Advances in Applied Mathematics 应用数学进展, 2023, 12(4), 1574-1580
Published Online April 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.124162

Hans汉斯

混合幂级数分布及其应用

赵雯雪，屈志扬，侯　文*

辽宁师范大学数学学院，辽宁　大连

收稿日期：2023年3月19日；录用日期：2023年4月15日；发布日期：2023年4月24日

摘　要

幂级数分布是一类广泛的概率分布，包括了常用的二项分布、Poisson分布等多种离散型概率分布。由若干个幂级数分布构成的混合幂级数分布，能够灵活地拟合如零膨胀、0~*k*膨胀以及其他多种膨胀类型的计数数据。本文基于EM算法给出了混合幂级数分布的参数估计，并对混合幂级数分布的特例0~1膨胀Poisson分布进行了说明，通过实例分析，说明混合幂级数分布及估计方法是适用的。

关键词

计数数据，混合幂级数分布，EM算法

Mixture Power Series Distribution and Its Application

Wenxue Zhao, Zhiyang Qu, Wen Hou*

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: Mar. 19th, 2023; accepted: Apr. 15th, 2023; published: Apr. 24th, 2023

Abstract

Power series distribution is a kind of extensive probability distribution, including the commonly used binomial distribution, Poisson distribution and other many discrete distributions. The mixed power series distribution, which is composed of multiple power series distributions, can flexibly fit the counting data of zero inflated, 0~*k* inflated and other many of inflated types. In this paper, the parameter estimation of mixed power series distribution is given based on EM algorithm, and the special case of mixed power series distribution 0~1 inflated Poisson distribution is explained. Through case analysis, the mixed power series distribution and estimation method are proved to be applicable.

*通讯作者。

文章引用: 赵雯雪, 屈志扬, 侯文. 混合幂级数分布及其应用[J]. 应用数学进展, 2023, 12(4): 1574-1580.
DOI: 10.12677/aam.2023.124162

Keywords

Count Data, Mixture Power Series Distribution, EM Algorithm

1. 引言

幂级数分布最初是由 Noak [1]在研究了一类具有共同特性的离散型随机变量分布的基础上，给出了幂级数分布的定义，同时也研究了关于幂级数分布矩的性质，并对幂级数分布中的一些重要特例，如二项分布、Poisson 分布、负二项分布等分别进行了说明。幂级数分布属于离散型概率分布族，目前已推广到了广义幂级数分布[2]，如广义 Poisson 分布和广义负二项分布等都属于这一类型的分布。幂级数分布适用于对计数数据的统计分析，而计数数据是一类重要的数据类型，广泛存在于医学、保险精算、人口学等多个研究领域中。

在实际对计数数据的统计过程中，经常会出现观测值在某些特定数值上频数过高的现象，例如在某段时间内汽车保险索赔次数、人工流产的次数等等，就是数据中 0 出现的频数明显过高，高于预期分布中 0 的期望值，即所谓的零膨胀数据。对于该类型的数据统计方法的研究，解锋昌、韦博成和林金官[3]有比较系统的论述。

除了零膨胀现象外，还会遇到某个特定的正整数值 k 大量出现的情况，如某个地区居民一年内看牙医的次数，观测数据中大量出现了 0 和 1，这可能是因为许多人没有养成定期看牙医的良好习惯，或是牙齿有问题去看牙医之后，短期之内就不再去看牙医。为了更好地拟合此类数据，Lin 和 Tsai [4]提出了 0~k 膨胀的模型。

在目前已有的研究中，对零膨胀数据或 0~k 膨胀数据进行分析时，模型大多选用的是 Poisson 分布或负二项分布，是否存在其他分布更适合这种数据，是一个值得探讨的问题。

2. 混合幂级数分布

2.1. 幂级数分布

由幂级数展开式 $g(\theta)=\sum_{l=0}^{\infty}a(l)\theta^{l}$，其中，$a(l)$ 为常数序列，从而有

$$\sum_{l=0}^{\infty}\frac{a(l)\theta^{l}}{g(\theta)}=1$$

故定义离散型随机变量 X 的概率分布律为

$$P(X=l)=\frac{a(l)\theta^{l}}{g(\theta)},\quad l=0,1,2,\cdots \text{且}\ \theta>0,\quad a_{l}>0,$$

称 X 为幂级数分布。

如在幂级数分布定义中，$a_{l}=\frac{1}{l!}$，$g(\theta)=\mathrm{e}^{\theta}$，显然 X 为 Possion 分布，记为 $X\sim \text{Possion}(\theta)$。又如

