DOI:10.13203/j.whugis20200172 文章编号:1671-8860(2022)09-1432-07

EM算法在p范混合模型参数估计中的应用

彭 飞[1] 王 中[1] 孟庆旭[1] 潘 雄[2] 邱封钦[3] 杨玉锋[3]

1 海军工程大学舰船与海洋学院,湖北 武汉,430033
2 武汉纺织大学计算机与人工智能学院,湖北 武汉,430200
3 中国地质大学(武汉)地理与信息工程学院,湖北 武汉,430078

摘 要:针对多种分布形式混合的观测数据,建立了p范混合模型,考虑到模型中混合数属于不完全数据,引入期望最大化(expectation-maximum, EM)算法,对该混合模型的参数进行估计,详细推导了p范混合模型参数估计的迭代公式,并给出了相应的迭代步骤。采用混合高斯分布数据、拉普拉斯分布与高斯分布混合数据及实测GPS观测值残差数据,验证了公式的正确性和适应性。算例结果表明,与单一概率分布相比,p范混合模型能够准确反映数据分布的实际情况,同时利用EM算法估计的模型参数具有较高的精度。

关键词:混合模型;参数估计;p范分布;期望最大化算法

中图分类号:P207 **文献标志码**:A

高精度测量技术在航空、航天和船舶建造领域有着广泛的应用,而数据的处理是精度控制的关键技术。在传统的数据处理过程中,认为误差服从正态分布,但在数据的采集、录入及处理过程中,不可避免地会出现异质数据,即数据来源于不同的子群体,而不是同质的单一的群体,如多个正态分布的组合、正态分布和拉普拉斯分布的组合或者多个其他分布的组合等,即数据中混合了多种分布形式的误差,从而形成了混合分布模型[1-3]。混合模型已成为数据分析中最常用的模型之一,期望最大化(expectation-maximum, EM)算法[4]为求解这些混合模型提供了一个较好的思路,许多学者对EM算法做了研究[5-13]。冯杭等[12]研究了混合高斯分布、混合指数分布参数估计的EM算法,给出了相应的迭代公式;赵杨璐等[10]研究了混合模型中总体个数的确定方法。也有学者将EM算法应用到测绘数据处理中,得到了一些有用的结果[8,13]。这些学者研究的都是同类型的加权混合模型,而对不同类型的加权混合模型的研究相对较少。

在数据处理过程中,观测值(误差)并不一定服从正态分布,研究表明,部分观测值服从更接近实际误差分布的p范分布[3,14-15],多名学者利用迭代法、分数矩、对数矩等方法,给出了更加符合实际情况的参数p的求解方法,提高了估计值的效率。在单个误差的假设条件下,通过选择合适的p值,可使误差分布的理论模式较正态分布更接近于误差的真实分布,估计结果的精度更高[2,14-15]。在误差为多个的情况下,对于p范分布混合模型的研究成果较少。

本文将正态混合模型推广到p范混合模型,借助于EM算法,推导了p范混合分布情况下参数估计的迭代公式,给出了相应的迭代步骤。同时提出利用EM算法实现参数的精确解算,使得扩展后的混合模型更符合实际的情况,提高了参数估计的精确性。

1 p范混合模型

设$l_i(i=1,2\cdots n)$是μ的一组独立观测值,观测值$L=(l_1,l_2\cdots l_n)^{\mathrm{T}}$服从一元$p$范分布,概率密度函数为[1,3]:

$$f(l_i|\theta)=\frac{p\lambda}{2\sigma\Gamma(1/p)}\exp\left\{-\left[\frac{\lambda}{\sigma}\left|l_i-\mu\right|\right]^p\right\} \tag{1}$$

式中,$\lambda=\sqrt{\Gamma(3/p)/\Gamma(1/p)}$;参数分量$\theta=(p,\mu,\sigma)$,其中,$p$为尺度参数(当$p=1$时,误差的概率密度函数为拉普拉斯分布;当$p=2$时,误差的概率密度函数为正态分布;当$p$值无限趋近于0

收稿日期:2020-04-17
项目资助:国家自然科学基金(42174010,41874009)。
第一作者:彭飞,博士,副教授,主要从事舰船建造工艺、舰船总体设计研究。pengfei75@qq.com
通讯作者:王中,博士,讲师。wangzhonghj@sohu.com

时,误差服从极限分布;当p值无限趋近于正无穷大时,误差服从均匀分布)。

若各观测值服从一元p范分布,假定模型的混合数为m,则构成了一元p范混合模型,该模型的概率密度函数表示如下:

$$f(L|\theta)=\sum_{i=1}^{n_1}\alpha_1 f(l_i|\theta_1)+\sum_{i=n_1+1}^{n_2}\alpha_2 f(l_i|\theta_2)+\cdots+\sum_{i=n_{m-1}+1}^{n}\alpha_m f(l_i|\theta_m) \tag{2}$$

