0.3,\mu_1=0.5,\mu_2=2$；(d) $\pi=0.4,\mu_1=1,\mu_2=3$ 和(e) $\pi=0.5,\mu_1=2,\mu_2=5$。考虑大小为 50、100、500、1000 和 5000 的子样本。参数的估计值及其标准误差的数值结果如表 1 所示。

Table 1. Numerical results of parameter estimates and associated standard errors
表 1. 参数的估计值及其标准误差的数值结果

参数	样本量	$\hat{\pi}$	$\hat{\mu}_1$	$\hat{\mu}_2$	Std($\hat{\pi}$)	Std($\hat{\mu}_1$)	Std($\hat{\mu}_2$)
(a)	50	0.1815	0.1203	0.2398	0.2181	0.0613	0.1300
	100	0.1538	0.1177	0.2153	0.1815	0.0486	0.0749
	500	0.1236	0.1040	0.2092	0.1149	0.0206	0.0399
	1000	0.1099	0.1025	0.2072	0.0850	0.0147	0.0276
	5000	0.0982	0.1004	0.1997	0.0492	0.0069	0.0139
(b)	50	0.1887	0.3196	1.0364	0.1189	0.1179	0.2662
	100	0.1737	0.2951	0.9616	0.1009	0.0753	0.1857
	500	0.1840	0.2891	0.9719	0.0506	0.0373	0.0937

续表

	1000	0.2035	0.3001	1.0103	0.0291	0.0214	0.0560
	5000	0.2025	0.3018	1.0033	0.0144	0.0111	0.0289
(c)	50	0.2737	0.4803	1.9369	0.1106	0.1590	0.4793
	100	0.2963	0.4965	2.0178	0.0660	0.1149	0.3155
	500	0.3017	0.5086	2.0170	0.0345	0.0486	0.1390
	1000	0.3017	0.5039	2.0078	0.0232	0.0413	0.0962
	5000	0.2997	0.4983	2.0015	0.0099	0.0146	0.0461
(d)	50	0.4020	1.0100	3.0581	0.0895	0.3143	0.6710
	100	0.3956	0.9867	2.9900	0.0638	0.1756	0.5017
	500	0.3963	0.9941	3.0163	0.0268	0.0893	0.2204
	1000	0.4020	1.0050	3.0093	0.0186	0.0560	0.1620
	5000	0.4006	1.0020	2.9999	0.0079	0.0560	0.0693
(e)	50	0.5072	1.9448	4.8357	0.0833	0.5421	1.0323
	100	0.5068	2.0276	5.1379	0.0563	0.3658	0.8407
	500	0.5007	1.9810	5.1294	0.0214	0.1533	0.3910
	1000	0.4991	2.0063	4.9908	0.0163	0.1229	0.2368
	5000	0.4994	1.9971	4.9922	0.0080	0.0540	0.1138

由表 1 的结果可知，随着样本量的增加，所有参数的估计值逐渐趋近于其真实值，表明估计量具有良好的一致性。例如，在(a)中，当样本量从 50 增至 5000 时，$\hat{\pi}$ 从 0.1815 收敛至 0.0982 (接近真实值 0.1)，$\hat{\mu}_1$ 从 0.1203 收敛至 0.1004，$\hat{\mu}_2$ 从 0.2398 收敛至 0.1997，均表现出明显的收敛趋势。类似地，在其他参数设定下也观察到一致的行为。此外，估计的标准差随样本量增大而显著减小，说明估计精度随样本规模提升而改善。例如，在(e)中，当样本量由 50 增至 5000 时，Std($\hat{\pi}$)从 0.0883 降至 0.0080，Std($\hat{\mu}_1$)从 0.5421 降至 0.0540，Std($\hat{\mu}_2$)从 1.0323 降至 0.1138，这体现了估计量的渐近有效性。尽管在小样本情况下(如 $n=50$)存在一定的估计偏倚，但随着样本量增长，偏倚迅速减小，验证了估计方法在大样本下的优良统计性质。不难得出结论，在大多数情况下，EM 方法表现出最好的性能。

