DOI:10.13546/j.cnki.tjyjc.2021.13.005

MCMC算法在多维混合数据参数识别中的应用

石　凯[1a],李　杰[2],刘洪江[1b]

(1.乐山师范学院a.数理学院;b.旅游学院,四川 乐山 614000;2.西南财经大学 统计学院,成都 611130)

摘　要:混合数据参数识别问题的研究一直备受关注,由于混合类型和混合权重无法直接观测,因此其实质是含有隐变量的不完全数据或者缺失数据。处理此类数据的难点在于参数的估计和识别,尤其是多维取值空间,会面临待估参数多、似然函数复杂等情况。文章给出了多维高斯分布假设下MCMC算法具体实施流程,并通过一个计算机模拟的三类别二维混合数据进行了实证研究,结果显示:MCMC算法达到了高度精确的区分效果,参数的样本估计值接近模拟生成的真值,能够为混合数据参数估计问题提供有效解决途径。

关键词:MCMC算法;多维混合数据;贝叶斯统计;Gibbs抽样

中图分类号:TP274.2　　**文献标识码:**A　　**文章编号:**1002-6487(2021)13-0024-04

0　引言

混合数据的研究最早源于1894年Pearson对两个生物物种数据混合的识别处理,因其适用范围广,现已被大量应用于医学、物理、生物、计算机、天文、经济、社会等学科之中。Hastie和Tibshirani(1996)[1]运用混合分布对手写数字进行了分类识别,Bailey和Elkan(1995)[2]将二元混合模型用于生物聚合物的基序发现,Makenna等(1998)[3]以高斯混合模型模拟人体脸部特征的跟踪,童佳宁(2010)[4]将混合数据用于语音信号处理中抗噪声性能判别,任长宏等(2011)[5]采用线性混合模型解决了生物科学中植物育种无重复实验数据的统计测验问题。

若以自然数 k 表示混合数据中类别的个数,$w_j(j=1,2,\cdots,k$,且 $\sum_{j=1}^{k}w_j=1)$ 表示混合的构成比例,$f(x;\theta_j)(j=1,2,\cdots,k$,且 $f(x;\theta_j)\geq 0$,$\int_{-\infty}^{\infty}f(x;\theta_j)\mathrm{d}x=1)$ 为各混合类别关于参数 θ_j 的分布密度函数,则混合数据背后对应的概率分布模型为:

$$f(x)=\sum_{j=1}^{k}w_j f(x;\theta_j)\ ,x\in R \tag{1}$$

容易验证式(1)中 $f(x)$ 满足概率分布的非负性、归一性等基本性质。根据数据分布情况,在给定混合类别个数 k 和设定了 $f(x;\theta_j)$ 类型后(对于连续型随机变量,常以高斯正态分布作为各类别分量分布假定,对应混合模型称为GMM),待估参数为 $(w_j,\ \theta_j)$ 。处理混合数据的难点在于参数的估计识别上,由于在一组容量为 n 的混合样本数据

基金项目:四川省教育厅人文社会科学基金资助项目(18SB0223);乐山师范学院校级学科建设重点科研项目(WZD016)

作者简介:石　凯(1976—),男,四川泸州人,博士,副教授,研究方向:应用统计。

李　杰(1986—),男,四川阆中人,博士研究生,研究方向:经济统计。

刘洪江(1971—),男,重庆人,博士,教授,研究方向:灾害大数据。

Likelihood Test and Model Comparison of Zero-and-one-inflated Poisson Model

Liu Yu[1], An Bowen[1], Tian Maozai[1,2a,2b,3]

(1.School of Statistics and Data Sciences, Xinjiang University of Finance and Economics, Urumqi 830012, China; 2.a.School of Statistics, b. Center for Applied Statistics, Renmin University of China, Beijing 100872, China; 3.School of Statistics, Lanzhou University of Finance and Economics, Lanzhou 730020, China)

Abstract: Aiming at the problem that it is difficult to determine the data type due to the simultaneous inflation of 0 and 1 values in complex counting data, this paper constructs Wald test, LR test and Score test statistics based on zero-and-one-inflated Poisson distribution model. In practical application, the EM algorithm is used to solve the parameter estimation values and confidence interval in order to popularize the zero-and-one-inflated Poisson regression model. The paper also makes analyses on the occurrence times of snowstorms. The results of the test statistics prove that the data has simultaneous inflation of 0 and 1 values, and determine that the optimal model suitable for this data is the zero-and-one-inflated Poisson model.