式中,参数分量$\theta=(\theta_1,\theta_2\cdots\theta_i\cdots\theta_m)$;$\theta_i=(p_i,\mu_i,\sigma_i)$;$\alpha_j$为混合的权重,表示满足第$j$种分布的数据所占的比例,为了满足密度函数的性质,必须满足$0\leqslant\alpha_j\leqslant1$且$\sum_{j=1}^{m}\alpha_j=1$,第$j$个总体的密度函数为$f_j(L)$。

该混合模型具有较强的灵活性,其难点在于如何求解参数的估计量。p值可以根据一定的迭代方法进行确定[1,3],也可以采用直接计算公式快速估计出p值[2,15],从而待估计的参数只有方差σ_i、均值μ_i以及混合数α_i。

2　似然函数构造

设L为混合分布的观测数据,由于无法分辨出哪个样本来自哪个分布,因此,观测数据中没有包含数据的全部信息,是不完全数据。引入分量z_{ij},当$z_{ij}=1$时,表示第i个观测数据来自第j个分布,即$P(z_{ij}=1)=\alpha_j$;当$z_{ij}=0$时,表示第i个观测数据不是来自第j个分布,则观测值L的条件分布密度函数为:

$$f_{L|Z}(L|z_{ij}=1)=f_j(L) \tag{3}$$

由于混合数α_j是无法观测的,因此称为不完全数据或缺失数据,设缺失数据向量为$Z=(z_1,z_2\cdots z_n)^{\mathrm{T}}$,$z_i=(z_{i1},z_{i2}\cdots z_{im})$,则$(L,Z)$称为完整数据。则有:

$$f(L,Z)=\sum_{j=1}^{m}f(L,z_{ij})=\sum_{j=1}^{m}P(z_{ij}=1)f_{L|Z}(L|z_{ij}=1)=\sum_{j=1}^{m}\alpha_j f_j(L)=\sum_{j=1}^{m}\alpha_j\frac{p_j\lambda_j}{2\Gamma(1/p_j)\sigma_j}\exp\{-(\frac{\lambda_j}{\sigma_j})^{p_j}|L-u_j|^{p_j}\} \tag{4}$$

完全数据(L,Z)的似然函数为:

$$f(L,Z)=\prod_{j=1}^{m}(\alpha_j f_j(L))^{z_{ij}} \tag{5}$$

缺失数据Z的条件分布为:

$$f_{L|Z}(z_{ij}=1|L)=\frac{f(L,z_{ij}=1)}{f(L)}=\frac{\alpha_j f_j(L)}{f(L)} \tag{6}$$

设$l_1,l_2\cdots l_n$为取自上述p范混合模型的一组独立观测数据,对应的缺失数据为$z_1,z_2\cdots z_m$,记$\theta=(\alpha,\mu,\sigma)$,则完全数据的对数似然函数为:

$$\ln(f(L,Z|\theta))=\sum_{i=1}^{n}\ln(f(l_i|z_i,\theta)f(z_i|\theta))=\sum_{i=1}^{n}\ln(\alpha_{z_i}f(l_i|\theta_{z_i}))=\sum_{i=1}^{n}\sum_{j=1}^{m}z_{ij}(\ln\alpha_j+\ln(\frac{p_j\lambda_j}{2\Gamma(1/p_j)\sigma_j}\exp\{-(\frac{\lambda_j}{\sigma_j})^{p_j}|l_i-u_j|^{p_j}\}))=$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m}z_{ij}(\ln\alpha_j+\ln p_j+\frac{1}{2}\ln\Gamma(3/p_j)-\frac{3}{2}\ln\Gamma(1/p_j)-\ln2-\ln\sigma_j-(\lambda_j/\sigma_j)^{p_j}|l_i-u_j|^{p_j}) \tag{7}$$

3　EM算法下的参数解算

对含有m个子体的p范混合模型来说,EM算法是迭代算法。先给定参数的初始值$\alpha_1^{(0)}$,$\alpha_2^{(0)}\cdots\alpha_m^{(0)}$;$(\mu_1^{(0)},\sigma_1^{(0)})$,$(\mu_2^{(0)},\sigma_2^{(0)})\cdots(\mu_m^{(0)},\sigma_m^{(0)})$,由它求出缺失数据的值;再根据此数据估计出新的参数估计值,根据这一估计值对缺失数据的值进行更新;如此反复迭代,直到收敛为止。

应用EM算法求解式(7)。求解第k次各参数表达式的步骤如下:

1)E步:构造$Q(\theta,\theta^{(k)})$。

缺失数据Z的条件分布期望为:

$$E(Z_{ik}|L,\theta^{(k)})=P(Z_{ik}=1|L,\theta^{(k)})=[P(Z_{ik}=1|L_i,\theta_i^{(k)})]/[\sum_{j=1}^{m}P(Z_{ik}=1|L_i,\theta_j^{(k)})]=$$
$$[P(L_i|Z_{ik}=1,\theta_i^{(k)})P(Z_{ik}=1,\theta_i^{(k)})]/[\sum_{j=1}^{m}P(L_i|Z_{ik}=1,\theta_j^{(k)})P(Z_{ik}=1,\theta_j^{(k)})]=$$
$$[\alpha_i^{(k)}f_i(l_i|\theta_i^{(k)})]/[\sum_{j=1}^{m}\alpha_j^{(k)}f_j(l_i|\theta_j^{(k)})] \tag{8}$$

