



2019,39A(3):638–648

*Acta  
Mathematica  
Scientia*  
数学物理学报

<http://actams.wipm.ac.cn>

## EM 算法的参数分辨率 \*

鲁纳纳 余旌胡 \*\*

(武汉理工大学理学院 武汉 430070)

**摘要:** 参数分辨率是在给定噪声情况下, 衡量两个相近信号能否区分开的一个标准, 为敏感参数、有效精度以及准确度的衡量提供了评估的“尺子”. 该文以 EM 算法为基础, 结合 Fisher 线性判别准则的思想, 给出 EM 算法参数分辨率的定义, 并以两正态混合模型为例进行验证. 实验表明两个方差为 0.1 的正态分布其均值距离大于 0.206 时, EM 算法在 90% 的置信度下可以区分这两个分布, 通过构建实验结果和理论推导之间的联系, 得到不同置信度下的比例因子图. 参数分辨率的提出, 为准确度的衡量提供一个定量指标, 也为相近信号的区分提供新的解决方案.

**关键词:** EM 算法; 分辨率; 判别准则; 变异系数.

**MR(2010) 主题分类:** 62P99; 94A12   **中图分类号:** O213   **文献标识码:** A

**文章编号:** 1003-3998(2019)03-638-11

### 1 引言

同一组数据, 采用不同的模型或者同一模型选择不同的算法进行参数估计, 得到的估计值可能不一致, 因此算法评估问题在数据分析中备受关注, 评估指标也变得越发重要, 常用指标有简单性、鲁棒性、计算速度以及准确度. 简单性亦称可行性, 通俗地讲, 直观表达各属性在预测中的重要性; 鲁棒性是控制系统在一定的参数扰动下, 维持某些性能的特性; 计算速度代表一个算法的运算量; 准确度指估计值与真实值之间的接近程度. 算法评估的首要指标是准确度, 而衡量“准”的标准至关重要. 残差在模型验证和回归分析中起着关键作用 [2], Shi 等<sup>[11]</sup> 提出以残差信息准则 (RIC) 作为选择标准, 此标准取决于残差方差的估计, 如果噪声很大, 有效尾分布没有有界二阶矩, 此时无法估计方差残差, Tan 等<sup>[12]</sup> 采用基于残差最大信息系数的方法, 可以有效的拟合数据.

分辨率是衡量准确度的有效指标, 并且在很多地方具有不可忽视的作用, 例如, 敏感参数的测量: 假如已知非线性高斯的相对误差是 0.01, 通过拟合得到的值是 0.001, 此时模型选择不准, 或者拟合值 1.01 与 0.99 是否准? 有效精度的度量: 估计到小数点后 5 位与后 10 位是否有区别? 准确度的衡量: 如果能够得知算法的分辨率, 则可以对算法的准确度给出有效

---

收稿日期: 2018-07-16; 修订日期: 2018-10-13

E-mail: 1358242069@qq.com; yujh67@126.com

\* 基金项目: 中央高校基本科研业务费专项资金 (2017IVA073) 和中央高校基本科研业务费 (2018IB016)

Supported by the Fundamental Research Funds for the Central Universities (2017IVA073) and the Fundamental Research Funds for the Central Universities (2018IB016)

\*\* 通讯作者

评价. 分辨率的概念很早已被提出, 如时间分辨率是时态特征划分的最小单元, 将连续过程离散化的最小时时间间隔, min、s、ms 等可以定义时间. 空间分辨率是图像中能够辨别临界距离的最小极限, m、um、nm 等可以定义空间. 此外, 分辨率在遥感以及 CT 处理等方面得到广泛应用. 参数分辨率的引入, 为衡量算法准确度提供了有效手段. 由于不同参数或者同一参数其分辨率可能不同, 因此需要研究局部分辨率和整体分辨率.

EM(Expectation-maximization) 算法又称期望最大化算法, 1977 年由 Dempster 等<sup>[4]</sup> 提出用于求解含有隐变量数学模型的参数极大似然估计, 其基本思想: 给定不完全数据初始值情形下, 估计出模型参数值; 然后再根据参数值估计出缺失数据的值, 根据估计出的缺失数据的值再对参数值进行更新, 反复迭代直至收敛. 自 Dempster 等人提出 EM 算法以来, 国内外取得了很多研究成果, 如 Wu 和 Xu<sup>[13,15]</sup> 总结出 EM 算法的一些收敛性质. Wei 和 Liu<sup>[7,14]</sup> 提出了改进 E 步、M 步. 后来许多学者引入加速算法<sup>[1,5-6,16-17]</sup> 研究高维 EM 算法的迭代性能. 正态分布因具有灵活、高效拟合的特点, 很多随机现象在足够大样本下都可以用此分布来逼近<sup>[10]</sup>. EM 算法易受初始值影响, 不能保证找到全局最优, 往往容易陷入局部最优, 因此关于初始值设置提出了许多方法, Zhai<sup>[18]</sup> 选取最远点为初始值, 易使边界点收敛效果受到影响; Li 和 Chen<sup>[9]</sup> 使用 k- 最近邻法删除异常值, 然后用 k-means 来初始化, 此估计效果显著优于原始效果. 近几年有学者采用基于密度的方法选择初始值, 这样可以避免噪声点对参数估计的影响, 从而提高迭代结果的稳健性.

