一类泊松-多项模型参数估计的 EM 算法

李　苗，魏立力
（宁夏大学数学统计学院，银川　750021）

摘要：研究一类泊松-多项模型参数估计的期望最大化（expectation maximization，EM）算法。首先，针对完全数据情形推导出了极大似然估计（maximum likelihood estimate，MLE）的解析表达式；其次，设计适用于缺失数据情形参数估计的 EM 算法，给出了算法迭代序列的表达式；最后，通过一个算例说明了所提方法的可行性。结果表明，EM 算法解决了极大似然估计法无法得到缺失数据对应参数估计值的问题，适用于不完全数据情形的泊松-多项模型参数估计，体现了一定的优越性。

关键词：数理统计学；EM 算法；缺失数据；泊松-多项分布
中图分类号：O212　　**文献标识码：**A　　**文章编号：**1674-2850(2018)19-1943-05

EM algorithm for parameter estimation of a class of Poisson-multinomial models

LI Miao, WEI Lili
(*School of Mathematics and Statistics*, *Ningxia University*, *Yinchuan*　750021, *China*)

Abstract: An expectation maximization (EM) algorithm for parameter estimation of a class of Poisson-multinomial models is studied in this paper. Firstly, the analytical expression of maximum likelihood estimate (MLE) is derived for the complete data case. Then the EM algorithm for parameter estimation is designed for the missing data case, and the expression of iterative sequence of the algorithm is given. Finally, an example is given to illustrate the feasibility of the proposed method. The results show that the EM algorithm not only solves the problem that the maximum likelihood estimation method cannot obtain the estimated value of corresponding parameters in the missing data, but also applies to the parameter estimation of Poisson-multinomial models for the incomplete data case, so that the superiority of the EM algorithm is reflected.

Key words: mathematical statistics; EM algorithm; missing data; Poisson-multinomial distribution

0　引言

疾病分布模型的研究可为探索病因提供线索，发病频数是卫生统计中的一个重要指标，通常表示为一个离散型随机变量[1]。有很多因素都有可能影响疾病的发生，例如不同时期、不同地方等都需要不同的统计模型描述相应的发病频数。一个合适的模型可以更好地描述疾病，有利于疾病的预防和控制。实践证明，单一参数分布模型（如 Poisson 分布）往往不能很好地描述疾病分布[2]，因而寻找适宜的多分布混合模型是一个重要的研究方向。泊松-多项分布能够刻画不同子区域人口的不同发病因子，因而是一个良好的发病频数统计分布。

在某区域i观测到数据(Y_i, X_i)，$i=1,\cdots,n$，它们相互独立，其中，Y_i为观测到的该区域发病人数，X_i为

基金项目：国家自然科学基金（11261044）
作者简介：李苗（1994—），女，硕士研究生，主要研究方向：应用统计与数据分析
通信联系人：魏立力，教授，主要研究方向：应用统计与数据分析. E-mail: weill866@163.com

对应的人口数。假定对人口计数使用多项分配模型，则有如下泊松-多项模型[3]：

$$Y_i \sim \text{Poisson}(m\beta\tau_i),$$

$$(X_1,\cdots,X_n) \sim \text{multinomial}(m;\boldsymbol{\tau}),$$

其中，β 为一个总效应；τ_i 为区域 i 的特有因子，表示该影响发病率的区域性测度，$\boldsymbol{\tau}=(\tau_1,\tau_2,\cdots,\tau_n)$，且 $\sum_{i=1}^{n}\tau_i=1$；$m=\sum_{i=1}^{n}x_i$ 视为已知。τ_i 无法被直接观测到，但可通过 X_i 来获取其信息。该模型可用于刻画某种疾病的发病率。

本文研究泊松-多项模型的参数估计问题，推导出完全数据情形的 MLE 的解析表达式，设计适用于缺失数据情形参数估计的 EM 算法，并通过算例验证本文方法的可行性。

1　极大似然估计

在泊松-多项模型的参数估计问题中，似然函数为

$$L(\beta,\tau_1,\cdots,\tau_n\,|\,(x_1,y_1),\cdots,(x_n,y_n))=\prod_{i=1}^{n}\frac{\mathrm{e}^{-m\beta\tau_i}(m\beta\tau_i)^{y_i}}{y_i!}m!\frac{\tau_i^{x_i}}{x_i!},\tag{1}$$

其中，$(x_1,y_1),\cdots,(x_n,y_n)$ 为已观测数据，其对数似然函数为

$$\log L=\sum_{i=1}^{n}[y_i\log(m\beta\tau_i)+x_i\log\tau_i-m\beta\tau_i+\log m!-\log(y_i!x_i!)].\tag{2}$$