由式(7)、(8)得:

$$Q(\theta,\theta^{(k)})=E(\ln f(L,Z|L,\theta^{(k)}))=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln(\alpha_j^{(k)})W_{ij}^{(k)}+\sum_{i=1}^{n}\sum_{j=1}^{m}\ln f(l_i,z_{ij}|\theta_j^{(k)}))W_{ij}^{(k)} \tag{9}$$

式中，$W_{ij}^{(k)}=(\alpha_j^{(k)}f_j^{(k)}(l_i))/(f^{(k)}(l_i))$。

2)M步：将对数似然函数$Q(\theta,\theta^i)$极大化，求取相应参数的参数估计值。在$\sum_{j=1}^{m}\alpha_j=1$的限制条件下，由式(7)求$Q(\theta,\theta^{(k)})$关于α_j、μ_j和σ_j的最大值，对似然函数各参数求导，令其等于0，有：

$$\frac{\partial}{\partial\alpha_j}\Big(\sum_{i=1}^{n}\sum_{j=1}^{m}\ln(\alpha_j^{(k)})W_{ij}^{(k)}+\sum_{i=1}^{n}\sum_{j=1}^{m}\ln f(l_i,z_{ij}|\theta_j^{(k)}))W_{ij}^{(k)}-\lambda\Big(\sum_{j=1}^{m}\alpha_j-1\Big)=0 \quad (10)$$

$$\sum_{i=1}^{n}W_{ij}=\lambda\hat{\alpha}_j \quad (11)$$

由于$\sum_{j=1}^{m}\alpha_j=1$，可以得到$\lambda=\sum_{j=1}^{m}\sum_{i=1}^{n}W_{ij}=n$，故得到$\alpha_j$的估计值，从而得到该参数的第$k+1$次迭代表达式为：

$$\hat{\alpha}_j^{(k+1)}=\frac{1}{n}\sum_{i=1}^{n}W_{ij}^{(k)}=\frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_j^{(k)}f_j^{(k)}(l_i|\theta_j^{(k)})}{f^{(k)}(l_i|\theta_j^{(k)})} \quad (12)$$

分别对μ_j、σ_j求偏导，化简，同理可得到参数的第$k+1$次迭代更新式如下：

$$\sum_{i=1}^{n}W_{ij}^{(k)}|l_i-\hat{\mu}_j^{(k+1)}|^{p_j-2}(l_i-\hat{\mu}_j^{(k+1)})=0 \quad (13)$$

$$\hat{\sigma}_j^{p_j(k+1)}=\Big[\,p_j\lambda_j^{p_j}\sum_{i=1}^{n}W_{ij}^{(k)}|l_i-\hat{\mu}_j^{(k)}|^{p_j}\,\Big]/\sum_{i=1}^{n}W_{ij}^{(k)}=\frac{p_j\lambda_j^{p_j}}{n\hat{\alpha}_j^{(k)}}\sum_{i=1}^{n}W_{ij}^{(k)}|l_i-\hat{\mu}_j^{(k)}|^{p_j} \quad (14)$$

式(12)～(14)是参数的非线性方程，可以采用迭代的方法计算，计算步骤可总结为：(1)选择合适的初始值，令$\theta_j^{(k)}=(\alpha_j^{(k)},\mu_j^{(k)},\sigma_j^{(k)})$；(2)进行第$k+1$次迭代，求得新的混合系数$\alpha_j^{(k+1)}$；(3)计算均值的估计值，通过迭代解方程(13)，求得第$k+1$次迭代的均值$\mu_j^{(k+1)}$；(4)计算方差的估计值，将第(2)步、第(3)步中得到的混合系数$\alpha_j^{(k+1)}$和$\mu_j^{(k+1)}$代入式(14)，求得第$k+1$次迭代的方差值$\sigma_j^{(k+1)}$；(5)比较迭代后得到的各参数估计值与迭代前相应参数估计值的差值是否充分小，若不满足，则将此次迭代值作为下一次迭代的初始值进行迭代运算，直到差值充分小停止循环。

4 算例与分析

假定模型的混合系数为2，采用拉普拉斯分布子样、高斯分布子样以及实测GPS观测值残差数据作为实验数据。首先，利用矩估计求解出混合数据的p值[16]；然后，利用本文的EM算法解算混合p范分布的参数值；最后，通过分析实验的结果验证本文算法的可行性。

4.1 算例1：混合同分布数据

假定p范混合模型为高斯混合模型，即混合数据由两组高斯分布数据组成。通过Matlab软件随机生成服从$L_1\sim N(3,1)$，$L_2\sim N(4,3^2)$的数据，L_1中取600个数据作为样本1，L_2中取400个数据作为样本2，即样本总数$n=1\,000$进行实验。样本1和样本2的数据分布直方图如图1所示(横坐标和纵坐标均无单位)。