取 $a_l = C_n^l$， $g(\theta) = (1+\theta)^n = \sum_{l=0}^{\infty} C_n^l \theta^l$ 时，X 的概率分布律为

$$P(X=l) = C_n^l \frac{\theta^l}{(1+\theta)^n} = C_n^l \left(\frac{\theta}{1+\theta}\right)^l \left(\frac{1}{1+\theta}\right)^{n-l}, \quad l = 0,1,2,\cdots,n,$$

是二项分布，记为 $X \sim \mathrm{b}\left(n, \frac{\theta}{1+\theta}\right)$。

2.2. 混合幂级数分布

设离散型随机变量 Y，概率分布律为

$$f(y|\theta) = \sum_{k=1}^{m} \alpha_k f_k(y|\theta_k) = \alpha_1 f_1(y|\theta_1) + \alpha_2 f_2(y|\theta_2) + \cdots + \alpha_m f_m(y|\theta_m) \tag{1}$$

其中，$f_k(y|\theta_k) = \frac{a_k(y)\theta_k^y}{g_k(\theta_k)}$，$\theta = (\theta_1,\cdots,\theta_m)$，$y = 0,1,\cdots$，$\alpha_k$ 为系数且 $\alpha_k > 0$，$\sum_{k=1}^{m} \alpha_k = 1$，称为混合幂级数分布。

混合幂级数分布的概率母函数为

$$G_Y(s) = E(s^Y) = \sum_{k=1}^{m} \alpha_k \sum_{y=0}^{\infty} f_k(y|\theta_k) s^y = \sum_{k=1}^{m} \alpha_k \sum_{l=0}^{\infty} \frac{a_k(l)(\theta_k s)^l}{g_k(\theta_k)}$$

通过概率母函数可以求得混合幂级数分布的各阶矩。

3. 混合幂级数分布的参数估计

若构成混合分布的幂级数分布较多，则需要估计的参数也多，用矩估计法是比较困难的。因此，对混合幂级数分布参数估计主要讨论 EM 算法[5]。

设观测 $Y_1, Y_2, \cdots, Y_n$ 是来自总体分布如式(1)的混合幂级数分布的样本，引入隐变量 Z_{ki}，当 Y_i 来自分布 $f_k(y|\theta_k)$ 时，有 $f(Z_{ki}=1) = \alpha_k$，$k = 1,2,\cdots,m$，记

$$y = (y_1,\cdots,y_n), \quad Z_k = (Z_{k1},\cdots,Z_{kn}), \quad \vartheta = (\theta_1,\cdots,\theta_m,\alpha_1,\cdots,\alpha_m),$$

完全数据的似然函数为

$$L(\vartheta|y,z_k) = \prod_{k=1}^{m}\prod_{i=1}^{n} \left[\alpha_k f_k(y_i|\theta_k)\right]^{z_{ki}},$$

对数似然函数为

$$l(\vartheta|y,z_k) = \log L(\vartheta|y,z_k) = \sum_{k=1}^{m}\sum_{i=1}^{n} z_{ki}\left[\log \alpha_k f_k(\theta_k|y_i)\right]$$

E 步：

$$\hat{z}_{ki} = E(z_{ki}|y_i,\theta_k) = f_k(z_{ki}|y_i,\theta_k) = \frac{f_k(z_{ki},y_i|\theta_k)}{\sum_{k=1}^{m} f_k(z_{ki},y_i|\theta_k)} = \frac{\alpha_k f_k(y_i|\theta_k)}{\sum_{k=1}^{m} \alpha_k f_k(y_i|\theta_k)} = \frac{\alpha_k \frac{a_k(y_i)\theta_k^{y_i}}{g_k(\theta_k)}}{\sum_{k=1}^{m} \alpha_k \frac{a_k(y_i)\theta_k^{y_i}}{g_k(\theta_k)}}$$

令

$$Q\left(\vartheta,\vartheta^{(s)}\right)=E\left\{\sum_{k=1}^{m}\sum_{i=1}^{n}z_{ki}\left[\log\alpha_k f_k\left(\theta_k\mid y_i\right)\right]\right\}$$

其中，$\vartheta^{(s)}$ 表示第 s 次迭代的各参数，则

$$\begin{aligned}Q\left(\vartheta,\vartheta^{(s)}\right)&=\sum_{k=1}^{m}\sum_{i=1}^{n}\hat{z}_{ki}\left[\log\alpha_k+\log f_k\left(y_i\mid\theta_k\right)\right]\\&=\sum_{k=1}^{m}\sum_{i=1}^{n}\hat{z}_{ki}\left[\log\alpha_k+\log a_k\left(y_i\right)+y_i\log\theta_k-\log g_k\left(\theta_k\right)\right]\end{aligned}$$