下面求式（2）的驻点，求导并令其为零，得到方程组：

$$\begin{cases}\dfrac{\partial}{\partial\beta}\log L=\sum\limits_{i=1}^{n}\left(\dfrac{y_i}{\beta}-m\tau_i\right)=0,\\[2ex]\dfrac{\partial}{\partial\tau_i}\log L=\dfrac{x_i+y_i}{\tau_i}-m\beta=0,\quad i=1,2,\cdots,n.\end{cases}$$

解方程组得其唯一解 $\beta=\dfrac{\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i}$，$\tau_i=\dfrac{x_i+y_i}{\sum_{i=1}^{n}(x_i+y_i)}$，$i=1,2,\cdots,n.$

可以进一步证明，上述解是最大值点[4]，故得 MLE 为

$$\hat{\beta}=\frac{\sum_{i=1}^{n}Y_i}{\sum_{i=1}^{n}X_i},\quad \hat{\tau}_i=\frac{X_i+Y_i}{\sum_{i=1}^{n}(X_i+Y_i)},\quad i=1,2,\cdots,n.\tag{3}$$

可见，当数据完全可观测时，MLE 的计算并不困难。然而，当数据缺失（假设数据 x_1 缺失）时，似然函数为

$$L(\beta,\tau_1,\tau_2,\cdots,\tau_n\,|\,y_1,(x_2,y_2),\cdots,(x_n,y_n))=\prod_{i=1}^{n}\frac{\mathrm{e}^{-m\beta\tau_i}(m\beta\tau_i)^{y_i}}{y_i!}\prod_{i=2}^{n}m!\frac{\tau_i^{x_i}}{x_i!}.\tag{4}$$

要对上述函数求最大值是很困难的，一种不得已的办法是弃掉 y_1 而仅用 $n-1$ 容量的样本，β 和 τ_i 的 MLE 为

$$\hat{\beta}=\frac{\sum_{i=2}^{n}Y_i}{\sum_{i=2}^{n}X_i},\quad \hat{\tau}_i=\frac{X_i+Y_i}{\sum_{i=2}^{n}(X_i+Y_i)},\quad i=2,3,\cdots,n.\tag{5}$$

结合式（3）和式（5）可知，当舍弃 y_1 时，关于区域 1 的信息将被完全抹去，而参数 β 和 τ_i 的估计

均依赖于所得的观测数据，这会导致无法通过极大似然估计法得到参数 τ_1 的估计值 $\hat{\tau}_1$，同时剩余参数 $(\beta,\tau_2,\cdots,\tau_n)$ 的估计也将受到影响。

2 EM 算法

出于不完全数据参数估计的需要，DEMPSTER 等首次提出了 EM 算法[5]。该算法是一种局部最优算法[6]，其迭代过程分为两步，即 E 步（期望步）和 M 步（最大化步）。

令 $\boldsymbol{y}=(y_1,\cdots,y_n)$ 表示来自于分布 $f(\boldsymbol{y};\boldsymbol{\theta})$ 的不完全已观测数据，其中 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_p)$ 为参数。当一部分数据 $\boldsymbol{z}$ 缺失时，有必要“填充”缺失数据 $\boldsymbol{z}$，以得到完全数据 $\boldsymbol{x}=(\boldsymbol{y},\boldsymbol{z})$，这时在完全数据的对数似然函数 $\log L(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ 下通过 EM 算法得到的 MLE 就可以被确定。

E 步：将不完全数据 $\boldsymbol{y}$ 和第 $k-1$ 次迭代中的参数估计值 $\boldsymbol{\theta}^{(k-1)}$ 视为常数，对 $\boldsymbol{z}$ 求 $\log L(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ 的条件期望，$E[\log L(\boldsymbol{\theta}\,|\,\boldsymbol{x})\,|\,\boldsymbol{y},\boldsymbol{\theta}^{(k-1)}]\equiv Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k-1)})$.

