Statistics and Application 统计学与应用, 2017, 6(4), 482-491
Published Online October 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.64054

Hans汉斯

Comparative Study on Effects of Parameter Estimation of Mixture Models under Different Types of Data

Xiaoying Wang, Yinghua Li, Xuemei Yang

School of Mathematics and Physics, North China Electric Power University, Beijing
Email: liyinghua327@126.com

Received: Oct. 8th, 2017; accepted: Oct. 23rd, 2017; published: Oct. 30th, 2017

Abstract

The normal mixture model has more applications in describing data. But it is easily influenced by the outlier, and the maximum likelihood estimation of parameters is not robust estimation. T-distribution mixture model has better robustness than Gauss mixture model to analyze data with longer time than normal tails because of its heavy-tails. In this paper, we studied a univariate t mixture model primarily. Based on EM algorithm, we derived the iteration steps of maximum likelihood estimation of the model's unknown parameters. Furthermore, we did a comparative analysis by three types of simulated data. Simulation study shows that this model has an advantage in fitting data with longer time than normal tails. The initial value is given by k-means method.

Keywords

EM Algorithm, Mixture T-Distribution Model, K-Means Initialization

不同类型数据下混合模型参数估计效果的对比研究

王小英，李迎华，杨雪梅

华北电力大学数理学院，北京
Email: liyinghua327@126.com

收稿日期：2017年10月8日；录用日期：2017年10月23日；发布日期：2017年10月30日

摘　要

混合高斯模型在描述数据方面应用较多，但它易受离群点的影响，其参数的极大似然估计不是稳健估计。

文章引用: 王小英, 李迎华, 杨雪梅. 不同类型数据下混合模型参数估计效果的对比研究[J]. 统计学与应用, 2017, 6(4): 482-491. DOI: 10.12677/sa.2017.64054

混合t-分布模型由于其重尾分布的特性，相对于混合高斯分布，在分析重尾数据上更具稳健性。文章首先研究一元混合t-分布模型，利用标准EM算法给出了该模型参数极大似然估计的迭代步骤，并分别在三类模拟数据下与混合高斯模型进行了对比分析，验证了该模型的有效性以及在拟合重尾数据上的优势。算法初始化采用k-means方法。

关键词

EM算法，混合t-分布模型，k-Means初始化

1. 引言

混合分布模型是一种基于概率和统计的建模工具，它提供了一个用简单结构模拟复杂密度的灵活有效的方法，从而受到了统计学界、模式识别、图像处理等诸多领域的广泛关注。它的基本策略是，把研究数据看作是从很大的单个或多个总体上抽取出的一部分，通过潜在的分布或密度函数来描述。混合高斯模型应用的较多，但通常我们收集到的很多数据并不是严格的服从正态分布，而是较明显的服从重尾分布。混合 t-分布模型由于其具有较长的尾巴，可对重尾点和异常点有效的降低权值，因此，相对于混合高斯分布模型，可以获得较强的精度和稳健性。

对于模型参数的求解，1977 年 Dempster 等人在文献[1]中提出的 EM 算法成为了混合模型参数极大似然估计的极有效的工具。Peel 和 McLachlan 在文献[2]中指出 EM 算法可以获得有限混合任意分布模型的极大似然估计。对于单一的 t-分布，为了使 M 步更好求解，Meng 和 Rubin 在文献[3]中用受限制的最大化 CM 步来代替 M 步，得到 ECM 算法；Peel 和 McLachlan 在文献[2] [4]中提出多元混合 t-分布模型，用标准 EM 算法求解混合 t-分布模型参数的极大似然估计，并给出了 ECM 算法的一个应用；而后，Liu 和 Rubin 在文献[5]中对 ECM 算法进行两处修改，得到收敛速度更快的 ECME 算法。在算法初始化方面，冉延平在文献[6]中用 k-means 方法确定混合高斯分布的最大混合子分布数目以及混合比例；史鹏飞在文献[7]中通过 k-means 方法先给出混合数据的一个粗糙分组，然后根据分组数据给出参数的一个粗略估计值，作为混合高斯分布 EM 算法的迭代初始值。随着计算机性能的快速发展，混合 t-分布模型已应用于诸多领域。如杨云飞在文献[8]中提出了自适应均值滤波的多元 t-分布混合模型，对医学脑图像分割进行了研究；熊太松在文献[9]中对伯克利图像数据和微软剑桥研究院提供的图像集用视觉和量化对比的评估方式证明了基于空间平滑的 t-分布混合模型在灰度图像分割中的有效性；朱志娥在文献[10]中针对偏 t 正态数据、异方差和线性回归提出了偏 t 正态数据下混合线性联合位置与尺度模型，详细介绍了该模型下的 EM 算法并进行了有效的模拟验证。