M 步：求函数 $Q\left(\vartheta,\vartheta^{(s)}\right)$ 的极大值点，计算的迭代公式为

$$\begin{cases}\hat{\alpha}_k^{(s+1)}=\dfrac{\sum\limits_{j=1}^{n}\hat{z}_{jk}^{(s)}}{\sum\limits_{k=1}^{m}\sum\limits_{j=1}^{n}\hat{z}_{jk}^{(s)}}\\[2ex]\hat{\theta}_k^{(s+1)}=\dfrac{\sum\limits_{j=1}^{n}\hat{z}_{jk}^{(s)}y_j f_k\left(\theta_k^{(s)}\right)}{\sum\limits_{j=1}^{n}\hat{z}_{jk}^{(s)}f_k\left(\theta_k^{(s)}\right)}\end{cases}\tag{2}$$

重复上述步骤，直到两次迭代值之差小于给定值为止。通过幂级数分布的一些特例，如二项分布、Poisson 分布、几何分布、负二项分布等可以构造出各种形式灵活的混合幂级数分布。

4. 混合幂级数分布的特例及应用

本节主要通过 0~1 分布和 Poisson 分布构造出混合幂级数分布的特例 0~1 膨胀 Poisson 分布。

4.1. 0~1 膨胀 Poisson 分布

设随机变量 $X_1\sim \mathrm{b}\left(1,\dfrac{\theta_1}{1+\theta_1}\right)$，$X_2\sim \mathrm{Possion}\left(\theta_2\right)$，$Z=\begin{cases}1 & \text{当}Y=X_1\text{时}\\0 & \text{当}Y=X_2\text{时}\end{cases}$，

且 $f\left(Z=1\right)=\alpha$，$f\left(Z=0\right)=1-\alpha$，即 $Y=ZX_1+\left(1-Z\right)X_2$，$Y$ 的概率分布律为

$$P\left(Y=k\right)=\begin{cases}\alpha\dfrac{1}{1+\theta_1}+\left(1-\alpha\right)\mathrm{e}^{-\theta_2} & k=0\\[2ex]\alpha\dfrac{\theta_1}{1+\theta_1}+\left(1-\alpha\right)\theta_2\mathrm{e}^{-\theta_2} & k=1\\[2ex]\left(1-\alpha\right)\dfrac{\theta_2^k}{k!}\mathrm{e}^{-\theta_2} & k=2,3,\cdots\end{cases}\tag{3}$$

称 Y 服从 0~1 膨胀 Poisson 分布。

4.2. 0~1 膨胀 Poisson 分布的参数估计

4.2.1. 矩估计

0~1 膨胀 Poisson 分布的概率母函数为

$$G_Y\left(t\right)=E\left(t^Y\right)=\sum_{k=0}^{\infty}t^k p\left(Y=k\right)=\alpha\left(\frac{1}{1+\theta_1}+\frac{t\theta_1}{1+\theta_1}\right)+\left(1-\alpha\right)\mathrm{e}^{\theta_2\left(t-1\right)},$$

求它的 1 阶、2 阶、3 阶导数分别为

$$G_Y'(1)=\alpha\frac{\theta_1}{1+\theta_1}+(1-\alpha)\theta_2,\quad G_Y''(1)=(1-\alpha)\theta_2^2,\quad G_Y'''(1)=(1-\alpha)\theta_2^3,$$

因此

$$E(Y)=G'(1)=\alpha\frac{\theta_1}{1+\theta_1}+(1-\alpha)\theta_2,$$

$$E(Y^2)=G''(1)+G'(1)=(1-\alpha)\theta_2^2+(1-\alpha)\theta_2+\alpha\frac{\theta_1}{1+\theta_1},$$

$$E(Y^3)=G'''(1)+3G''(1)+G'(1)=(1-\alpha)\theta_2^3+3(1-\alpha)\theta_2^2+(1-\alpha)\theta_2+\alpha\frac{\theta_1}{1+\theta_1}。$$

设 $Y_1,Y_2,\cdots,Y_n$ 是从式(3)的总体 Y 中抽取的样本，样本的 1 阶、2 阶和 3 阶原点矩分别记为

$$m_1=\frac{1}{n}\sum_{i=1}^{n}Y_i,\quad m_2=\frac{1}{n}\sum_{i=1}^{n}Y_i^2,\quad m_3=\frac{1}{n}\sum_{i=1}^{n}Y_i^3,$$

由矩估计的原理，可得

$$\hat{\theta}_2=\frac{m_3-3m_2+2m_1}{m_2-m_1},\quad \hat{\alpha}=1-\frac{(m_2-m_1)^3}{(m_3-3m_2+2m_1)^2} \tag{4}$$