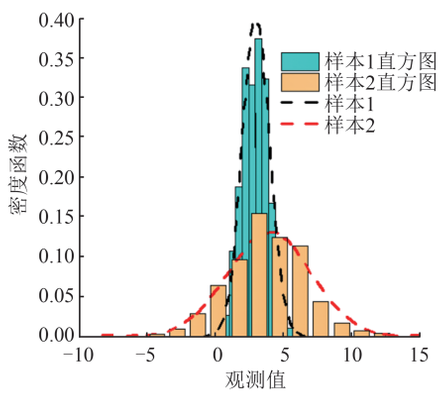

图1 算例1的样本直方图及分布曲线

Fig.1 Sample Histogram and Distribution Curves of Example 1

利用矩估计迭代法求解出混合数据的模型分布参数p值，并以直方图的形式给出模型参数p值求解的可靠性，如图2所示。

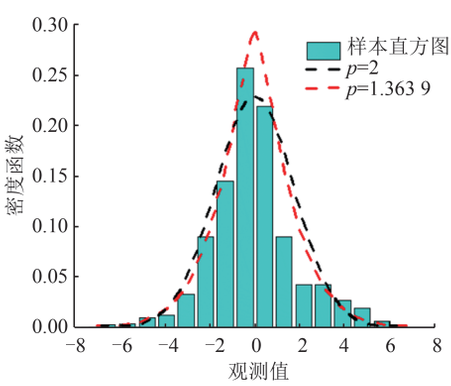

图2 直方图及真实和估计的分布曲线(算例1)

Fig.2 Histogram and True and Estimated Distribution Curves (Example 1)

从图2可以看出，两种高斯数据混合后，计算出的p值为1.363 9，可见混合高斯分布的数据不再服从高斯分布。与高斯分布相比，采用p范分布($p=1.363\,9$)估计出的概率密度与真实密度更为一致，也更加符合实际分布的情况。

为了获得更高的模型参数精度，利用上述样本数据进行12次实验(p选取1.363 9)，将本文算法EM_p估计的结果列入表1，取12次结果的平均值作为模型参数的最终值。为了说明混合模

型参数估计的准确性,将计算结果与EM解算高斯混合模型参数(EM_G)结果进行对比,并采用均方根误差(root mean square error,RMSE)来表征参数估计的效果。

表 1　高斯分布混合下的EM算法估计结果

Tab.1　Estimation Results of EM Algorithm Under Gaussian Distribution Mixing

次数	$\mu_1=3$		$\mu_2=4$		$\sigma_1=1$		$\sigma_2=3$		$\alpha_1=0.6$		$\alpha_2=0.4$	
	EM_p	EM_G	EM_p	EM_G	EEM_p	EM_G	EM_p	EM_G	EM_p	EM_G	EM_p	EM_G
1	3.019 1	2.904 8	3.946 3	4.072 0	0.992 8	1.171 2	2.903 1	2.708 0	0.601 7	0.585 7	0.398 3	0.414 3
2	3.019 1	2.881 8	3.946 8	4.099 2	0.993 1	1.126 7	2.903 5	2.717 4	0.601 9	0.583 9	0.398 1	0.416 1
3	3.019 1	2.904 7	3.946 0	4.070 0	0.992 7	1.170 2	2.902 9	2.707 2	0.601 6	0.584 9	0.398 4	0.415 1
4	3.019 1	2.880 4	3.945 8	4.089 7	0.992 6	1.122 2	2.902 8	2.712 8	0.601 5	0.580 0	0.398 5	0.420 0
5	3.019 0	2.878 7	3.946 4	4.079 2	0.992 9	1.117 0	2.903 0	2.707 5	0.601 7	0.575 5	0.398 3	0.424 5
6	3.019 1	2.879 4	3.945 8	4.083 2	0.992 6	1.119 1	2.902 8	2.709 5	0.601 5	0.577 2	0.398 5	0.422 8
7	3.019 0	2.885 7	3.946 4	4.125 3	0.992 9	1.138 9	2.903 0	2.729 8	0.601 7	0.594 5	0.398 3	0.405 5
8	3.019 1	2.888 4	3.946 0	4.143 6	0.992 7	1.147 3	2.902 9	2.738 3	0.601 6	0.601 7	0.398 4	0.398 3
9	3.019 0	2.886 3	3.946 4	4.129 4	0.992 9	1.140 7	2.903 1	2.731 7	0.601 7	0.596 1	0.398 3	0.403 9
10	3.019 1	2.883 4	3.946 6	4.110 0	0.993 0	1.131 7	2.903 4	2.722 5	0.601 8	0.588 3	0.398 2	0.411 7
11	3.019 0	2.885 3	3.946 7	4.122 6	0.993 0	1.137 5	2.903 3	2.728 5	0.601 8	0.593 4	0.398 2	0.406 6
12	3.019 1	2.887 7	3.946 5	4.138 2	0.993 0	1.144 9	2.903 3	2.735 7	0.601 8	0.599 6	0.398 2	0.400 4
均值	3.019 1	2.887 2	3.945 8	4.105 2	0.992 8	1.139 0	2.903 1	2.720 7	0.601 7	0.588 4	0.398 3	0.411 6