在前人研究的基础上，本文研究了基于 EM 算法的三总体一元混合 t-分布模型参数的极大似然估计，克服了多元混合 t-分布模型中协方差矩阵向一元混合 t-分布模型中尺度参数转变的困难，并首次将 k-means 方法用于该模型下算法初值的选取。此外，我们引进了混合高斯模型，并分别在三种不同类型数据下进行了对比仿真实验，验证了本文研究的模型和方法的有效性及其在处理重尾数据上的优势。

2. 有限 t-分布混合模型

2.1. 一元学生 t-分布

设随机变量 y 服从一元学生 t-分布，记做 $y\sim t\left(y\left|\mu,\sigma,\nu\right.\right)$，概率密度函数定义为[11]：

$$t\left(y\left|\mu,\sigma,\nu\right.\right)=\frac{\Gamma\left(\frac{1+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{1}{\pi\sigma^2\nu}\right)^{\frac{1}{2}}\left[1+\frac{\left(y-\mu\right)^2}{\sigma^2\nu}\right]^{-\left(\frac{1+\nu}{2}\right)} \tag{1}$$

其中参数 μ 和 σ 分别表示 t-分布的位置参数和尺度参数，$\Gamma\left(\bullet\right)$ 表示伽马函数，参数 ν 称为 t-分布的自由度。

2.2. t-分布有限混合模型

设随机向量服从一元 t-分布且由 m 个子分布混合而成，混合比例为 π_k 且满足 $\sum_{k=1}^{m}\pi_k=1$，$t\left(y\left|\theta_k\right.\right)$ 为第 k 个子分布的概率密度函数，$\theta_k=\left\{\mu_k,\sigma_k,\nu_k\right\}$ 为对应的子分布的概率密度函数参数，$\Psi=\left\{\pi_1,\cdots,\pi_m,\theta_1,\cdots\theta_m\right\}$ 为参数空间。因此，有限混合 t-分布模型可以表示为

$$t\left(y\left|\theta\right.\right)=\sum_{k=1}^{m}\pi_k t\left(y\left|\theta_k\right.\right) \tag{2}$$

本文研究三个子分布的情况，即取 $m=3$，则式(2)化为

$$t\left(y\left|\theta\right.\right)=\pi_1 t\left(y\left|\theta_1\right.\right)+\pi_2 t\left(y\left|\theta_2\right.\right)+\left(1-\pi_1-\pi_2\right)t\left(y\left|\theta_3\right.\right) \tag{3}$$

其中，$t\left(y\left|\theta_k\right.\right)$ 为第 k 个子分布的概率密度函数，具体形式见式(1)。

3. 模型参数极大似然估计的 EM 算法

本文要研究的模型为上文所提到的式(3)，并假设三个子分布的自由度相同，即 $\nu_1=\nu_2=\nu_3=\nu$。我们采用标准 EM 算法来求解模型参数，它提供了一种近似计算含有隐变量概率模型的极大似然估计的方法。在 EM 算法的基本框架下，我们引入隐变量以得到完整数据集。完整数据集定义为 $Y_c=\left\{Y,Z,U\right\}$，其中，Z 为标签变量 $Z=\left\{z_1,z_2,\cdots,z_N\right\}$，且 $z_{ik}=\begin{cases}1, & \text{第}i\text{个样本来自第}k\text{个子分布}\\ 0, & \text{其他}\end{cases}$ $\left(i=1,2,\cdots,N,\ k=1,2,3\right)$；$U$ 为引进的另一个隐变量 $U=\left\{u_1,u_2,\cdots,u_N\right\}$，给定 $z_{ik}=1$ 时，u_i 是独立同分布的，且满足 $u_i\left|z_{ik}=1\sim G\left(\frac{\nu}{2},\frac{\nu}{2}\right)\right.$；$Y$ 为可观测数据集 $Y=\left\{y_1,\cdots,y_N\right\}$，且有 $y_i\left|u_i,z_{ik}=1\sim N\left(\mu_k,\sigma_k^2/u_i\right)\right.$。由(1) (3)两式建立完整数据的似然函数为：