M 步：将 E 步得到的期望函数 $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k-1)})$ 最大化，即求 $\boldsymbol{\theta}^{(k)}$，使得

$$\boldsymbol{\theta}^{(k)}=\arg\max_{\Theta} Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k-1)}),$$

其中，Θ 为参数空间。

记泊松−多项模型的参数 $\boldsymbol{\theta}=(\beta,\tau_1,\tau_2,\cdots,\tau_n)$，$(\boldsymbol{x},\boldsymbol{y})=((x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n))$ 表示完全数据，若数据 x_1 缺失，记数据为 $(\boldsymbol{x}_{(-1)},\boldsymbol{y})=(y_1,(x_2,y_2),\cdots,(x_n,y_n))$，由式（3）可得泊松−多项模型中的完全对数似然期望为

$$E\left[\log L(\boldsymbol{\theta}\,|\,(\boldsymbol{x},\boldsymbol{y}))\middle|\boldsymbol{\tau}^{(r)},(\boldsymbol{x}_{(-1)},\boldsymbol{y})\right]=\sum_{x_1=0}^{\infty}\left[\log\left(\prod_{i=1}^{n}\frac{\mathrm{e}^{-m\beta\tau_i}(m\beta\tau_i)^{y_i}}{y_i!}\cdot\frac{m!(\tau_i)^{x_i}}{x_i!}\right)\cdot\frac{m!(\tau_1^{(r)})^{x_1}}{x_1!}\right],\tag{6}$$

化简式（6），将其分组为含有参数 β 和 τ_i 的项与不含这些参数的项，得

$$\begin{aligned}
&E\left[\log L(\boldsymbol{\theta}\,|\,(\boldsymbol{x},\boldsymbol{y}))|\boldsymbol{\tau}^{(r)},(\boldsymbol{x}_{(-1)},\boldsymbol{y})\right]\\
&=\sum_{i=1}^{n}[-m\beta\tau_i+y_i(\log\beta+\log\tau_i+\log m)-\log(y_i!)]+\sum_{i=2}^{n}[\log(m!)+x_i\log\tau_i-\log(x_i!)]+\\
&\sum_{x_1=0}^{\infty}\left\{[x_1\log\tau_1-\log(x_1!)]\frac{m!(\tau_1^{(r)})^{x_1}}{x_1!}\right\}\\
&=\sum_{i=1}^{n}[-m\beta\tau_i+y_i(\log\beta+\log\tau_i)]+\sum_{i=2}^{n}\left(x_i\log\tau_i\right)+\sum_{x_1=0}^{\infty}\left(x_1\log\tau_1\right)\frac{m!(\tau_1^{(r)})^{x_1}}{x_1!}+\\
&\left[\sum_{i=1}^{n}y_i\log m-\sum_{i=1}^{n}\log(y_i!)-\sum_{i=2}^{n}\log(x_i!)+\log(m!)-\sum_{x_1=0}^{\infty}\log(x_1!)\frac{m!(\tau_1^{(r)})^{x_1}}{x_1!}\right].
\end{aligned}\tag{7}$$

忽略最后一对中括号中的项（因其不含参数 β 和 τ_i），利用 $X_1\sim\text{binomial}(m;\tau_1)$ 这一事实，重写第三个和式中的项，得

$$\log\tau_1\sum_{x_1=0}^{\infty}x_1\frac{m!(\tau_1^{(r)})^{x_1}}{x_1!}=m\tau_1^{(r)}\log\tau_1,$$

结合式（7）可知，当把 $m\tau_1^{(r)}$ 换成 x_1 时，式（6）就是完全数据似然。第 r 步迭代中的 MLE 与式（3）相

比仅有很小的差别，式（8）给出了 EM 算法迭代序列：

$$\hat{\beta}^{(r+1)}=\frac{\sum_{i=1}^{n}y_i}{m\hat{\tau}_1^{(r)}+\sum_{i=2}^{n}x_i},\quad \hat{\tau}_i^{(r+1)}=\frac{x_i+y_i}{m\hat{\tau}_1^{(r)}+\sum_{i=1}^{n}y_i+\sum_{i=2}^{n}x_i},\quad i=1,2,\cdots,n. \tag{8}$$

3　实例分析

表 1 中的数据来源于文献[7]，它记录了美国纽约州部分区域的白血病患者数和相应地区的居民人口数。

表 1　白血病患者计数

Tab. 1　Population count of leukemia patients

区域	1	2	3	4	5	6	7	8	9
人口数/人	*	3 560	3 739	2 784	2 571	2 729	3 952	993	1 908
患者数/人	3	4	1	1	3	1	2	0	2
区域	10	11	12	13	14	15	16	17	18
人口数/人	948	1 172	1 047	3 138	5 485	5 554	2 943	4 969	4 828
患者数/人	0	1	3	5	4	6	2	5	4

注：*表示该处数据缺失，下同

显然，表 1 提供的白血病患者计数数据是不完全数据，因为区域 1 对应的人口数 x_1 缺失。分别通过极大似然估计法与 EM 算法对参数 $(\beta,\tau_1,\tau_2,\cdots,\tau_n)$ 进行估计，结果如表 2 所示。