通过表1可以发现,不论是EM_p还是EM_G算法,均能较好地估计出混合模型的6个参数。但从最终估计的结果来看,EM_p估计出的6个模型参数十分接近真值,估计精度远远高于EM_G算法。同时,每次估计出的模型参数变化均较小,且与真值符合度较高,从而验证了本文算法估计混合多峰数据的有效性和稳定性。结合图3可以看出,EM_p算法估计的模型参数的RMSE均小于0.05,远远优于EM_G算法,进一步说明利用EM_p算法估计的混合高斯分布模型参数具有较好的精确度和稳定性。

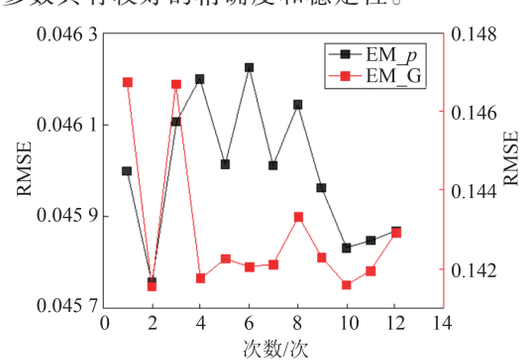

图3　模型参数估计的均方根误差(算例1)

Fig.3　Root Mean Square Error of Model Parameter Estimates(Example 1)

4.2　算例2:混合异分布数据

假定混合模型中p值分别取1和2,混合数据由一组拉普拉斯分布数据($L_1\sim L(u=0,\sigma=1)$)和一组高斯分布数据$L_2\sim N(1,4^2)$组成。分别从L_1和L_2中各取1 000个数据作为样本数据,即样本总数$n=2\,000$进行实验,样本直方图见图4。

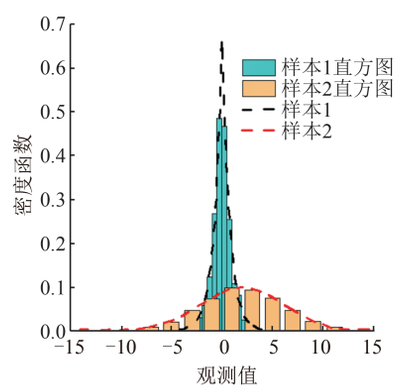

图4　算例2的样本直方图及分布曲线

Fig.4　Sample Histogram and Distribution Curves of Example 2

利用矩估计法求解出混合数据的模型参数p值,由图5可以看出,当p取0.952 6时,估计出的概率密度与真实密度十分接近,说明此时的模型更加符合该实验样本数据分布的真实情况。

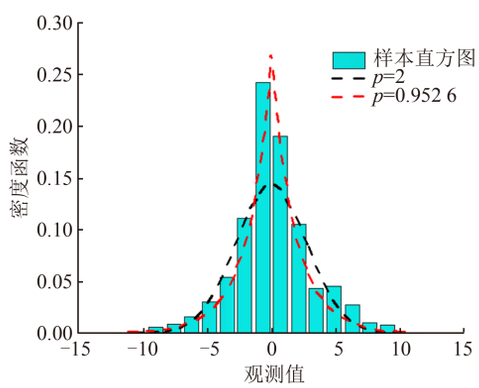

图5　直方图及真实和估计的分布曲线(算例2)

Fig.5　Histogram and True and Estimated Distribution Curves (Example 2)

表2统计了混合数据的模型参数，取12次计算结果的平均值作为模型参数的最终值。

表2 拉普拉斯分布与高斯分布混合下的EM算法估计结果

Tab.2 Estimation Results of EM Algorithm Under the Mixture of Laplace and Gaussian Distributions