从 EM 算法的相关介绍知, 其改进集中于提高收敛速度和稳健性, 关于参数分辨率问题, 目前没有相关文献涉及. 本文以两分量混合模型参数估计的 EM 算法为例, 给出 EM 算法的参数分辨率的定义 (详见 2.3 节). 我们发现, 两个方差  $\sigma = 0.1$  的正态分布的均值距离大于 0.206 时, EM 算法在 90% 的置信度下可以把这两个分布区分开; 方差与相对误差满足线性关系, 拟合知斜率为 2.018; 局部分辨率和整体分辨率保持一致. 此外, 分辨率的模拟计算与理论推导一致.

## 2 准备工作

### 2.1 EM 算法的相关原理

用 Y 表示可被观测的随机变量,  $N \geq 2$ , 对应 Y 的 N 次简单随机抽样观测值记为  $y = (y_1, y_2, \dots, y_N)$ , 用 Z 表示隐含随机变量或被遗漏观测的随机变量. 通常情况下, Z 隐含或遗漏在观测数据中, y 称为不完全数据, 若对 Z 补上取值  $z = (z_1, z_2, \dots, z_N)$ , 则  $(y, z)$  为完全数据. 给定观测数据  $Y = y$ , 其概率函数记为  $P(y|\theta)$ , 其中  $\theta \in \Theta$ ,  $\Theta$  是需要估计的参数空间. Y 和 Z 的联合分布记为  $P(y, z|\theta)$ . EM 算法是求某一对数似然函数的极大似然估计, 每次迭代由 E 步 (期望步) 和 M 步 (极大化) 组成, 反复迭代 E 步和 M 步直至收敛<sup>[8]</sup>.

EM 算法具体为: 第  $i$  次迭代结果记  $\theta^{(i)}$ , 第  $i+1$  次迭代开始的参数设置为  $\theta^{(i)}$ , 则第  $i+1$  次迭代 E 步和 M 步分别为:

E 步: 求关于概率函数  $P(z|y, \theta^{(i)})$  的条件期望  $E_{\theta^{(i)}}[\log P(y, Z|\theta)|\theta^{(i)}]$ , 并记

$$Q(\theta|\theta^{(i)}) = E_{\theta^{(i)}}[\log P(y, Z|\theta)|\theta^{(i)}],$$

如果随机变量 Z 是连续型, 则

$$Q(\theta|\theta^{(i)}) = \int_{z \in Z} \log P(y, z|\theta) P(z|y, \theta^{(i)}) dz,$$

如果随机变量  $Z$  是离散型, 则

$$Q(\theta|\theta^{(i)}) = \sum_Z \log P(y, z|\theta)P(z|y, \theta^{(i)}),$$

这里  $\{P(z|y, \theta^{(i)}): z \in \mathbb{R}^d\}$  表示在给定参数  $\theta = \theta^{(i)}$  和观测数据  $Y = y$  下隐变量  $Z$  的条件密度.

M 步: 极大化  $Q(\theta|\theta^{(i)})$ , 通过极大似然估计找到  $\theta^{(i+1)}$  即

$$\theta^{(i+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta|\theta^{(i)}),$$

这样就实现了一次  $\theta^{(i)} \rightarrow \theta^{(i+1)}$  的迭代, 反复重复 E 步和 M 步, 直到对给定的正数  $\varepsilon_1$  和  $\varepsilon_2$  下列条件成立:

$$\|\theta^{(i+1)} - \theta^{(i)}\| < \varepsilon_1$$

或

$$\|Q(\theta^{(i+1)}|\theta^{(i)}) - Q(\theta^{(i)}|\theta^{(i)})\| < \varepsilon_2,$$

则迭代停止.

## 2.2 两分量模型 EM 算法的迭代公式

对于二分量模型  $Y = \Delta X_1 + (1 - \Delta)X_2$ , 其中  $\Delta \sim B(1, \pi)$ ,  $X_1 \sim F_1(x; \theta_1)$ ,  $X_2 \sim F_2(x; \theta_2)$  ( $\theta_1, \theta_2 \in \Theta$ ), 在  $X_1, X_2, \Delta$  相互独立的前提下, 可以推导出 EM 算法中  $\pi$  及  $F_1, F_2$  中参数的迭代公式, 对二分量高斯混合模型 [4], 其参数迭代公式如下:

$$\left\{ \begin{array}{l} \mu_1^{(i+1)} = \frac{\sum_{k=1}^N \gamma_{k1}^{(i+1)} y_k}{\sum_{k=1}^N \gamma_{k1}^{(i+1)}}, \quad \sigma_1^{2(i+1)} = \frac{\sum_{k=1}^N \gamma_{k1}^{(i+1)} (y_k - \mu_1)^2}{\sum_{k=1}^N \gamma_{k1}^{(i+1)}}, \\ \mu_2^{(i+1)} = \frac{\sum_{k=1}^N (1 - \gamma_{k1}^{(i+1)}) y_k}{\sum_{k=1}^N (1 - \gamma_{k1}^{(i+1)})}, \quad \sigma_2^{2(i+1)} = \frac{\sum_{k=1}^N (1 - \gamma_{k1}^{(i+1)}) (y_k - \mu_2)^2}{\sum_{k=1}^N (1 - \gamma_{k1}^{(i+1)})}, \\ \pi^{(i+1)} = \frac{\sum_{k=1}^N \gamma_{k1}^{(i+1)}}{N}, \end{array} \right.$$