Key words: zero-and-one-inflated Poisson model; Wald test; LR test; Score test; extreme weather

$(x_1, x_2, \cdots, x_n)$ 中，各样本点 $x_i(i=1,2,\cdots,n)$ 所对应的类别是缺失的（用 z_i 表示，其中 $z_i \in (1,2,\cdots,k)$，对应的分布即为 $P(z_i=j)=w_j$），所以混合数据也称为缺失数据，或者不完全数据。

目前，关于混合模型的参数估计方法主要有EM算法，即通过期望E步求缺失变量分布，然后最大化样本似然函数M步求得参数的估计值，反复迭代直至收敛[6]。EM算法实施简单，但存在的缺点是可能收敛较慢，且不能保证收敛到全局最大。本文尝试引入近年来关注热点的MCMC算法（Markov Chain Monte Carlo），以期为解决混合数据参数识别问题提供更为丰富的计算方法。虽然Rasmussen（2000）[7]、尤芳（2006）[8]、王平（2011）[9]曾做过此类问题的研究，但大都停留在一维数据的推导和实验上，本文更进一步在多维空间上给出MCMC算法的实施过程，并以多维数据集进行实验验证，使其更具有普遍的适用价值。

1 多维混合高斯分布的MCMC算法

1.1 多维混合高斯分布的MCMC算法原理

在 p 维实数空间中，将多维高斯分布的密度函数代入式（1），则式（1）可改写为如下形式：

$$f(x)=\sum_{j=1}^{k} w_j (2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right],\ x \in R^p \tag{2}$$

式（2）中 μ_j 为均值向量，Σ_j 为协方差矩阵，待估参数为 $(w_j, \mu_j, \Sigma_j; j=1,2,\cdots,k)$。给定一组容量为 n 的 p 维混合样本数据 $(x_1, x_2, \cdots, x_n)$，其所对应的样本似然函数为：

$$\prod_{i=1}^{n} f(x_i; w_j, \mu_j, \Sigma_j) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} w_j (2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)\right] \right\} \tag{3}$$

根据数理统计理论，极大化式（3）就能得到参数的估计值。然而，鉴于式（3）的复杂性，无法通过一阶条件求解显示解析解的具体形式，甚至由于不是单峰函数，常用的数值迭代算法（如Newton-Raphson算法、Quasi-Newton算法等）也会失效。而MCMC算法则为解决此类问题提供了一种行之有效的途径，其基本原理是以Bayes理论为基础，利用一个平稳分布的马尔科夫链来获取样本，从而以蒙特卡洛方法进行统计推断。常用于构造马尔科夫链的主要方法有Gibbs抽样和Metropolis-Hasting算法，其中Gibbs抽样特别适合处理本文涉及的缺失数据问题[10]。

根据Bayes理论，将混合模型中的待估参数 $(w_j, \mu_j, \Sigma_j; j=1,2,\cdots,k)$ 视为具有分布的随机向量，利用先验分布和样本似然函数则可以得出后验分布信息。然而多元随机向量的联合分布难以确定，Gibbs抽样的特点是通过一元（或者低维）分布的抽样来获得目标多元分布本身的样本。将Gibbs抽样运用于多维混合高斯分布模型的具体情形如下：

（1）定义完全样本数据为 $(x_i, z_i)=\{(x_1, z_1),(x_2, z_2), \cdots, (x_n, z_n)\}$，其中 $x_i(i=1,2,\cdots,n)$ 是可观测数据，$z_i \in (1,2,\cdots,k), i=1,2,\cdots,n$ 是不可观测的数据，k 为混合数据所属类型。

（2）w 与 (μ, Σ) 相互独立，w 的共轭先验分布为Dirichlet分布，即 $w \sim D(\alpha_{1.0}, \alpha_{2.0}, \cdots, \alpha_{k.0})$，其中 $(\alpha_{1.0}, \alpha_{2.0}, \cdots, \alpha_{k.0})$ 为先验分布中的超参数。

（3）各个正态类别的 $(\mu_j, \Sigma_j), j=1,2,\cdots,k$ 相互独立，共轭先验分布为正态-逆维希特分布（Normal-Inverse Wishart），给定 Σ_j 时，$\mu_j | \Sigma_j \sim N_p(\gamma_{j.0}, \Sigma_j/m_{j.0})$，其中 $(\gamma_{j.0}, m_{j.0})$ 是超参数，$m_{j.0}>0$。给定 μ_j 时，$\Sigma_j^{-1} | \mu_j \sim W(h_{j.0}, v_{j.0})$，其中 W 是维希特分布，超参数 $(h_{j.0}, v_{j.0})$ 中，$h_{j.0}$ 表示自由度，$v_{j.0}$ 是 $p \times p$ 维的正定矩阵。