次数	$\mu_1=0$		$\mu_2=1$		$\sigma_1=1$		$\sigma_2=4$		$\alpha_1=0.5$		$\alpha_2=0.5$	
	EM_p	EM_G	EM_p	EM_G	EM_p	EM_G	EM_p	EM_G	EM_p	EM_G	EM_p	EM_G
1	0.003 8	0.118 4	0.929 9	0.883 8	0.889 5	0.644 7	4.054 1	3.744 0	0.448 8	0.406 1	0.551 2	0.593 9
2	0.003 8	0.118 5	0.929 8	0.883 9	0.889 4	0.645 2	4.053 9	3.744 5	0.448 8	0.406 3	0.551 2	0.593 7
3	0.003 8	0.118 3	0.929 7	0.883 7	0.889 3	0.644 7	4.053 8	3.743 9	0.448 7	0.406 1	0.551 3	0.593 9
4	0.003 9	0.118 4	0.930 2	0.883 8	0.889 8	0.644 9	4.054 7	3.744 2	0.449 0	0.406 2	0.551 0	0.593 8
5	0.003 9	0.118 3	0.930 0	0.883 7	0.889 6	0.644 6	4.054 3	3.743 8	0.448 9	0.406 0	0.551 1	0.594 0
6	0.003 8	0.118 3	0.929 9	0.883 7	0.889 5	0.644 6	4.054 1	3.743 8	0.448 8	0.406 0	0.551 2	0.594 0
7	0.003 9	0.118 5	0.930 1	0.883 9	0.889 8	0.645 1	4.054 6	3.744 4	0.449 0	0.406 2	0.551 0	0.593 8
8	0.003 9	0.118 4	0.930 0	0.883 8	0.889 7	0.644 9	4.054 4	3.744 1	0.448 9	0.406 2	0.551 1	0.593 8
9	0.003 9	0.118 5	0.930 2	0.884 0	0.889 9	0.645 4	4.054 8	3.744 7	0.449 0	0.406 4	0.551 0	0.593 6
10	0.003 9	0.118 5	0.930 2	0.883 9	0.889 9	0.645 3	4.054 8	3.744 6	0.449 0	0.406 3	0.551 0	0.593 7
11	0.003 8	0.118 4	0.929 7	0.883 8	0.889 3	0.645 0	4.053 7	3.744 3	0.448 7	0.406 2	0.551 3	0.593 8
12	0.003 8	0.118 3	0.929 7	0.883 7	0.889 3	0.644 6	4.053 8	3.743 7	0.448 7	0.406 0	0.551 3	0.594 0
均值	0.003 9	0.118 4	0.930 0	0.883 8	0.889 6	0.644 9	4.054 2	3.744 2	0.448 9	0.406 2	0.551 1	0.593 8

通过表2可以发现，当样本数据是由拉普拉斯分布与高斯分布混合组成时，EM_G算法估计出的模型参数精度较差，不能将混合数据分类出来，其原因是样本数据中存在不同于高斯分布的数据，致使其算法失效。同时可以看出EM_p算法虽然模型参数估计的精度不如混合同分布数据时那么精确，但从最终估计的结果来看，EM_p算法仍能够较好地估计出混合模型的6个参数，估计出的模型参数十分接近真值，估计精度也远远高于EM_G算法。结合图6可以看出，EM_G算法估计的模型参数的RMSE在4左右，估计的精度达不到模型参数估计所需的精度，而EM_p算法估计的模型参数的均方根误差在0.06附近，精度较高，充分说明EM_p算法能够有效地估计出混合异分布模型的参数。

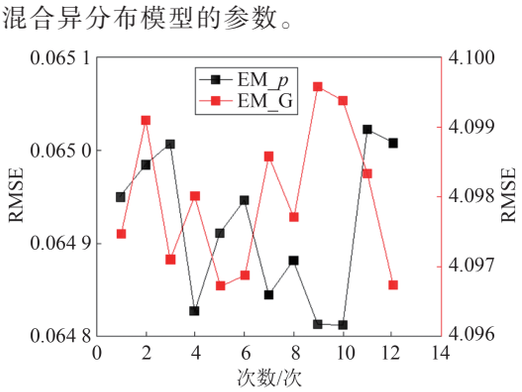

图6 模型参数估计的均方根误差(算例2)

Fig.6 Root Mean Square Error of Model Parameter Estimates (Example 2)

4.3 算例3：实测GPS观测值残差数据

数据来自加拿大Algonquin Park的ALGO测站点，利用TPS NET-G3A接收机采集获得2013-04-28的观测数据。在获得的32颗卫星对地观测数据中，选取某颗卫星伪距的精密单点定位双频无电离层组合观测值残差进行分析。取其中200个误差值作为样本数据，利用矩估计求出样本数据的p值为1.398。假设样本数据由两种分布数据组成，利用EM_p算法进行参数解算，结果如表3所示。

表3 观测值残差的EM算法估计结果

Tab.3 Estimation Results of EM Algorithm of Observed Value Residuals

次数	μ_1	μ_2	σ_1	σ_2	α_1	α_2
1	−0.066	−0.028	0.543	0.425	0.488	0.511
2	−0.075	−0.024	0.556	0.436	0.433	0.566
3	−0.065	−0.029	0.543	0.428	0.485	0.515
4	−0.071	−0.026	0.549	0.431	0.464	0.536
5	−0.075	−0.025	0.556	0.437	0.434	0.566
6	−0.073	−0.025	0.552	0.434	0.449	0.551
7	−0.067	−0.028	0.545	0.430	0.478	0.523
8	−0.070	−0.026	0.549	0.431	0.467	0.534
9	−0.076	−0.025	0.557	0.437	0.433	0.567
10	−0.068	−0.027	0.546	0.427	0.480	0.520
11	−0.072	−0.025	0.552	0.433	0.452	0.548
12	−0.071	−0.026	0.551	0.432	0.458	0.542
均值	−0.071	−0.026	0.550	0.432	0.460	0.540