表 2　参数估计结果对比

Tab. 2　Comparison of parameter estimation results

参数	MLE	EM	参数	MLE	EM
$\hat{\beta}$	0.000 840 978 6	0.000 841 296 2	$\hat{\tau}_{10}$	0.018 104 04	0.016 954 86
$\hat{\tau}_1$	*	0.063 476 52	$\hat{\tau}_{11}$	0.022 400 89	0.020 978 96
$\hat{\tau}_2$	0.068 062 03	0.063 741 69	$\hat{\tau}_{12}$	0.020 051 94	0.018 779 12
$\hat{\tau}_3$	0.071 423 12	0.066 889 42	$\hat{\tau}_{13}$	0.060 022 15	0.056 212 16
$\hat{\tau}_4$	0.053 185 39	0.049 809 37	$\hat{\tau}_{14}$	0.104 823 92	0.098 170 07
$\hat{\tau}_5$	0.049 155 91	0.046 035 66	$\hat{\tau}_{15}$	0.106 179 82	0.099 439 89
$\hat{\tau}_6$	0.052 135 05	0.048 825 70	$\hat{\tau}_{16}$	0.056 240 93	0.052 670 95
$\hat{\tau}_7$	0.075 509 89	0.070 716 79	$\hat{\tau}_{17}$	0.094 988 92	0.088 959 36
$\hat{\tau}_8$	0.018 963 41	0.017 759 68	$\hat{\tau}_{18}$	0.092 277 14	0.086 419 71
$\hat{\tau}_9$	0.036 475 44	0.034 160 11			

注：此处的 MLE 是指舍弃 y_1，将剩余观测值视为样本总体的 MLE

结果显示，由于舍弃了数据 y_1，极大似然估计法无法得到参数 τ_1 的估计值 $\hat{\tau}_1$，其他参数的估计相对于 EM 算法的估计结果而言，呈现了不同程度的差异。EM 算法解决了极大似然估计法无法得到缺失数据对应的参数估计值的问题，体现了 EM 算法的优越性，表明 EM 算法适用于不完全数据的泊松-多项模型的参数估计。

4　结论

本文介绍了泊松-多项模型，利用 EM 算法研究了泊松-多项模型参数的极大似然估计问题，给出了

参数估计的迭代方程。针对缺失数据情形，对比了极大似然估计法和 EM 算法的参数估计结果，结果表明，EM 算法表现较好，适用于不完全数据的泊松-多项模型的参数估计。

[参考文献] (References)

[1] 陈锋，杨树勤. 混合 Poisson 分布及其应用——疾病的统计分布（五）[J]. 中国卫生统计，1997，14（2）：9-12.
CHEN F, YANG S Q. The mixed Poisson distribution and its application: statistical distribution of diseases (V)[J]. Chinese Journal of Health Statistics, 1997, 14(2): 9-12. (in Chinese)

[2] 陈锋，杨树勤. 疾病的复合分布模型研究——疾病的统计分布（一）[J]. 中国卫生统计，1995，12（6）：12-15.
CHEN F, YANG S Q. Study on compound distribution model of disease: statistical distribution of diseases (I)[J]. Chinese Journal of Health Statistics, 1995, 12(6): 12-15. (in Chinese)

[3] CASELLA G, BERGER R L. 统计推断[M]. 张忠占，傅莺莺，译. 北京：机械工业出版社，2010.
CASELLA G, BERGER R L. Statistical inference[M]. Translated by ZHANG Z Z, FU Y Y. Beijing: China Machine Press, 2010. (in Chinese)

[4] 胡毓达. 最优化理论与算法[M]. 北京：清华大学出版社，1989.
HU Y D. Optimization theory and algorithm[M]. Beijing: Tsinghua University Press, 1989. (in Chinese)

[5] McLACHLAN G J. The EM algorithm and extensions[M]. 2nd ed. New York: Wiley & Sons, Inc., 2008.

[6] 全星澄，李巍. 基于 EM 算法的有限维混合分布参数估计研究[J]. 统计与决策，2017（12）：25-29.
QUAN X C, LI W. Research on parameter estimation of finite dimension mixture distribution based on EM algorithm[J]. Statistics and Decision, 2017(12): 25-29. (in Chinese)

[7] LANGE N, BILLARD L, CONQUEST L, et al. Case studies in biometry[M]. New York: Wily-Interscience, 1994.