## 2.3 EM 算法的参数分辨率

本文以两分量混合模型参数估计的 EM 算法为例, 给出参数分辨率的定义: 对两正态混合模型:  $Y = \Delta X_1 + (1 - \Delta)X_2$ , 其中  $\Delta \sim B(1, p)$ ,  $X_1 \sim N(\mu, \sigma^2)$ ,  $X_2 \sim N(\nu, \sigma^2)$ ,  $\mu \neq \nu$  以及  $\sigma^2$  为已知参数. EM 算法对参数  $(\mu, \nu)$  的分辨率定义准备如下:

(1) 对混合模型进行  $n$  轮抽样, 对每轮抽样结果, 运用 EM 算法对参数  $(\mu, \nu)$  重新进行估计给出估计值,  $k$  次估计值记为  $(\mu^{(k)}, \nu^{(k)}), k = 1, 2, \dots, n$ ;

(2) 采用判别准则, 判别每轮估计是否将参数  $\mu$  与  $\nu$  区分开, 如能区分开, 则此轮 EM 算法是成功的;

(3) 令  $d_n \triangleq n$  轮估计中成功次数 /  $n$ .

其中成功分开的标准由 Fisher 判别准则和阈值 (置信度)  $\alpha$  决定: Fisher 判别准则 [3]

$$\left\{ \begin{array}{l} |\mu_i^* - \mu| \leq |\nu_i^* - \mu|, \\ |\nu_i^* - \nu| \leq |\mu_i^* - \nu|, \end{array} \right. \quad (2.1)$$

其中  $\mu_i^*$  表示 EM 算法  $\mu_i$  估计值, 其它参数类似; 若分辨率达到所给阈值则终止计算.

**定义 2.1** 设置阈值  $\alpha(0 < \alpha < 1)$ , 当  $d_n \geq \alpha$  时, 称 EM 算法能将此对参数  $(\mu, \nu)$  是可以分开, 并称相对误差 (RE)  $\frac{|\mu-\nu|}{\mu}$  或者绝对误差 (AE)  $|\mu - \nu|$  为阈值等于  $\alpha$  时, EM 算法对模型 (两正态混合) 的参数分辨率.

### 3 实验与分析

EM 算法在参数估计中存在随机性, 每轮迭代结果可能不同, 为了避免这种随机性, 首先, 在每组模型下循环 100 轮, 得到 100 组 EM 算法估计值. 其次, 分辨率计算时, 把运行 50 次结果进行升序排列, 依次去除前三个和后三个数值后取平均值. 这样不仅克服了一定的偶然性, 而且也提高了 EM 算法的鲁棒性. 除此, 考虑到 EM 算法易受初始值影响, 初始值均在真实值的 10% 左右, 精度为  $10^{-6}$ ,  $p = 0.5$ , 绝对误差先取 0.01 确定出大致区间, 然后取 0.001 进行精选区间, 针对阈值、相对误差、变异系数三个方面进行相应的结果分析.

#### 3.1 相对误差与变异系数的影响分析

固定  $\sigma = 0.1$ , 选择不同模型, 实验按照固定  $\mu$  调整  $\nu$  和固定  $\nu$  调整  $\mu$  这两种方式进行, 阈值设置 85%、90%、95% 时, 计算对应的相对误差和变异系数, 变异系数是衡量各观测值离散程度的一个统计量, 是方差与均值的比值, 简记为 CV, 计算公式:  $CV = \frac{\sigma}{\mu}$ . 如表 1 当模型为  $\mu = 1, \nu = 1.189$ , 此时  $d_n$  为 84.6818% 未达到阈值 85%, 则需要固定  $\mu = 1$  来增加  $\nu$  的值进行估计, 直至  $d_n$  大于等于阈值. 各表中 RE1 和 RE2 表示固定  $\mu$  调整  $\nu$  和固定  $\nu$  调整  $\mu$  所对应的相对误差;  $\delta$  表示平均相对误差, 如表 1 中  $\delta = \frac{RE1+RE2}{2} = \frac{0.1900+0.1860}{2}$ ; CV1 和 CV2 表示固定  $\mu$  调整  $\nu$  和固定  $\nu$  调整  $\mu$  所对应的变异系数.