$$\begin{aligned}p\left(Y,Z,U\left|\Psi\right.\right)&=\prod_{i=1}^{N}p\left(y_i,z_{i1},z_{i2},u_i\left|\Psi\right.\right)\\&=\prod_{i=1}^{N}\prod_{k=1}^{3}\left\{\frac{\sqrt{u_i}}{\sqrt{2\pi}\sigma_k}\mathrm{e}^{-\frac{u_i\left(y_i-\mu_k\right)^2}{2\sigma_k^2}}\left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\mathrm{e}^{-\frac{\nu}{2}u_i}u_i^{\frac{\nu}{2}-1}\right]\pi_k\right\}^{z_{ik}}\end{aligned} \tag{4}$$

则完整数据的对数似然函数为：

$$\begin{aligned}\log p\left(Y,Z,U\left|\psi\right.\right)&=\sum_{k=1}^{3}\sum_{i=1}^{N}z_{ik}\left\{\frac{1}{2}\log u_i-\frac{1}{2}\log 2\pi-\log\sigma_k-\frac{u_i}{2\sigma_k^2}\left(y_i-\mu_k\right)^2\right.\\&\left.+\frac{\nu}{2}\log\frac{\nu}{2}-\log\Gamma\left(\frac{\nu}{2}\right)-\frac{\nu}{2}u_i+\log\pi_k\right\}\\&=\sum_{k=1}^{3}\sum_{i=1}^{N}z_{ik}\log\pi_k\\&+\sum_{k=1}^{3}\sum_{i=1}^{N}z_{ik}\left\{-\log\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\left(\log u_i-u_i\right)\right\}\\&+\sum_{k=1}^{3}\sum_{i=1}^{N}z_{ik}\left\{-\frac{1}{2}\log 2\pi-\frac{1}{2}\log u_i-\log\sigma_k-\frac{u_i\left(y_i-\mu_k\right)^2}{2\sigma_k^2}\right\}\end{aligned}\tag{5}$$

EM 算法是一种迭代近似求解算法，它主要分两步进行：E 步是对对数似然函数求期望，M 步是最大化对数似然函数以获得新的参数值。

应用 EM 算法于上式，求解第 j 次各参数的极大似然更新表达式。

E 步：$Q\left(\Psi\left|\Psi^{(j)}\right.\right)=E_{Z,U}\left[\log p\left(Y,Z,U\left|\Psi\right.\right)\left|Y,\Psi^{(j)}\right.\right]$

首先计算关于隐变量 Z，U 的条件分布的期望：

$$E_{Z,U}\left[z_{ik}\left|Y,\Psi^{(j)}\right.\right]=\tau_{ik}^{(j)}=\frac{\pi_k^{(j)}f_k\left(y_i\left|\theta_k^{(j)}\right.\right)}{\sum_{k=1}^{3}\pi_k^{(j)}f_k\left(y_i\left|\theta_k^{(j)}\right.\right)}\tag{6}$$

$$E_{Z,U}\left[u_{ik}\left|y,\Psi^{(j)}\right.\right]=u_{ik}^{(j)}=\frac{\nu^{(j)}+1}{\nu^{(j)}+\left(y_i-\mu_k^{(j)}\right)^2\Big/\sigma_k^{2(j)}}\tag{7}$$

$$E_{Z,U}\left[\log u_{ik}\left|y,\Psi^{(j)}\right.\right]=l_{ik}^{(j)}=\log u_{ik}^{(j)}+\left\{\psi\left(\frac{\nu^{(j)}+1}{2}\right)-\log\left(\frac{\nu^{(j)}+1}{2}\right)\right\}\tag{8}$$