进而得

$$\hat{\theta}_1=\frac{m_1-(1-\hat{\alpha})\hat{\theta}_2}{\hat{\alpha}-m_1+(1-\hat{\alpha})\hat{\theta}_2} \tag{5}$$

4.2.2. EM 估计

总体 Y 为 0~1 膨胀 Poisson 分布，在混合幂级数分布式(1)的定义，取 $m=2$，只需引入隐变量为 Z_i，当 Y_i 来自分布 $Y_i\sim \mathrm{b}\left(1,\frac{\theta_1}{1+\theta_1}\right)$ 时，有 $f(Z_i=1)=\alpha$，当 Y_i 来自分布 $\mathrm{Possion}(\theta_2)$ 时，有 $f(Z_i=0)=1-\alpha$，$i=1,2,\cdots,n$。

记 $z=(z_1,\cdots,z_n)$，$y=(y_1,\cdots,y_n)$，完全数据的似然函数为

$$L(\theta\mid y,z)=\prod_{i=1}^{n}\left[\alpha\frac{\theta_1^{y_i}}{1+\theta_1}\right]^{z_i}\left[(1-\alpha)\frac{\theta_2^{y_i}}{x_i!}\mathrm{e}^{-\theta_2}\right]^{1-z_i},$$

对数似然函数为

$$\ell(\theta\mid y,z)=\sum_{i=1}^{n}\left\{z_i\left[\ln\alpha+y_i\ln\theta_1-\ln(1+\theta_1)\right]+(1-z_i)\left[\ln(1-\alpha)+y_i\ln\theta_2-\theta_2-\ln y_i!\right]\right\},$$

然后根据本文第 2 部分 EM 算法中的迭代公式(2)可以得到参数估计结果。

事实上，对于 0~1 膨胀 Poisson 分布，在 E 步中得到的 $\hat{z}_i$ 更简单，只有三种情形。具体地

当 $y_i=0$ 时，$\hat{z}_i=\dfrac{\frac{\alpha}{1+\theta_1}}{\frac{\alpha}{1+\theta_1}+(1-\alpha)\mathrm{e}^{-\theta_2}}=\dfrac{1}{1+(1+\theta_1)(1/\alpha-1)\mathrm{e}^{-\theta_2}}$，记为 $\tilde{p}_0$；

当 $y_i=1$ 时，$\hat{z}_i=\dfrac{\frac{\alpha\theta_1}{1+\theta_1}}{\frac{\alpha\theta_1}{1+\theta_1}+(1-\alpha)\theta_2\mathrm{e}^{-\theta_2}}=\dfrac{1}{1+(1+1/\theta_1)(1/\alpha-1)\theta_2\mathrm{e}^{-\theta_2}}$，记为 $\tilde{p}_1$；

当 $y_i=2,3$ 时， $\hat{z}_i=0$ 。

记 n_0 为样本中 $y_i=0$ 的个数， n_1 为样本中 $y_i=1$ 的个数， $\theta^{(0)}$ 为初始参数值，故在 M 步中，

$$\begin{aligned}Q\left(\theta\mid y,\theta^{(0)}\right)=&\left\{\tilde{p}_0\left[\ln\alpha-\ln\left(1+\theta_1\right)\right]+\left(1-\tilde{p}_0\right)\left[\ln\left(1-\alpha\right)-\theta_2\right]\right\}n_0\\&+\left\{\tilde{p}_1\left(\ln\alpha+\ln\frac{\theta_1}{1+\theta_1}\right)+\left(1-\tilde{p}_1\right)\left[\ln\left(1-\alpha\right)+\ln\theta_2-\theta_2\right]\right\}n_1\\&+\sum_{y_i>2}\left[\ln\left(1-\alpha\right)+y_i\ln\theta_2-\ln y_i-\theta_2\right]\end{aligned}$$

从而，得

$$\begin{cases}\hat{\alpha}^{(s+1)}=\dfrac{n_0\tilde{p}_0^{(s)}+n_1\tilde{p}_1^{(s)}}{n}\\\hat{\theta}_1^{(s+1)}=\dfrac{n_1\tilde{p}_1^{(s)}}{n_0\tilde{p}_0^{(s)}}\\\hat{\theta}_2^{(s+1)}=\dfrac{\sum y_i-n_1\tilde{p}_1^{(s)}}{n-n_0\tilde{p}_0^{(s)}-n_1\tilde{p}_1^{(s)}}\end{cases}\tag{6}$$