表 1 阈值为 85% 时的参数分辨率

固定 $\mu$ 调整 $\nu$	RE1	AE1	固定 $\nu$ 调整 $\mu$	RE2	$\delta$	CV1	CV2
$\mu = 1.0, \nu = 1.190$	0.1900	0.190	$\mu = 0.814, \nu = 1.0$	0.1860	0.1880	0.1840	0.2229
$\mu = 1.2, \nu = 1.391$	0.1592	0.191	$\mu = 1.010, \nu = 1.2$	0.1583	0.1588	0.1552	0.1823
$\mu = 1.6, \nu = 1.792$	0.1200	0.192	$\mu = 1.407, \nu = 1.6$	0.1206	0.1203	0.1183	0.1336
$\mu = 1.8, \nu = 1.991$	0.1061	0.191	$\mu = 1.606, \nu = 1.8$	0.1078	0.1070	0.1058	0.1178
$\mu = 2.0, \nu = 2.192$	0.0960	0.192	$\mu = 1.806, \nu = 2.0$	0.0970	0.0965	0.0956	0.1054
$\mu = 3.0, \nu = 3.192$	0.0640	0.192	$\mu = 2.809, \nu = 3.0$	0.0637	0.0639	0.0647	0.0689

表 2 阈值为 90% 时的参数分辨率

固定 $\mu$ 调整 $\nu$	RE1	AE1	固定 $\nu$ 调整 $\mu$	RE2	$\delta$	CV1	CV2
$\mu = 1.0, \nu = 1.200$	0.2000	0.200	$\mu = 0.798, \nu = 1.0$	0.2020	0.2010	0.1833	0.2253
$\mu = 1.2, \nu = 1.400$	0.1667	0.200	$\mu = 1.000, \nu = 1.2$	0.1667	0.1667	0.1548	0.1833
$\mu = 1.6, \nu = 1.804$	0.1275	0.204	$\mu = 1.395, \nu = 1.6$	0.1281	0.1278	0.1179	0.1342
$\mu = 1.8, \nu = 2.005$	0.1139	0.205	$\mu = 1.597, \nu = 1.8$	0.1128	0.1134	0.1054	0.1182
$\mu = 2.0, \nu = 2.204$	0.1020	0.204	$\mu = 1.795, \nu = 2.0$	0.1025	0.1023	0.0954	0.1057
$\mu = 3.0, \nu = 3.206$	0.0687	0.206	$\mu = 2.796, \nu = 3.0$	0.0680	0.0684	0.0645	0.0691

表 3 阈值为 95% 时的参数分辨率

固定 $\mu$ 调整 $\nu$	RE1	AE1	固定 $\nu$ 调整 $\mu$	RE2	$\delta$	CV1	CV2
$\mu = 1.0, \nu = 1.220$	0.2200	0.220	$\mu = 0.780, \nu = 1.0$	0.2200	0.2200	0.1820	0.2282
$\mu = 1.2, \nu = 1.420$	0.1833	0.220	$\mu = 0.980, \nu = 1.2$	0.1833	0.1833	0.1538	0.1854
$\mu = 1.6, \nu = 1.827$	0.1419	0.227	$\mu = 1.376, \nu = 1.6$	0.1400	0.1410	0.1172	0.1352
$\mu = 1.8, \nu = 2.025$	0.1250	0.225	$\mu = 1.576, \nu = 1.8$	0.1244	0.1247	0.1049	0.1190
$\mu = 2.0, \nu = 2.226$	0.1130	0.226	$\mu = 1.778, \nu = 2.0$	0.1110	0.1120	0.0949	0.1062
$\mu = 3.0, \nu = 3.224$	0.0747	0.224	$\mu = 2.775, \nu = 3.0$	0.0750	0.0749	0.0644	0.0694

由表 1、表 2、表 3 可知, 阈值设置为 85% 时, 绝对误差  $0.186 \leq AE \leq 0.194$ ; 阈值设置为 90% 时, 绝对误差  $0.20 \leq AE \leq 0.206$ ; 阈值设置为 95% 时, 绝对误差  $0.22 \leq AE \leq 0.227$ ; 由于迭代结果的随机性和抽取样本的数量有限, 导致绝对误差得到的是范围值. 若阈值设置为 90%, 可知绝对误差大于 0.206 时, 两个相近的信号一定可以分开.

阈值设置 85%、90%、95% 时, 固定  $\mu$  调整  $\nu$  和固定  $\nu$  调整  $\mu$  这两种方式, 通过计算对应的相对误差和变异系数的值, 下面给出直观上的结果分析.

由图 1(a)~(c) 可知, 方差  $\sigma = 0.1$  固定时, 同一模型, 增加或减少一定绝对误差, 达到相同阈值时, 所对应的相对误差相差很近或达到相同值, 则下面只研究绝对误差增加时的情形, 大于上限值表示两个相近参数可以分开, 小于下限值表示两个相近参数分不开.