其中，$\psi\left(x\right)=\dfrac{\mathrm{d}\left(\log\Gamma\left(x\right)\right)}{\mathrm{d}x}=\dfrac{\Gamma'\left(x\right)}{\Gamma\left(x\right)}$，于是

$$\begin{aligned}Q\left(\Psi\left|\Psi^{(j)}\right.\right)&=\sum_{k=1}^{3}\sum_{i=1}^{N}\tau_{ik}^{(j)}\log\pi_k\\&+\sum_{k=1}^{3}\sum_{i=1}^{N}\tau_{ik}^{(j)}\left\{-\log\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\left(\log u_{ik}^{(j)}+\Psi\left(\frac{\nu^{(j)}+1}{2}\right)-\log\left(\frac{\nu^{(j)}+1}{2}\right)-u_{ik}^{(j)}\right)\right\}\\&+\sum_{k=1}^{3}\sum_{i=1}^{N}\tau_{ik}^{(j)}\left\{-\frac{1}{2}\log 2\pi-\log\sigma_k-\frac{u_{ik}^{(j)}\left(y_i-\mu_k\right)^2}{2\sigma_k^2}-\frac{1}{2}\left[\log u_{ik}^{(j)}+\Psi\left(\frac{\nu^{(j)}+1}{2}\right)-\log\left(\frac{\nu^{(j)}+1}{2}\right)\right]\right\}\end{aligned}\tag{9}$$

M 步：$\theta^{(j+1)}=\arg\max_{\theta}Q\left(\theta,\theta^{(j)}\right)$

利用 Q 函数对各参数求偏导数并令其等于零，求解得到各参数的第 $j+1$ 次迭代更新表达式：

$$\pi_k^{(j+1)}=\sum_{i=1}^{N}\tau_{ik}^{(j)}\Big/N\tag{10}$$

$$\mu_k^{(j+1)}=\frac{\sum_{i=1}^{N}\tau_{ik}^{(j)}u_{ik}^{(j)}y_i}{\sum_{i=1}^{N}\tau_{ik}^{(j)}u_{ik}^{(j)}}\tag{11}$$

$$\sigma_k^{(j+1)}=\text{sqrt}\left(\frac{\sum_{i=1}^{N}\tau_{ik}^{(j)}u_{ik}^{(j)}\left(y_i-\mu_k^{(j+1)}\right)^2}{\sum_{i=1}^{N}\tau_{ik}^{(j)}u_{ik}^{(j)}}\right) \tag{12}$$

自由度$\nu^{(j+1)}$是非线性方程

$$\log\frac{\nu}{2}-\psi\left(\frac{\nu}{2}\right)+1+\frac{1}{\sum_{i=1}^{N}\tau_{ik}^{(j)}}\sum_{i=1}^{N}\tau_{ik}^{(j)}\left(l_{ik}^{(j)}-u_{ik}^{(j)}\right)=0 \tag{13}$$

的解。上式是关于ν的非线性方程，文献[5]中采用搜索ν的空间求出ν的估计值，但计算量大。文献[12]中作者给出了一个计算量相对较小的可直接计算多元混合 t-分布模型中参数ν近似解的方法，但其不稳定。因此，方便起见，在接下来的数值模拟中我们只考虑自由度是已知的情形，即每一子分布的自由度都相同且固定为ν。

4. 数值模拟

为了验证上述参数估计方法的有效性，我们共采用三种不同类型数据进行模拟研究，并引进混合高斯分布模型[13]与之作对比。由 t-分布的方差与尺度参数的关系$\upsilon^2=\frac{\nu}{\nu-2}\sigma^2$，将混合 t-分布 EM 算法参数估计结果中的尺度参数σ转化为标准差υ，再与混合高斯分布 EM 算法估计的参数υ作比较。算法的初始化均采用 k-means 方法。参数估计的精确度采用均方误差来衡量，如混合比例π_1的均方误差定义为：

$$MSE\left(\pi_1\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\pi_{1i}-\pi_{1(0)}\right)^2$$

其中，$\pi_{1(0)}$是π_1的真值，n 为模拟次数。

4.1. 混合高斯分布数据

给定真值$\pi_{1(0)}=0.5$和$\pi_{2(0)}=0.3$，$\mu_{1(0)}=2$、$\mu_{2(0)}=7$、$\mu_{3(0)}=11$、$\upsilon_{1(0)}=\upsilon_{2(0)}=\upsilon_{3(0)}=1$，分别取样本量$N=500$，1000，共产生 2 组混合高斯分布数据。对混合 t-分布模型，分别取自由度$\nu=3$ [14]，15，30。重复模拟 1000 次，模拟结果如下：