（4）令 G_j 为样本数据属于第 j 类别的集合，$j=1,2,\cdots,k$，d_j 表示第 j 类包含样本数据的个数，则在上述设定下，随机向量 $(w_j, \mu_j, \Sigma_j, z_i; j=1,2,\cdots,k, i=1,2,\cdots,n)$ 各分量对应的后验满条件分布分别为：

$$(\mu_j | \Sigma_j, x_i, z_i) \sim N_p(\gamma_j, \Sigma_j/m_j) \tag{4}$$

其中，$\gamma_j=(m_{j.0}\gamma_{j.0}+d_j\bar{x}_j)$，$\bar{x}_j=\frac{1}{d_j}\sum_{i \in G_j} x_i$，$m_j=m_{j.0}+d_j$。

$$(\Sigma_j | \mu_j, x_i, z_i) \sim W^{-1}(h_j, v_j) \tag{5}$$

其中，$h_j=h_{j.0}+d_j$，$v_j=v_{j.0}+S_j+d_j m_{j.0}(\bar{x}_j-\gamma_j)(\bar{x}_j-\gamma_j)^T/m_j$，$S_j=\sum_{i \in G_j}(x_i-\bar{x}_j)(x_i-\bar{x}_j)^T$。

$$(w | z_i) \sim D(\alpha_1, \alpha_2, \cdots, \alpha_k) \tag{6}$$

其中，$\alpha_j=\alpha_{j.0}+d_j$。

$$P(z_i=j | \mu, \Sigma, w, x_i) \propto w_j N_p(x_i | \mu_j, \Sigma_j) \tag{7}$$

其中，$\propto$ 表示正比于符号，z_i 的分布是一个离散型分布。

1.2 基于Gibbs抽样的MCMC算法实施步骤

通过以上各个参数的后验满条件分布，在给定初始值后，就可以进行后验抽样，具体实施步骤如下：

步骤1：由式（6）在Dirichlet分布中抽取混合权重 w 的值。

步骤2：由式（7）离散型分布抽取 z_i，$i=1,2,\cdots,n$ 的值。

步骤3：由式（4）在 p 维高斯分布中抽取 $\mu_j, j=1,2,\cdots,k$ 的值。

步骤4：由式（5）在逆维希特分布中抽取 $\Sigma_j, j=1,2,\cdots,k$ 的值。

步骤5：对步骤1至步骤4反复迭代生成马尔科夫链，为消除初值设定的影响，舍弃开始一段的样本，取其波动稳定的余下样本作为参数分布的取样。

步骤6：依据Monte Carlo原理，计算遍历的样本均值即可得到参数的估计值。

2 实验模拟分析

2.1 数据设计与描述

在计算机上模拟生成三个类别 ($k=3$) 的二维 ($p=2$) 混合数据，样本容量 $n=2000$，将各参数真值设定为：

$w=(w_1, w_2, w_3)=(0.35, 0.5, 0.15)$

$\mu_1=(-3, 3)$，$\mu_2=(2, 2)$，$\mu_3=(-1, 0)$

$\Sigma_1=\begin{pmatrix}1 & 0.5\\ 0.5 & 0.5\end{pmatrix}$，$\Sigma_2=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$，$\Sigma_3=\begin{pmatrix}0.2 & 0\\ 0 & 0.2\end{pmatrix}$

以 x_i 的两个分量绘制二维平面图形，以反映样本的分布情况（见图1）。可以发现，样本数据有3个集中区域（由位置参数决定，即均值向量 μ_j），各个区域的样本点聚集数量不同（由比例参数决定，即 w_j），而且分布形状和分散程度各异（由尺度参数决定，即协方差矩阵 Σ_j）。