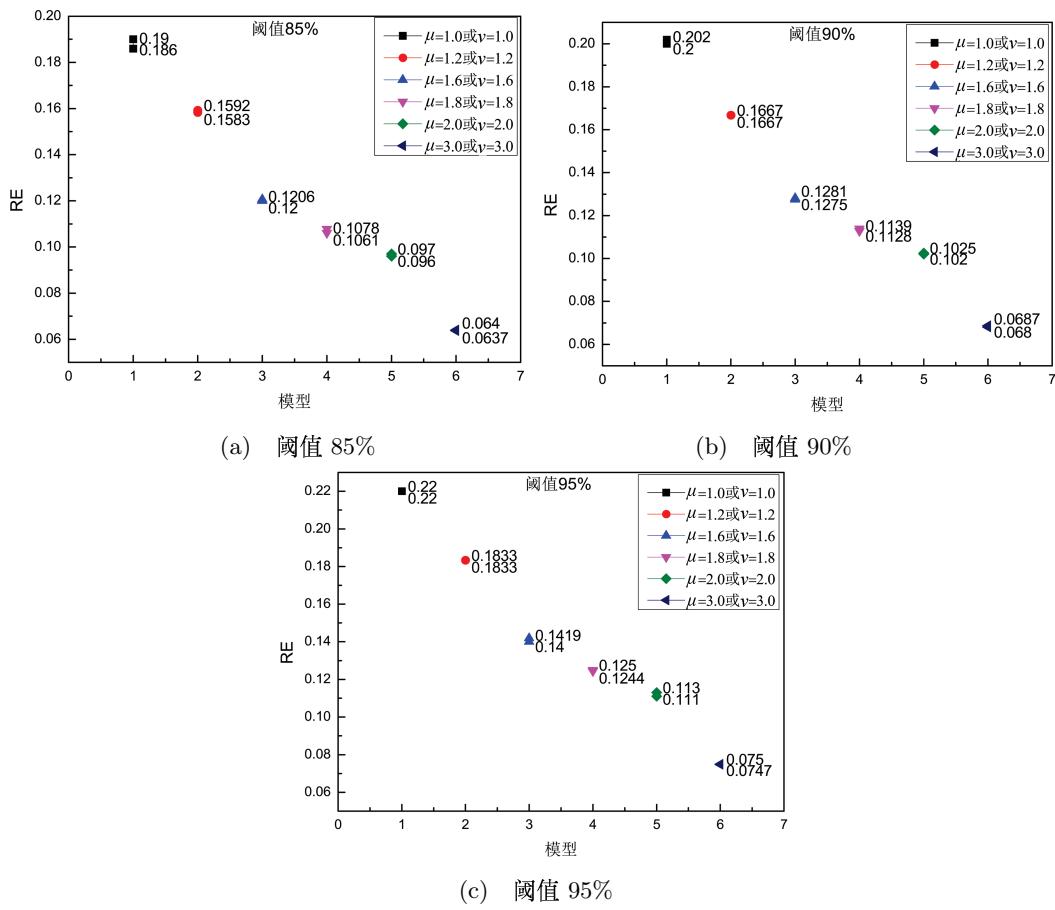


图 1

由图 2(a)~(c) 可知, 首先, 阈值依次设置 85%、90%、95% 时, 相对误差单调性与变异系数的单调性保持一致, 均是递减的趋势. 除此, 变异系数的倒数表示信号的强度, 从图中可知, 信号越强, 相对误差越小.