由表 1、表 2、表 3 知：

$\nu=3$时，混合高斯模型参数估计的均方误差均比混合 t-分布模型参数估计的均方误差小，这一点在υ_1、υ_2、υ_3上更为明显；在$\nu=15,30$时，除了对υ_1、υ_2、υ_3的估计结果混合高斯模型略好于混合 t-分布模型外，两种方法对其他参数的估计的均方误差，几乎无差。此外还有，随着自由度的增大，混合 t-分布模型参数估计的均方误差变小；整体来看，样本量越大，MSE 越小。

4.2. 混合 t-分布数据

给定真值$\pi_{1(0)}=0.5$和$\pi_{2(0)}=0.3$，$\mu_{1(0)}=2$、$\mu_{2(0)}=7$、$\mu_{3(0)}=11$、$\upsilon_{1(0)}=\upsilon_{2(0)}=\upsilon_{3(0)}=1$，取$\nu=3$ [14]，15，30。分别取样本量$N=500$，1000，共产生 6 组混合 t-分布数据。重复模拟 1000 次，模拟结果见下表：

由表 4、表 5、表 6 知：

混合 t-分布模型可以较好地拟合该数据，参数估计值与真值十分接近。当$\nu=3$时，对所有参数的估计，混合 t-分布模型参数估计的均方误差均比混合高斯分布模型参数估计的均方误差小；$\nu=15$时，除μ_2、

Table 1. The simulation results of $\nu = 3$
表 1. $\nu = 3$ 的模拟结果

参数	N	EST1	EST2	MSE1	MSE2
$\pi_{1(0)} = 0.5$	500	0.4974	0.5011	0.0000	0.0000
	1000	0.4976	0.5014	0.0000	0.0000
$\pi_{2(0)} = 0.3$	500	0.2984	0.3063	0.0001	0.0002
	1000	0.2979	0.3057	0.0001	0.0001
$\mu_{1(0)} = 2$	500	1.9840	1.9974	0.0045	0.0050
	1000	1.9881	2.0033	0.0025	0.0026
$\mu_{2(0)} = 7$	500	6.9485	7.0348	0.0119	0.0120
	1000	6.9493	7.0357	0.0084	0.0072
$\mu_{3(0)} = 11$	500	10.9590	11.0208	0.0174	0.0178
	1000	10.9515	11.0123	0.0124	0.0092
$\upsilon_{1(0)} = 1$	500	0.9816	1.4078	0.0022	0.1730
	1000	0.9855	1.4043	0.0013	0.1723
$\upsilon_{2(0)} = 1$	500	0.9921	1.5314	0.0053	0.3081
	1000	0.9917	1.5311	0.0036	0.2960
$\upsilon_{3(0)} = 1$	500	1.0230	1.3980	0.0061	0.1803
	1000	1.0279	1.4030	0.0049	0.1750

Table 2. The simulation results of $\nu = 15$
表 2. $\nu = 15$ 的模拟结果

参数	N	EST1	EST2	MSE1	MSE2
$\pi_{1(0)} = 0.5$	500	0.4974	0.4998	0.0000	0.0000
	1000	0.4976	0.5001	0.0000	0.0000
$\pi_{2(0)} = 0.3$	500	0.2984	0.3019	0.0001	0.0001
	1000	0.2979	0.3013	0.0001	0.0001
$\mu_{1(0)} = 2$	500	1.9840	1.9967	0.0045	0.0044
	1000	1.9881	2.0018	0.0025	0.0023
$\mu_{2(0)} = 7$	500	6.9485	7.0102	0.0119	0.0100
	1000	6.9493	7.0114	0.0084	0.0056
$\mu_{3(0)} = 11$	500	10.9590	11.0054	0.0174	0.0171
	1000	10.9515	10.9984	0.0124	0.0088
$\upsilon_{1(0)} = 1$	500	0.9816	1.0127	0.0022	0.0029
	1000	0.9855	1.0172	0.0013	0.0015
$\upsilon_{2(0)} = 1$	500	0.9921	1.0407	0.0053	0.0124
	1000	0.9917	1.0391	0.0036	0.0042
$\upsilon_{3(0)} = 1$	500	1.0230	1.0116	0.0061	0.0099
	1000	1.0279	1.0158	0.0049	0.0059