其中， $\tilde{p}_0^{(s)}$ 、 $\tilde{p}_1^{(s)}$ 为第 s 步迭代值， $\hat{\alpha}^{(s+1)}$ 、 $\hat{\theta}_1^{(s+1)}$ 、 $\hat{\theta}_2^{(s+1)}$ 为第 $s+1$ 步的迭代值。通过上述迭代，得到参数的最大似然估计值。

4.3. 实例分析

本小节分析的数据集是新西兰白兔出产死亡的数据，最初是由 Morgan 等[6]研究得分检验时使用过。表 1 的第 1 列和第 2 列列出了新西兰白兔每窝死产数量和观测频数。数据分别用零膨胀 Poisson 分布(ZIP)和 0~1 膨胀 Poisson (ZIOP)分布进行拟合的，参数估计的方法采用矩估计(ME)和基于 EM 算法的最大似然估计(MLE)。

Table 1. Fitting results of stillbirth numbers in New Zealand white rabbits
表 1. 新西兰白兔死产数量数据的拟合结果

死产数量	观测频数	期望频数			
		ME		MLE	
		ZIP	ZOIP	ZIP	ZOIP
0	314	328	314	314	316
1	48	22	52	36	48
2	20	23	11	28	16
3	7	16	10	15	11
4	5	8	7	6	6
5	2	3	4	2	2
6^+	6	2	4	1	3
合计	402	402	402	402	402

Continued

参数	α	0.790	0.882	0.724	0.851
	θ_1	-	0.142	-	0.106
	θ_2	2.069	2.768	1.578	2.110
	χ^2	63.90	11.14	35.72	5.63
	自由度	4	3	4	3
	P 值	0.0001	0.015	0.0001	0.076

ZOIP 模型的参数的矩估计通过式(4)和式(5)，计算结果分别列于表 1 的第 4 列中的第 12 行至第 14 行，最大似然估计通过式(6)得到，计算结果分别列于表 1 的第 6 列中的第 12 行至第 14 行。作为对照，将 ZIP 模型的参数的矩估计和极大似然估计结果分别列于表 1 的第 3 列和第 5 列中的第 12 行和第 14 行。

根据模型参数估计值拟合的新西兰白兔死产的频数列于表 1 上半部分，ZIP 模型和 ZOIP 模型的拟合优度检验统计量列于表 1 的下半部分。由于 ZIP 模型包括 2 个参数，ZOIP 模型 3 个参数，拟合优度检验中 ZIP 模型自由度为 $7-2-1=4$，ZOIP 模型自由度为 $7-3-1=3$。拟合优度检验结果表明，只有基于 EM 算法得到的 ZOIP 模型的拟合优度检验 P 值大于 0.05，故可以认为基于 EM 算法的 ZOIP 模型拟合该数据集是适合的。

5. 讨论

事实上，为了适应更多类型的数据，需要尝试选择多种分布来拟合数据。因此，对幂级数分布及相关内容的研究具有重要意义。如在幂级数分布基础上可以选择加入一个比例参数，构成零膨胀幂级数分布，也可以选择若干个幂级数分布的构成混合幂级数分布。混合幂级数分布可以构造出很多灵活的模型，包括零膨胀或 0~k 膨胀以及多种膨胀类型的计数数据，拓宽了幂级数分布的应用范围，也为多点膨胀计数模型提供了一个新思路。

基金项目

本研究由 2022 年度辽宁省教育厅高校基本科研项目(LJKMZ20221412)和 2022 年度辽宁省研究生教育教学改革研究项目(2022-180-39510165)资助。

参考文献

[1] Noak, A. (1950) A Class of Random Variable with Discrete Distribution. *Annals of the Institute of Statistical Mathematics*, **21**, 127-132. https://doi.org/10.1214/aoms/1177729894

[2] Patil, G.P. (1962) On Certain Properties of the Generalized Power Series Distribution. *Annals of the Institute of Statistical Mathematics*, **14**, 179-182. https://doi.org/10.1007/BF02868639

[3] 解锋昌, 韦博成, 林金官. 零过多数据的统计分析及其应用[M]. 北京: 科学出版社, 2013.

[4] Lin, T.H. and Tsai, M.H. (2013) Modeling Health Survey Data with Excessive Zero and K Responses. *Statistics in Medicine*, **32**, 1572-1583. https://doi.org/10.1002/sim.5650

[5] Dempster, A., Laird, N. and Rubin, D. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[6] Morgan, B.J.T., Palmer, K.J. and Ridout, M.S. (2007) Negative Score Test Statistics. *The American Statistician*, **61**, 285-288. https://doi.org/10.1198/000313007X242972