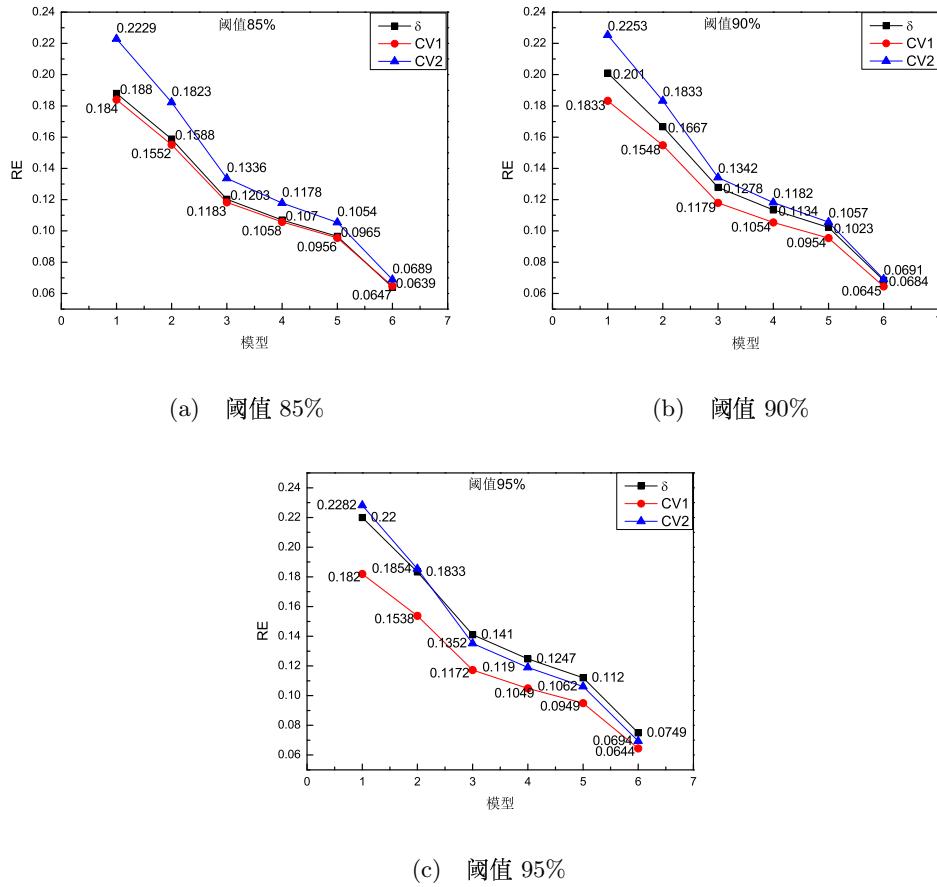


图 2

### 3.2 相对误差与分辨率的影响分析

基于变异系数影响参数分辨率的计算结果, 本小节将进行两方面的研究: 一方面将方差依次设置为 0.1、0.2、0.3、0.4、0.5, 分析方差变化对相对误差的影响, 另一方面由 EM 算法估计出 100 组参数, 利用判别准则 (2.1) 运行 50 次后把结果进行升序排名, 把最大数值记为极大分辨率、最小数值记为极小分辨率、并依次去除前三个和后三个数值后取平均值得到平均分辨率即整体分辨率, 原始模型为  $0.5N(1, 0.1) + 0.5N(1.1, 0.1)$ , 固定  $\sigma = 0.1$ , 随着相对误差增大, 对极值分辨率与整体分辨率的变化趋势进行分析.

由图 3(a) 可知, 阈值设置 90%、95% 时相对误差与方差满足线性关系, 通过拟合得斜率分别为 2.018、2.172. 图 3(b)~3(d) 采用旋转和归一化后, 两线段重合验证了达到不同阈值时, 相对误差的变化值不大. 其中图 3(c) 是权重归一化, 图 3(d) 是 0~1 归一化.

由图 4 可知, 当相对误差增大, 极值分辨率与整体分辨率均以不减的趋势变化, 说明局部分辨率与整体分辨率保持一致; 也符合信号越强, 相对误差越小的特点.

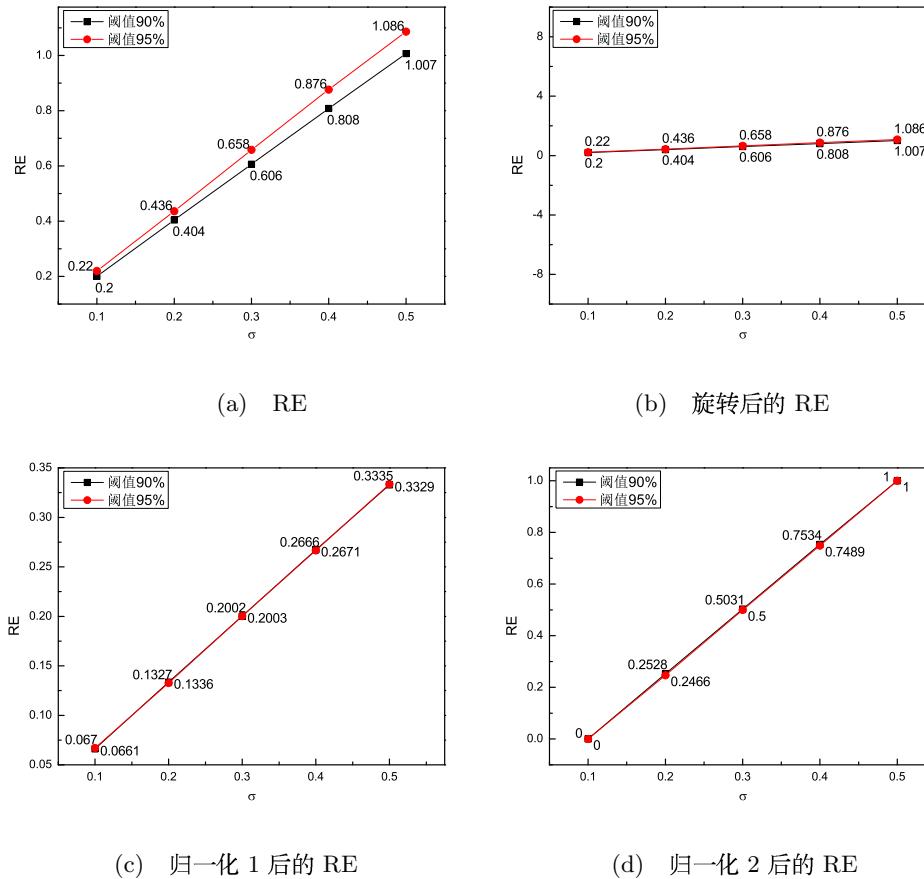


图 3

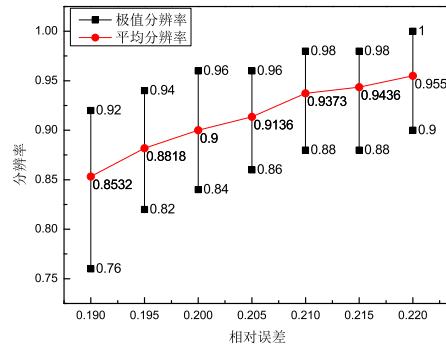


图 4 整体分辨率与局部分辨率

### 3.3 Fisher 信任分布的相关理论

基于 Fisher 的判别准则、混合高斯分布的相关原理以及正态分布密度函数的性质，推导出当  $\sigma$  固定时，两个相近参数分开的信任分布值。首先，引入信任分布的概念。