4.2. 零膨胀检验的模拟

在本小节中，通过模拟比较不同样本量的 LRT 和得分检验之间相应的错误率($H_0:\pi=0$ 时)和功效($H_1:\pi>0$ 时)，其中 H_1 中的 π 值选择为 0.01、0.02、0.03、0.04、0.05、0.06、0.08、0.10。样本大小设置为 $n=50(50)500$。对于给定的 (n,π)，首先设 $Z_1,\cdots,Z_n \sim \text{Bernoulli}(1-\pi)$ 对于 $i=1,\cdots,L$ $(L=1000)$，然后独立生成

$$X_{11}^{(l)},\cdots,X_{1n}^{(l)} \sim \text{Geometric}(\mu_1),$$

$$X_{21}^{(l)},\cdots,X_{2n}^{(l)} \sim \text{Geometric}(\mu_2),$$

$$y_j^{(l)}=\begin{pmatrix}Y_{1j}^{(l)}\\ Y_{2j}^{(l)}\end{pmatrix}=Z_j^{(l)}\begin{pmatrix}X_{1j}^{(l)}\\ X_{2j}^{(l)}\end{pmatrix},\quad j=1,\cdots,n\ 。$$

当 $\mu_1=1$ 且 $\mu_2=3$ 时，所有假设检验均在显著性水平 $\alpha=0.05$ 下进行。令 r_k 表示分别通过检验统计量 $T_k\ (k=1,2)$ 拒绝原假设 $H_0:\pi=0$ 的次数。因此，当 $\pi=0$ 时实际显著性水平可以通过 r_k/L 来估计。当 $\pi>0$ 时检验统计量 T_k 的功效可以通过 r_k/L 估计。

如图 1 所示，可以观察到 LRT(实线)和得分检验(虚线)结果非常接近。LRT 测试在将错误率控制在预先选择的标准水平附近具有良好的性能。如图 2 所示，可以观察到当 $0.00<\pi<0.08$ 时，LRT (实线)总是比得分检验(虚线)稍强一些。当 $\pi=0.1$ 且样本量大于 200 时，两个检验的功效几乎相同。

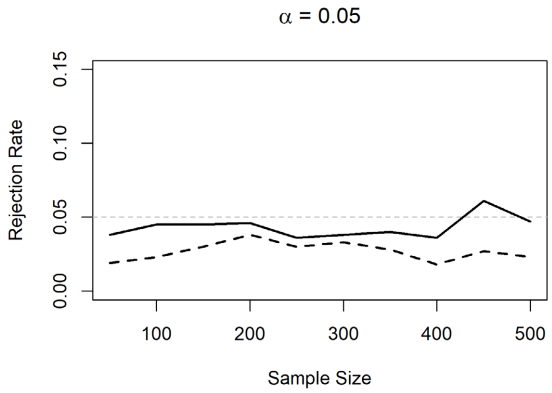

Figure 1. Comparison of error rates between the likelihood ratio test (solid line) and the score test (dotted line) (m = 2)
图 1. 似然比检验(实线)与评分检验(虚线)的错误率比较(m = 2)

5. 实例分析

本节的实例分析基于 Li (1999) [22]等的真实三元计数数据集，其中包含 72 个观察样本。约 83.3%的样本在三个变量中同时取零值，并且非零观测值经常同时出现在多个变量中，这表明数据存在明显的零膨胀性且变量之间存在内在相关性。令 $y_1,\cdots,y_n\sim \mathrm{ZIG}(\pi;\mu_1,\mu_2,\mu_3)$，其中 $y_j=(Y_{1j},Y_{2j},Y_{3j})$，$j=1,\cdots,n$ $(n=72)$。参数的最大似然估计值及置信区间如表 2 所示。

Table 2. The estimated values, standard errors, and confidence intervals of the parameters
表 2. 参数的估计值与标准误差及其置信区间