由表3可以看出，EM_p算法计算出的两类分布数据十分相似，几乎为同一种分布。所以将其模型分量数设为1，重新利用EM_p算法进行解算，取12个计算结果的平均值作为真值。最终求得GPS观测值残差服从p为1.398、μ为−0.052、σ为0.446的p范分布。

通过GPS伪距单点定位的精度来验证计算

结果的正确性。利用武汉大学精密单点定位软件对观测数据进行解算,得到观测站点的精密坐标值,并以此作为伪距单点定位中测站点的真实坐标。现以$p_{0371180}$点的观测数据为例,对其进行解算,通过精密单点定位软件解算得到的测站点高精度坐标为:$[X,Y,Z]=[-1\,304\,152.045\,9,-4\,831\,831.378\,5,3\,943\,232.966\,1]$。分别假设误差服从高斯分布和$p$范分布($p=1.398$),利用最小二乘(least square,LS)和p范平差(least p-norm adjustment,Lp)求解伪距单点定位误差方程,得到待求参数的估计值。计算出定位点的三维坐标估计值,采用各历元解得的坐标估计值与真实坐标值进行对比,以每个方向上的坐标中误差作为GPS伪距单点定位的精度,见表4。

表4　伪距单点定位结果/m

Tab.4　Pseudorange Single Point Positioning Results/m

坐标	LS		Lp	
	估值	中误差	估值	中误差
X	−1 304 154.105	2.582	−1 304 153.996	2.472
Y	−4 831 834.409	4.474	−4 831 833.913	4.154
Z	3 943 236.838	5.344	3 943 236.658	5.212
精度	1.138		0.920	

从表4可以看出,假定GPS观测值的误差服从高斯分布,利用传统的LS求解伪距单点定位的精度较低,其定位误差(1.138 m)达到米级。采用本文EM算法求解出的GPS观测值的误差分布模型进行p范平差所得到的坐标在3个方向上的精度均优于LS,定位效果达到分米级,从而进一步验证了本文EM_p算法估计模型参数具有较高的准确度与可靠性。

5　结　语

p范混合模型作为一种新的分布模型,考虑了测量误差的不确定性和误差分布的多样性。本文探讨了该模型参数的EM算法的迭代公式,利用模拟数据验证了EM算法结合p范分布可以有效解决误差的参数估计问题,并将其应用到GPS观测值误差分析中。从仿真和实测数据的算例计算结果可以看出,相比高斯混合模型,p范混合模型能够更好地反映出混合数据的实际分布情况,同时利用EM算法求解出的混合p范模型的参数值也更加准确。

本文扩充了p范分布理论,对进一步提高测量数据处理的精度具有一定的实用价值。同时还发现混合数的确定、尺度参数的确定和参数的初始值对计算结果影响较大,如何减小其影响是下一步需要考虑的问题。

参　考　文　献

[1] Pan Xiong. The Estimation Theory and Application Research in Semi-Parametric Model[D]. Wuhan: Wuhan University, 2005(潘雄. 半参数模型的估计理论及其应用[D]. 武汉: 武汉大学, 2005)

[2] Pan Xiong, Cheng Shaojie, Zhao Chunru. A Fast Parameter Estimation in p-Norm Distribution[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(2): 189-192(潘雄, 程少杰, 赵春茹. 一元p范分布的参数快速估计方法[J]. 武汉大学学报·信息科学版, 2010, 35(2): 189-192)

[3] Sun Haiyan. p-Distribution Theory and Its Application in Modern Survey Data Processing[D]. Wuhan: Wuhan University of Surveying and Mapping, 1995(孙海燕.p范分布理论及其在现代测量数据处理中的应用[D]. 武汉:武汉测绘科技大学,1995)

[4] Booth J G, Hobert J P. Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm[J]. *Journal of the Royal Statistical Society*, 1999, 61(1): 265-285

[5] Lian Junyan. The Application Research of EM Algorithm and Its Improvement in Mixed Model Parameter Estimation[D]. Xi'an: Chang'an University, 2006(连军艳. EM算法及其改进在混合模型参数估计中的应用研究[D]. 西安: 长安大学, 2006)

[6] Tuaç Y, Güney Y, Arslan O. Parameter Estimation of Regression Model with AR(p) Error Terms Based on Skew Distributions with EM Algorithm[J]. *Soft Computing*, 2020, 24(5): 3309-3330

[7] Wu Ke, He Tan, Yang Yetao. Change Detection Method Based on Pixel Unmixing and EM Algorithm for Low and Medium Resolution Remote Sensing Imagery[J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(4): 555-562(吴柯, 何坦, 杨叶涛. 基于混合像元分解与EM算法的中低分辨率遥感影像变化检测[J]. 武汉大学学报·信息科学版, 2019, 44(4): 555-562)