若样本 (容量是 1)， $\theta$  为未知参数。则  $X - \theta \sim N(0, 1)$ ，即对任意实数  $t$ ，有

$$P(X - \theta < t) = \Phi(t), \quad (3.1)$$

其中,  $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{y^2}{2}} dy$ . 此式可改写为

$$P(\theta > X - t) = \Phi(t), \quad P(\theta < X - t) = 1 - \Phi(t). \quad (3.2)$$

从概率的角度知 (3.1) 和 (3.2) 式没有实质性的变化. 但 Fisher 赋予 (3.2) 式信任意义: 确定样本  $X$  后, 将  $\theta$  看成随机变量, 而 (3.2) 式给出了  $\theta$  的分布, 这个分布称  $\theta$  的信任分布.

**定理 3.1** 设随机变量  $\xi_1 \sim N(\mu, \sigma^2)$ ,  $\xi_2 \sim N(\nu, \sigma^2)$ , 若置信区间为  $(\mu - 2\sigma, \nu + 2\sigma)$ , 则 EM 算法参数估计对应的分开区域的概率为

$$P = \frac{[\Phi(\frac{\nu-\mu}{2\sigma}) - (1 - \Phi(2))] - [\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})]}{\Phi(\frac{\nu-\mu}{2\sigma}) - (1 - \Phi(2))} + \lambda \frac{\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})}{\Phi(\frac{\nu-\mu}{2\sigma}) - (1 - \Phi(2))}, \quad (3.3)$$

其中,  $\lambda$  为弃真或存伪所占的概率, 随阈值的不同而不同.

**证** 令  $S$  表示整体区域、 $S_1$  表示可能误分的区域、 $S_2$  表示不可能误分的区域, 且满足  $S = S_1 + S_2$ .  $S_3$  表示 EM 算法参数估计对应的分开区域.

随机变量  $\xi_1$  和  $\xi_2$  的密度函数分别记为  $f_1$  和  $f_2$ . 从图 5(a) 易求得正态分布  $f_1 = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi_1-\mu)^2}{2\sigma^2}}$  与  $f_2 = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi_2-\nu)^2}{2\sigma^2}}$  的交点坐标为  $\frac{\mu+\nu}{2}$ , 由 (3.2) 式易知  $P(\xi_1 < \mu + 2\sigma) = \Phi(2) = 0.9772$ ,  $P(\xi_1 < \mu - 2\sigma) = P(\xi_1 > \mu + 2\sigma) = 1 - P(\xi_1 < \mu + 2\sigma) = 1 - \Phi(2) = 0.0228$ , 以及  $P(\xi_1 < \frac{\mu+\nu}{2}) = \Phi(\frac{\nu-\mu}{2\sigma})$ .

那么整体区域  $S$  的大小 ( $A(S)$ ):

$$A(S) = 2 \times P(\mu - 2\sigma < \xi_1 < \frac{\mu+\nu}{2}) = 2 \times [\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228],$$

可能误分区域  $S_1$  的大小 ( $A(S_1)$ ):

$$A(S_1) = 2 \times P(\frac{\mu+\nu}{2} < \xi_1 < \nu + 2\sigma) = 2 \times [\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})],$$

不可能误分区域  $S_2$  的大小 ( $A(S_2)$ ):

$$A(S_2) = A(S) - A(S_1) = 2 \times [\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228] - 2 \times [\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})],$$

$A(S_2)$  对应的概率大小 ( $P(S_2)$ )

$$\begin{aligned} P(S_2) &= \frac{A(S) - A(S_1)}{A(S)} \\ &= \frac{2 \times [\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228] - 2 \times [\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})]}{2 \times [\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228]} \\ &= \frac{[\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228] - [\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})]}{[\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228]}. \end{aligned}$$

记 EM 算法参数估计对应的分开区域  $S_3$  的大小  $A(S_3)$ , 则  $A(S_3) = A(S_2) + \lambda(A(S) - A(S_2))$ ,  $\lambda$  为比例因子, 随不同的阈值而不同. 故对应的分开区域的概率大小 ( $P(S_3)$ )

$$\begin{aligned} P(S_3) &= P(S_2) + \lambda(1 - P(S_2)) \\ &= \frac{[\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228] - [\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})]}{\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228} + \lambda \frac{\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{2\sigma})}{\Phi(\frac{\nu-\mu}{2\sigma}) - 0.0228}. \end{aligned}$$

故原命题成立.