Table 3. The simulation results of $\nu = 30$
表 3. $\nu = 30$ 的模拟结果

参数	N	EST1	EST2	MSE1	MSE2
$\pi_{1(0)} = 0.5$	500 1000	0.4974 0.4976	0.4998 0.5001	0.0000 0.0000	0.0000 0.0000
$\pi_{2(0)} = 0.3$	500 1000	0.2984 0.2979	0.3011 0.3005	0.0001 0.0001	0.0001 0.0001
$\mu_{1(0)} = 2$	500 1000	1.9840 1.9881	1.9965 2.0013	0.0045 0.0025	0.0043 0.0023
$\mu_{2(0)} = 7$	500 1000	6.9485 6.9493	7.0044 7.0059	0.0119 0.0084	0.0098 0.0054
$\mu_{3(0)} = 11$	500 1000	10.9590 10.9515	11.0012 10.9949	0.0174 0.0124	0.0168 0.0088
$\upsilon_{1(0)} = 1$	500 1000	0.9816 0.9855	1.0008 1.0053	0.0022 0.0013	0.0027 0.0013
$\upsilon_{2(0)} = 1$	500 1000	0.9921 0.9917	1.0161 1.0150	0.0053 0.0036	0.0098 0.0047
$\upsilon_{3(0)} = 1$	500 1000	1.0230 1.0279	1.0000 1.0040	0.0061 0.0049	0.0094 0.0054

注：EST1：混合高斯分布模型下的平均估计值。EST2：混合 t-分布模型下的平均估计值。MSE1：混合高斯分布模型参数估计的均方误差。MSE2：混合 t-分布模型参数估计的均方误差。

Table 4. The simulation results of $\nu = 3$
表 4. $\nu = 3$ 的模拟结果

参数	N	EST1	EST2	MSE1	MSE2
$\pi_{1(0)} = 0.5$	500 1000	0.4742 0.4792	0.4888 0.4952	0.0047 0.0024	0.0047 0.0019
$\pi_{2(0)} = 0.3$	500 1000	0.3119 0.3099	0.3021 0.3015	0.0018 0.0010	0.0013 0.0006
$\mu_{1(0)} = 2$	500 1000	1.7448 1.7567	1.9529 1.9369	0.5765 1.0027	0.2529 0.2239
$\mu_{2(0)} = 7$	500 1000	6.7244 6.7962	6.8751 6.9595	0.5162 0.2137	0.4994 0.2074
$\mu_{3(0)} = 11$	500 1000	11.1080 11.1236	10.9223 10.9733	0.3646 0.1072	0.2360 0.0816
$\upsilon_{1(0)} = 1.7321(\sigma_{1(0)} = 1)$	500 1000	1.3999 1.4025	1.7257 1.7216	0.2390 0.1964	0.0497 0.0184
$\upsilon_{2(0)} = 1.7321(\sigma_{2(0)} = 1)$	500 1000	1.2354 1.2257	1.7236 1.7178	0.5135 0.4336	0.0814 0.0352
$\upsilon_{3(0)} = 1.7321(\sigma_{3(0)} = 1)$	500 1000	1.4238 1.4863	1.7759 1.7539	0.2652 0.4228	0.1942 0.0740

Table 5. The simulation results of $\nu = 15$
表 5. $\nu = 15$ 的模拟结果

参数	N	EST1	EST2	MSE1	MSE2
$\pi_{1(0)} = 0.5$	500 1000	0.4953 0.4964	0.4995 0.5007	0.0006 0.0003	0.0006 0.0003
$\pi_{2(0)} = 0.3$	500 1000	0.2979 0.2987	0.2995 0.2996	0.0005 0.0003	0.0006 0.0003
$\mu_{1(0)} = 2$	500 1000	1.9711 1.9724	2.0004 1.9369	0.0056 0.0031	0.0047 0.0022
$\mu_{2(0)} = 7$	500 1000	6.9158 6.9216	6.9948 6.9982	0.0356 0.0121	0.0326 0.0058
$\mu_{3(0)} = 11$	500 1000	10.9621 10.9658	10.9958 10.9935	0.0270 0.0118	0.0291 0.0101
$\upsilon_{1(0)} = 1.0742(\sigma_{1(0)} = 1)$	500 1000	1.0343 1.0337	1.0716 1.0717	0.0041 0.0030	0.0036 0.0019
$\upsilon_{2(0)} = 1.0742(\sigma_{2(0)} = 1)$	500 1000	1.0198 1.0250	1.0700 1.0720	0.0085 0.0060	0.0132 0.0062
$\upsilon_{3(0)} = 1.0742(\sigma_{3(0)} = 1)$	500 1000	1.0713 1.0702	1.0715 1.0753	0.0079 0.0041	0.0156 0.0068