[8] Xiao Qinqin, Song Yingchun, Du Kun. Application of EM Algorithm to the Calculation of the Satellite Position Based on Broadcast Ephemeris[J]. *Engineering of Surveying and Mapping*, 2013, 22(6): 73-76(肖琴琴, 宋迎春, 杜琨. EM算法在广播星历计算卫星位置中的应用[J]. 测绘工程, 2013, 22(6): 73-76)

[9] Lu Nana, Yu Jinghu. Research on Resolution Based on EM Algorithm[J]. *Acta Mathematica Scientia*, 2019, 39(3): 638-648(鲁纳纳, 余旌胡. EM算法的参数分辨率[J]. 数学物理学报, 2019, 39(3): 638-648)

[10] Zhao Yanglu, Duan Dandan, Hu Raomin, et al. On the Number of Components in Mixture Model Based on EM Algorithm[J]. *Journal of Applied Statistics and Management*, 2020, 39(1): 35-50 (赵杨璐, 段丹丹, 胡饶敏, 等. 基于EM算法的混合模型中子总体个数的研究[J]. 数理统计与管理, 2020, 39(1): 35-50)

[11] Li Renzhong, Zhang Huanhuan, Jing Junfeng, et al. Fabric Defect Detection Based on Gaussian Mixture Models of EM Algorithm [J]. *Computer Engineering and Applications*, 2014, 50(10): 184-187 (李仁忠, 张缓缓, 景军锋, 等. 基于EM算法的高斯混合型的织物疵点检测研究[J]. 计算机工程与应用, 2014, 50(10): 184-187)

[12] Feng Hang, Wang Shengbing. Discrete-Continuous Mixed Distribution Parameter Estimation Based on EM Algorithm [J]. *Statistics & Decision*, 2019, 35(3): 85-88 (冯杭, 王胜兵. 基于EM算法的离散-连续型混合分布参数估计[J]. 统计与决策, 2019, 35(3): 85-88)

[13] Guo X, Li Q Y, Xu W L. Acceleration of the EM Algorithm Using the Vector Aitken Method and Its Steffensen Form[J]. *Acta Mathematicae Applicatae Sinica*, English Series, 2017, 33(1): 175-182

[14] Pan Xiong, Zhao Qilong, Wang Junlei, et al. Maximum Likelihood Adjustment of the Monadic Unsymmetrical P-Norm Distribution[J]. *Acta Geodaetica et Cartographica Sinica*, 2011, 40(1): 33-36 (潘雄, 赵启龙, 王俊雷, 等. 一元非对称p范分布的极大似然平差[J]. 测绘学报, 2011, 40(1): 33-36)

[15] Pan Xiong, Luo Jing, Wang Yao. Real Order and Logarithmic Moment Estimation Method of p-Norm Distribution[J]. *Acta Geodaetica et Cartographica Sinica*, 2016, 45(3): 302-309 (潘雄, 罗静, 汪耀. p范分布的实数阶与对数矩估计法[J]. 测绘学报, 2016, 45(3): 302-309)

Application of EM Algorithm in Parameter Estimation of p-Norm Mixture Model

PENG Fei[1] *WANG Zhong*[1] *MENG Qingxu*[1] *PAN Xiong*[2] *QIU Fengqin*[3] *YANG Yufeng*[3]

1 College of Naval Architecture and Ocean Engineering, Naval University of Engineering, Wuhan 430033, China

2 School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China

3 School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan 430078, China

Abstract: **Objectives**: Aiming at the mixed observation data of multiple distribution forms, a expectation-maximum (EM) combined p-norm distributed model(EM_p) is established. **Methods**: Considering that the mixed number in the mixture model belongs to incomplete data, the EM algorithm is introduced to estimate the parameters of the mixture model and the p-model mixture model parameters are derived in detail. The estimated iteration formula and the corresponding iteration steps are given. The mixture Gaussian distribution data, Laplace distribution and Gaussian distribution mixture data, and the residual data of measured global positioning system(GPS) observations are used to verify the correctness and adaptability of the formula in this paper. **Results and Conclusions**: The results of the calculation examples show that, compared with the single probability distribution, the p-norm mixture model can accurately reflect the actual situation of the data distribution, and the model parameters estimated by the EM algorithm have higher accuracy.

Key words: mixture model; parameter estimation; p-norm distribution; expectation-maximum(EM) algorithm

First author: PENG Fei, PhD, associate professor, specializes in ship building technology and ship overall design research. E-mail: pengfei75@qq.com

Corresponding author: WANG Zhong, PhD, lecturer. E-mail: wangzhonghj@sohu.com

Foundation support: The National Natural Science Foundation of China(42174010, 41874009).

引文格式: PENG Fei, WANG Zhong, MENG Qingxu, et al. Application of EM Algorithm in Parameter Estimation of p-Norm Mixture Model[J]. Geomatics and Information Science of Wuhan University, 2022, 47(9): 1432-1438. DOI: 10.13203/j.whugis20200172 (彭飞, 王中, 孟庆旭, 等. EM算法在p范混合模型参数估计中的应用[J]. 武汉大学学报·信息科学版, 2022, 47(9): 1432-1438. DOI: 10.13203/j.whugis20200172)