参数	估计值	Std	Bootstrap CIs
π	0.8871	0.0438	[0.7862, 0.9545]
μ_1	0.123	0.1133	[0.05, 0.4091]
μ_2	1.7219	0.5665	[0.4922, 2.6987]
μ_3	1.3529	0.8156	[0.0819, 3.0343]

零膨胀参数 π 的估计值为 0.8871，表明数据中存在显著的零膨胀现象。Bootstrap 置信区间范围为 [0.7862, 0.9545]，区间宽度较窄，表明零膨胀参数的估计具有较高的精度和可靠性。几何分布的三个均值参数的估计结果揭示了变量间的异质性。μ_1 的估计值为 0.1230，在三个变量中最小；μ_2 的估计值为 1.7219，计数水平最高；μ_3 的估计值为 1.3529，位于中间位置。同时 Bootstrap 方法提供了合理的区间估计。

采用 AIC 和 BIC 准则评估多元零膨胀几何(MZIG)分布的拟合优度，并与多元几何分布进行比较。参数估计采用 EM 算法，信息准则计算结果如表 3 所示，结果表明多元零膨胀几何分布在考虑参数复杂度

后显著优于传统几何分布。

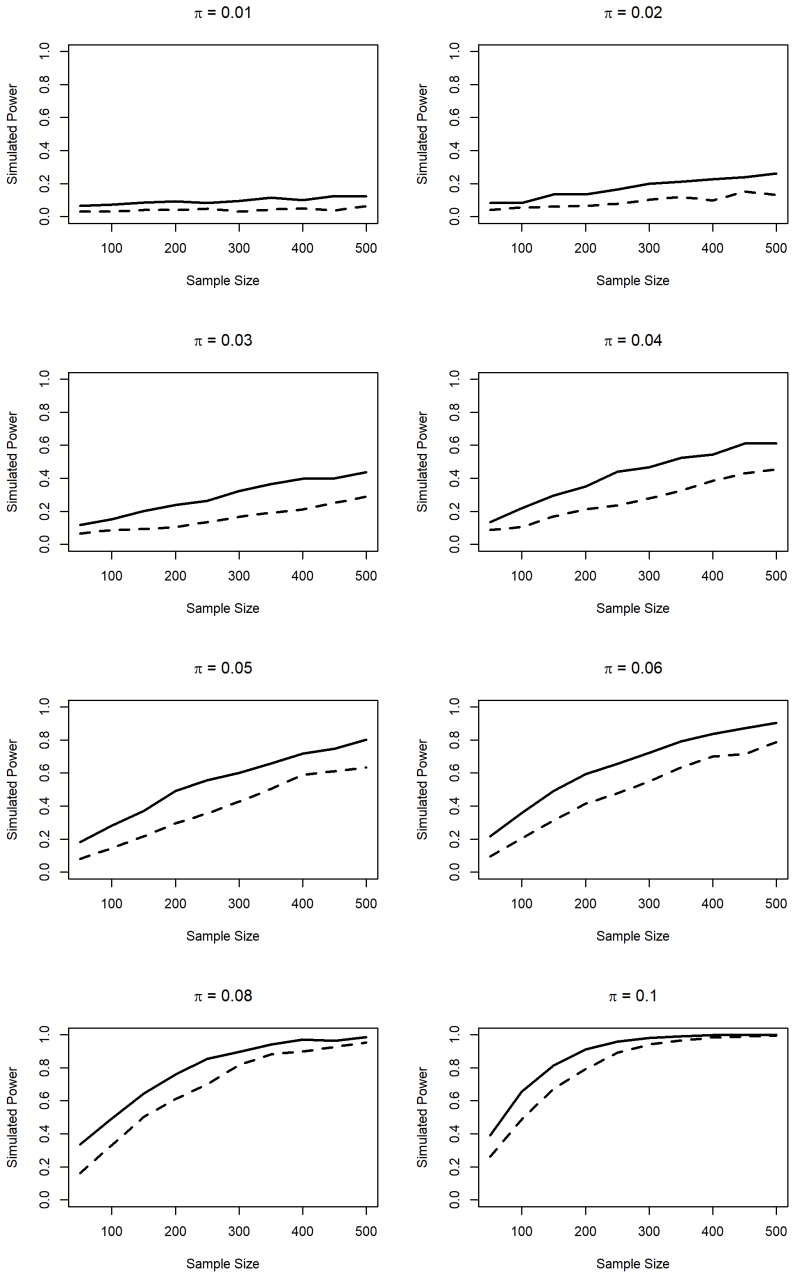

Figure 2. Comparison of powers between the likelihood ratio test (solid line) and the score test (dotted line) ($m = 2$))
图 2. 似然比检验(实线)与评分检验(虚线)的功效比较($m = 2$)

Table 3. Comparison of AIC and BIC between the two distributions
表 3. 两种分布的 AIC 与 BIC 比较

分布	AIC	BIC
多元几何分布	157.9153	164.7453
多元零膨胀几何分布	108.8718	117.9785

6. 结论与展望

6.1. 结论

本研究构建了一个针对多元零膨胀几何分布的综合性理论框架，并开发了相应的回归分析方法。所提出的模型成功解决了多元计数数据过多零值这一关键问题，从而为建模多个相关变量提供了可靠的分析工具。通过系统推导，本文阐明了该分布的基本性质，包括其联合概率质量函数、累积分布函数、矩结构以及条件分布。在参数估计方面，实现了期望最大化算法和 Fisher 评分算法，这两种算法在有限样本中均表现出良好的性能。模拟研究证实了估计量的一致性和似然比检验的优越性，似然比检验在保持名义误差率的同时具有较好的检验效能。

6.2. 展望

尽管本文提出的 MZIG 分布在理论和应用上展现出良好性能，但当前框架存在若干局限。其中最值得关注的是各分量在非零状态下条件独立的强假设，这一假设意味着在给定非零状态的条件下，各分量间的相关性完全由零膨胀机制所刻画，而排除了分量间其他形式的依赖关系。因此，当前模型在描述具有复杂依赖结构的多元计数数据时可能存在一定的拟合偏差。

针对上述局限，未来研究可从以下几个方向展开：首先，引入 Copula 函数构建基础向量的联合分布，通过连接函数可以将边缘几何分布与灵活的依赖结构相结合，从而允许分量间存在除零膨胀机制外的多种关联模式。其次，在多元回归框架内扩展协变量效应建模，特别是针对高维环境下可能需要的降维处理技术。第三，开发贝叶斯估计方法以充分利用先验信息，并处理小样本情况下的参数不确定性。最后，将时间依赖结构纳入模型，构建适用于零膨胀计数数据的动态模型。

基金项目

辽宁科技大学博士启动基金(6003000310)。

参考文献