Table 6. The simulation results of $\nu = 30$
表 6. $\nu = 30$ 的模拟结果

参数	N	EST1	EST2	MSE1	MSE2
$\pi_{1(0)} = 0.5$	500 1000	0.4954 0.4969	0.4992 0.5002	0.0009 0.0003	0.0006 0.0003
$\pi_{2(0)} = 0.3$	500 1000	0.2988 0.2980	0.2999 0.2999	0.0006 0.0003	0.0006 0.0003
$\mu_{1(0)} = 2$	500 1000	1.9782 1.9820	1.9990 2.0005	0.0071 0.0031	0.0056 0.0029
$\mu_{2(0)} = 7$	500 1000	6.9239 6.9310	6.9908 6.9984	0.0825 0.0121	0.0800 0.0166
$\mu_{3(0)} = 11$	500 1000	10.9558 10.9549	10.9939 10.9908	0.0354 0.0118	0.0350 0.0116
$\upsilon_{1(0)} = 1.0351(\sigma_{1(0)} = 1)$	500 1000	1.0085 1.0098	1.0313 1.0331	0.0042 0.0030	0.0033 0.0024
$\upsilon_{2(0)} = 1.0351(\sigma_{2(0)} = 1)$	500 1000	1.0116 1.0066	1.0309 1.0284	0.0113 0.0060	0.0110 0.0047
$\upsilon_{3(0)} = 1.0351(\sigma_{3(0)} = 1)$	500 1000	1.0484 1.0487	1.0317 1.0337	0.0102 0.0041	0.0168 0.0053

υ_2、υ_3 外，混合 t-分布模型参数估计的均方误差比混合高斯分布模型参数估计的均方误差小；在 $\nu = 30$ 时，除 υ_3 外，混合 t-分布模型参数估计的均方误差均比混合高斯分布模型参数估计的均方误差小。此外，随着自由度的增大，混合 t-分布模型参数估计的均方误差变小；整体来看，样本量越大，MSE 越小，估计结果越好。

4.3. 含噪声的混合高斯数据

混合 t-分布模型相对于混合高斯模型有着较好的稳健性，这种稳健性尤其体现在对重尾数据(含噪声点、异常点数据)的处理。而处理重尾数据的另一种方法是在高斯分布的基础上添加一个均匀分布的成分[6]。因此，我们在双高斯数据的基础上添加一个均匀分布的部分作为重尾数据，再分别用混合 t-分布模型和混合高斯模型进行拟合并作比较。因为前两小节已经对自由度、样本量进行了研究比较，在这一小节我们不再考虑此二者的影响。取噪声所占比例分别为 5%和 10%，混合比例 $\pi_1=0.5$、$\pi_2=0.3$，自由度 $\nu=15$，样本量 $N=1000$。重复模拟 1000 次，模拟结果如表 7 和表 8。

由表 7、表 8 知：

通过比较两种模型下参数的估计结果和均方误差我们可以得到，混合 t-分布模型对该类型数据拟合的较好，尤其对混合比例、位置参数的估计都较混合高斯分布模型估计的效果好。而对于尺度参数的估计，混合高斯模型拟合下得到的参数的均方误差略小，但相差不大。因此相对于混合高斯分布，混合 t-分布模型可以更好的拟合含噪声的混合高斯数据，这也正说明了混合 t-分布模型较于混合高斯模型能够更好地处理重尾数据。

Table 7. Mixed Gaussian data with 5% noises
表 7. 含噪声 5%的混合高斯数据

参数	EST1	EST2	MSE1	MSE2
$\pi_1=0.5$	0.4911	0.4944	0.0001	0.0001
$\pi_2=0.3$	0.2988	0.3001	0.0001	0.0001
$\mu_1=2$	1.9774	1.9960	0.0031	0.0026
$\sigma_1=1$	1.0187	1.0426	0.0019	0.0038
$\mu_2=7$	6.9379	7.0000	0.0100	0.0057
$\sigma_2=1$	1.0246	1.0412	0.0046	0.0075
$\mu_3=11$	10.9821	11.0173	0.0118	0.0101
$\sigma_3=1$	1.0699	1.0543	0.0098	0.0110