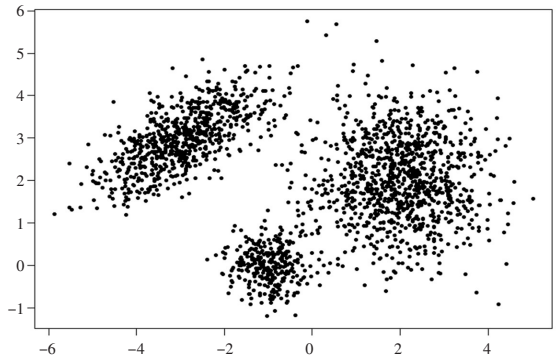

图1 混合样本数据分布图

2.2 MCMC算法的参数估计

对先验分布中的超参数进行设置，w 的先验分布为Dirichlet分布，即 $w\sim D(\alpha_{1.0}, \alpha_{2.0}, \alpha_{3.0})$，取 $(\alpha_{1.0}, \alpha_{2.0}, \alpha_{3.0})=(1, 1, 1)$；$\mu_j\big|\Sigma_j \sim N_p(\gamma_{j.0}, \Sigma_j/m_{j.0})$，取 $\gamma_{1.0}=\gamma_{2.0}=\gamma_{3.0}=\bar{x}$，$\bar{x}$ 为所有样本数据的均值向量；$\Sigma_j^{-1}\big|\mu_j$ 服从维希特分布 $W(h_{j.0}, v_{j.0})$，取 $(h_{1.0}, h_{2.0}, h_{3.0})=(1, 1, 1)$，$v_{1.0}=v_{2.0}=v_{3.0}=\mathrm{cov}(x)$，$\mathrm{cov}(x)$ 为所有样本数据的协方差矩阵。实验迭代次数设定为1000次，迭代初值从 $w^{(0)}=(1/3, 1/3, 1/3)$，$\mu_1^{(0)}=\mu_2^{(0)}=\mu_3^{(0)}=\bar{x}$，$\Sigma_1^{(0)}=\Sigma_2^{(0)}=\Sigma_3^{(0)}=\mathrm{cov}(x)$ 开始。

经过1000次迭代后，各参数后验取样的马尔科夫链用图2表示，限于篇幅，仅给出第一类混合元的均值向量各分量的马氏链，用 $(\mu_{1.1}, \mu_{1.2})$ 表示，第一类混合元的协方差矩阵中各分量的马氏链用 $(\Sigma_{1.11}, \Sigma_{1.12}, \Sigma_{1.21}, \Sigma_{1.22})$ 表示，(w_1, w_2, w_3) 各分量的马氏链以及其他参数后验取样的马氏链类似，仅是收敛的数值不同。

从图2可知，尽管各个类型给定的参数迭代初始值都一样，但在Gibbs抽样下，均收敛到其各自的参数真值附近，说明MCMC算法能够对多维混合分布数据进行有效区分和识别。去掉各参数马氏链前300个样本，用余下的样本计算均值，作为各参数的估计值，并与模拟生成参数的真值进行对比。根据有限样本计算的估计值与真值均相

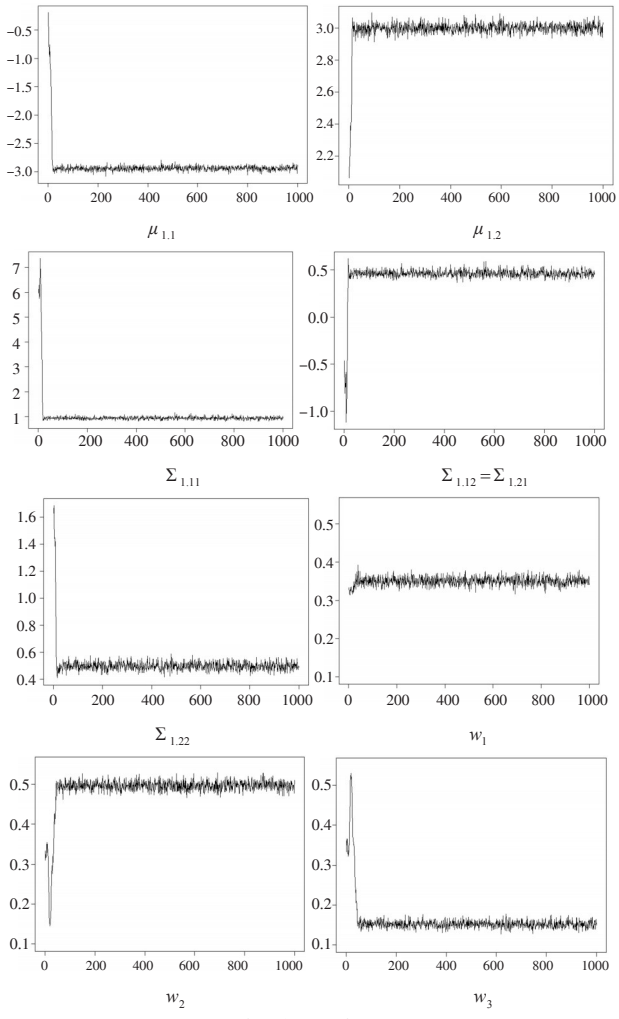

图2 Gibbs抽样的1000次迭代马氏链过程

当接近，进一步验证了利用MCMC算法进行区分和识别的有效性。将参数估值代入式(2)，即可得出多维混合分布密度函数 $f(x)$ 的具体形式，据此绘制二维投影图（见图3）。图3中阴影部分即为混合模型密度函数 $f(x)$ 的投影，可以看出，样本数据在各个区域上的分布情况的区分效果良好。