[1] Cameron, A.C. and Trivedi, P.K. (2013) Regression Analysis of Count Data. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9781139013567

[2] Warton, D.I. (2005) Many Zeros Does Not Mean Zero Inflation: Comparing the Goodness-Of-Fit of Parametric Models to Multivariate Abundance Data. *Environmetrics*, **16**, 275-289. https://doi.org/10.1002/env.702

[3] Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L. and Kirchner, U. (1999) The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society Series A*: *Statistics in Society*, **162**, 195-209. https://doi.org/10.1111/1467-985x.00130

[4] Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer, D.B. (1996) Modelling the Abundance of Rare Species: Statistical Models for Counts with Extra Zeros. *Ecological Modelling*, **88**, 297-308. https://doi.org/10.1016/0304-3800(95)00113-1

[5] Crépon, B. and Duguet, E. (1997) Estimating the Innovation Function from Patent Numbers: GMM on Count Panel Data. *Journal of Applied Econometrics*, **12**, 243-263. https://doi.org/10.1002/(sici)1099-1255(199705)12:3<243::aid-jae444>3.0.co;2-4

[6] Hu, M., Pavlicova, M. and Nunes, E.V. (2011) Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *The American Journal of Drug and Alcohol Abuse*, **37**, 367-375. https://doi.org/10.3109/00952990.2011.597280

[7] Min, Y. and Agresti, A. (2005) Random Effect Models for Repeated Measures of Zero-Inflated Count Data. *Statistical Modelling*, **5**, 1-19. https://doi.org/10.1191/1471082x05st084oa

[8] Greene, W.H. (1994) Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. New York University.