Table 8. Mixed Gaussian data with 10% noises
表 8. 含噪声 10%的混合高斯数据

参数	EST1	EST2	MSE1	MSE2
$\pi_1=0.5$	0.4841	0.4885	0.0003	0.0002
$\pi_2=0.3$	0.3005	0.2994	0.0001	0.0001
$\mu_1=2$	1.9663	1.9914	0.0044	0.0030
$\sigma_1=1$	1.0483	1.0776	0.0042	0.0086
$\mu_2=7$	6.9231	6.9857	0.0122	0.0062
$\sigma_2=1$	1.0624	1.0674	0.0086	0.0111
$\mu_3=11$	11.0148	11.0341	0.0144	0.0120
$\sigma_3=1$	1.1078	1.1083	0.0172	0.0226

5. 结论

本文主要研究了一元混合 t-分布模型，给出了 EM 算法下该模型参数的极大似然估计，并在模拟的三种类型的数据下与混合高斯模型进行了对比分析。从前两类数据的模型参数估计结果中可以看出，每个子分布的自由度固定且取相同的值的情况下，对于混合高斯数据，当自由度的取值足够大时，基于混合 t-分布模型的 EM 算法的参数估计结果并不比基于混合高斯模型的 EM 算法差；对于混合 t-分布数据，基于混合 t-分布模型的 EM 算法能够得到较好的估计结果并优于基于混合高斯模型的 EM 算法的估计结果，且随着自由度的增大，效果会更好；而在第三类含噪声的混合高斯分布数据下，整体而言，混合 t-分布模型比混合高斯分布模型拟合效果更好，说明了混合 t-分布模型在处理重尾数据上更具优势。以上结果验证了本文研究的模型和方法的有效性。

资助信息

国家自然科学基金(11601150)；国家自然科学基金(U1430103)；中央高校基本科研业务费专项资金资助(2016MS62)。

参考文献 (References)

[1] Dempster, P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1-38.

[2] McLachlan, G. and Krishnan, T. (2007) The EM Algorithm and Extensions. John Wiley & Sons, New York.

[3] Meng, X.L. and Rubin, D.B. (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80**, 267-278. https://doi.org/10.1093/biomet/80.2.267

[4] Peel, D. and McLachlan, G. (2000) Robust Mixture Modelling Using the t Distribution. *Statistics and Computing*, **10**, 339-348. https://doi.org/10.1023/A:1008981510081

[5] Liu, C. and Rubin, D.B. (1995) ML Estimation of the t Distribution Using EM and Its Extensions, ECM and ECME. *Statistica Sinica*, **5**, 19-39.

[6] 冉延平. 基于混合模型的聚类算法及其稳健性研究[D]: [硕士学位论文]. 郑州: 中国人民解放军信息工程大学, 2005.

[7] 史鹏飞. 基于改进 EM 算法的混合模型参数估计及聚类分析[D]: [硕士学位论文]. 西安: 西北大学, 2009.

[8] 杨云飞. 基于混合模型的医学图像分割算法应用研究[D]: [硕士学位论文]. 南京: 东南大学, 2015.

[9] 熊太松. 基于统计混合模型的图像分割方法研究[D]: [博士学位论文]. 成都: 电子科技大学, 2013.

[10] 朱志娥, 吴刘仓, 戴琳. 偏 t 正态数据下混合线性联合位置与尺度模型的参数估计[J]. 高校应用数学学报, 2016, 31(4): 379-389.

[11] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin, 423-435.

[12] Shoham, S. (2002) Robust Clustering by Deterministic Agglomeration EM of Mixtures of Multivariate T-Distributions. *Pattern Recognition*, **35**, 1127-1142. https://doi.org/10.1016/S0031-3203(01)00080-2

[13] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012(3).

[14] Coretto, P. and Hennig, C. (2010) A Simulation Study to Compare Robust Clustering Methods Based on Mixtures. *Advances in Data Analysis and Classification*, **4**, 111-135. https://doi.org/10.1007/s11634-010-0065-4

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org