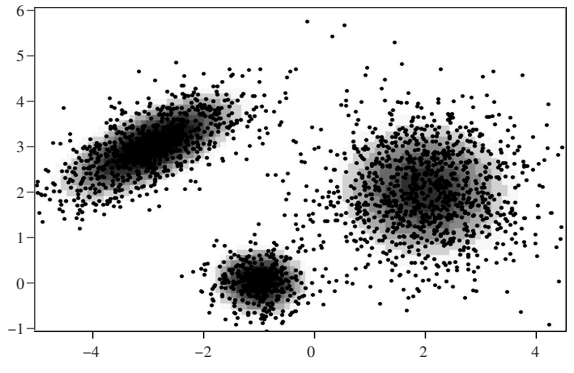

图3 样本数据与混合高斯分布密度函数的投影图

3 结束语

混合数据在自然科学和社会科学的众多领域中应用广泛，但其难点在于参数的估计和类型的识别上，而且由

于现实数据往往是多维的，实际工作中势必会面临似然函数复杂、待估参数量大等问题，导致一些传统的数值计算方法失效。鉴于此，本文采用MCMC算法为多维混合分布数据的参数估计和识别提供了一种有效的解决途径。MCMC算法本质是一种概率抽样计算，利用一个平稳分布的马尔科夫链来获取样本，进而做出统计推断，其中以Gibbs抽样最为常用。本文在给出多维混合高斯分布参数的后验Gibbs抽样流程基础上，通过一个计算机模拟的三类别二维混合数据进行了实证研究。结果显示，基于Gibbs抽样的MCMC算法尤其适合混合数据的处理，对各个类型混合比例的区分以及各个类型参数估计的效果良好，相比其他算法，具有理论基础较强、初始条件限定较少等优点，是一种行之有效的计算方法。进一步的研究方向是将其运用于更多领域多维数据的聚类和判别分析之中。

参考文献:

[1]Hastie T J, Tibshirani R J. Discriminate Analysis by Gaussian Mixtures [J].Journal of the Royal Statistical Society Series B,1996,(1).

[2]Bailey T L, Elkan C. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization [J].Machine Learning Journal,1995,(21).

[3]Makenna S J, Gong S, Raja Y. Modeling Facial Color Identity With Gaussian Mixtures [J].Pattern Recognition,1998,(12).

[4]童佳宁.基于HMM和PNN的混合语音识别模型研究[D].邯郸:河北工程大学学位论文,2010.

[5]任长宏,胡希远,李建平.线性混合模型在作物育种无重复试验数据实证分析中的应用[J].西北农林科技大学学报(自然科学版),2011,39(2).

[6]Figueiredo M A T, Jain A K. Unsupervised Learning of Finite Mixture Models [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,(3).

[7]Rasmussen C E. The Infinite Gaussian Mixtrure Model [J].Neural Information Processing Systems,2000,(12).

[8]尤芳.混合模型的参数估计[D].苏州:苏州大学学位论文,2006.

[9]王平.自然数集上的Dirichlet过程以及无限正态混合模型[D].南京:东南大学学位论文,2011.

[10]Geman S, Geman D. Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1984,(6).

（责任编辑/浩　天）

Application of MCMC Algorithm in Parameter Identification of Multidimensional Mixed Data

Shi Kai[1a], Li Jie[2], Liu Hongjiang[1b]

(1.a.School of Mathematics and Physics, b. School of Tourism, Leshan Normal University, Leshan Sichuan 614000, China;
2. School of Statistics, Southwest University of Finance and Economics, Chengdu 611130, China)

Abstract: The research on parameter identification of mixed data has been paid much attention. Because mixed type and mixed weight cannot be directly observed, the essence of mixed data is incomplete data or missing data with hidden variables. The difficulty in processing such data is the estimation and identification of parameters, and especially in the multidimensional value space, which will face the situation of many parameters to be estimated and complex likelihood function. This paper presents the specific implementation process of MCMC algorithm under the assumption of multidimensional Gaussian distribution, and conducts an empirical study through a computer simulation of three categories of 2D mixed data. The results show that the MCMC algorithm achieves highly accurate distinction effect, and that the sample estimate of parameters is close to the true value generated by simulation, which can provide an effective solution to the problem of parameter estimation of mixed data.

Key words: MCMC algorithm; multidimensional mixed data; Bayesian statistics; Gibbs sampling