[9] Lambert, D. (1992) Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*,

34, 1-14. https://doi.org/10.2307/1269547

[10] Ridout, M., Demétrio, C.G.B. and Hinde, J. (1998) Models for Count Data with Many Zeros. *Proceedings of the XIX International Biometric Conference*, Cape Town, 14-18 December 1998, 179-192.

[11] Bakouch, H.S. and Ristić, M.M. (2010) Zero Truncated Poisson Integer-Valued AR(1) Model. *Metrika*, **72**, 265-280. https://doi.org/10.1007/s00184-009-0252-5

[12] Jazi, M.A., Jones, G. and Lai, C. (2012) First-Order Integer Valued AR Processes with Zero Inflated Poisson Innovations. *Journal of Time Series Analysis*, **33**, 954-963. https://doi.org/10.1111/j.1467-9892.2012.00809.x

[13] Dietz, E. and Böhning, D. (2000) On Estimation of the Poisson Parameter in Zero-Modified Poisson Models. *Computational Statistics & Data Analysis*, **34**, 441-459. https://doi.org/10.1016/s0167-9473(99)00111-5

[14] Srisuradetchai, P. and Dangsupa, K. (2023) On Interval Estimation of the Geometric Parameter in a Zero-Inflated Geometric Distribution. *Thailand Statistician*, **21**, 93-109.

[15] Mallick, A. and Joshi, R. (2018) Parameter Estimation and Application of Generalized Inflated Geometric Distribution. *Journal of Statistical Theory and Applications*, **17**, 491-519. https://doi.org/10.2991/jsta.2018.17.3.7

[16] Pandya, M., Pandya, H. and Pandya, S. (2012) Bayesian Inference on Mixture of Geometric with Degenerate Distribution: Zero Inflated Geometric Distribution. *International Journal of Research and Reviews in Applied Sciences*, **13**, 53-66.

[17] Liu, Y. and Tian, G.L. (2015) Type I Multivariate Zero-Inflated Poisson Distribution with Applications. *Computational Statistics & Data Analysis*, **83**, 200-222. https://doi.org/10.1016/j.csda.2014.10.010

[18] Zhang, P., Chen, Z., Tzougas, G., *et al*. (2025) Multivariate Zero-Inflated INAR(1) Model with an Application in Automobile Insurance. *North American Actuarial Journal*, **29**, 310-328. https://doi.org/10.1080/10920277.2024.2381726

[19] Self, S.G. and Liang, K. (1987) Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association*, **82**, 605-610. https://doi.org/10.1080/01621459.1987.10478472

[20] Feng, Z. and McCulloch, C.E. (1992) Statistical Inference Using Maximum Likelihood Estimation and the Generalized Likelihood Ratio When the True Parameter Is on the Boundary of the Parameter Space. *Statistics & Probability Letters*, **13**, 325-332. https://doi.org/10.1016/0167-7152(92)90042-4

[21] Jansakul, N. and Hinde, J.P. (2002) Score Tests for Zero-Inflated Poisson Models. *Computational Statistics & Data Analysis*, **40**, 75-96. https://doi.org/10.1016/s0167-9473(01)00104-9

[22] Li, C., Lu, J., Park, J., Kim, K., Brinkley, P.A. and Peterson, J.P. (1999) Multivariate Zero-Inflated Poisson Models and Their Applications. *Technometrics*, **41**, 29-38. https://doi.org/10.1080/00401706.1999.10485593