**注 3.1** 图 5(b) 为阈值分别等于 85%、90%、95% 时, 比例因子的变化图, 从图中可知阈值为 90% 时,  $0.455 \leq \lambda \leq 0.4843$ .

**推论 3.1** 在同一阈值下, 当  $\frac{\nu-\mu}{\sigma} \geq 1.09$  时, 弃真或存伪所占的概率  $\lambda$  随样本  $n$  的增大而增大.

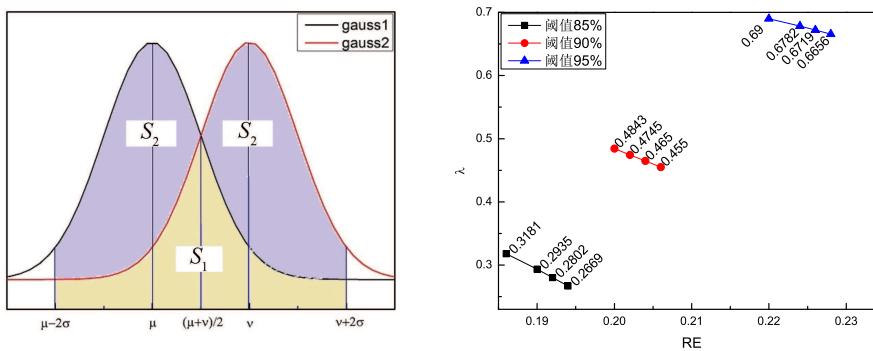
**证** 根据实验结果可知, 随着样本容量  $n$  增大, EM 算法对应的分辨率  $\nu - \mu$  的值会减小, 由定理 3.1 知, 当  $\frac{\nu-\mu}{\sigma} \geq 1.09$  时, 有

$$(1 - \lambda) \frac{\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{\sigma})}{\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228} = 1 - P,$$

这里  $P$  对应于参数分辨率定义中设置的阈值  $\alpha$ . 再由

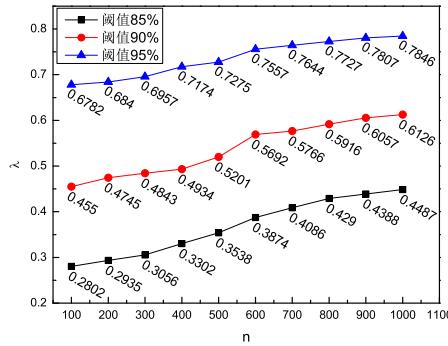
$$\begin{aligned} \frac{\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{\sigma})}{\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228} &= \frac{[\Phi(\frac{\nu-\mu}{\sigma} + 2) - 0.0228] - [\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228]}{\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228} \\ &= \frac{1 - 0.0228}{\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228} - 1 = \frac{0.9772}{\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228} - 1 \end{aligned}$$

可知  $\frac{\Phi(\frac{\nu-\mu}{\sigma} + 2) - \Phi(\frac{\nu-\mu}{\sigma})}{\Phi(\frac{\nu-\mu}{\sigma}) - 0.0228}$  的值增大. 从而. 随着样本容量  $n$  增大  $\lambda$  的值增大.



(a) 正态分布图

(b) 相对误差与比例因子图



(c) 样本量与比例因子图

图 5

**注 3.2** 图 5(c) 为  $\sigma = 0.1$ , 固定  $\mu = 3$  调整  $\nu$ , 随着样本容量  $n$  的增大, 阈值分别达到 85%、90%、95% 时, 比例因子的变化图, 从图中可知达到不同阈值时,  $\lambda$  均以递增的趋势变化.

## 4 结论

算法评估指标日益重要, 常用的评估指标有简单性、鲁棒性、计算速度及准确度, 前三者已取得很多成果, 而后者却没有学者研究. 由于分辨率是衡量准确度的有效指标, 如果可以提前预知分辨率, 那么就可以很好的衡量准确度, 本文以两分量高斯混合模型为例, 给出参数分辨率定义, 并运用 Fisher 思想给出 (2.1) 式的判别准则, 此准则不受随机性和主观因素的影响, 进而由 (3.2) 式推导出 (3.3) 式两个相近参数分开的信任分布. 以两正态混合模型为例进行验证, 得到了参数分辨率在 EM 算法中的可行性. 实验表明两个方差为 0.1 的正态分布其均值距离大于 0.206 时, EM 算法在 90% 的置信度下可以区分这两个分布; 方差与相对误差满足线性关系, 通过拟合知斜率为 2.018; 局部分辨率与整体分辨率一致; 除此, 通过构建实验结果和理论推导之间的联系, 得到不同置信度下的比例因子图. 参数分辨率的提出, 不仅为区分两个相近的信号提供了新的指导方案, 而且为评估 EM 算法的准确度提供了新的参考依据.

## 参 考 文 献

- [1] Berlinet A F, Roland C. Acceleration of the EM algorithm: P-EM versus epsilon algorithm. *Computational Statistics & Data Analysis*, 2012, **56**(12): 4122–4137
- [2] Cucker F, Zhou D X. Learning Theory: An Approximation Theory Viewpoint. Volume 24. Cambridge: Cambridge University Press, 2007
- [3] 陈希孺. 数理统计学教程. 安徽: 中国科技大学出版社, 2009  
Chen X R. Mathematical Statistics Course. Anhui: China University of Science and Technology Press, 2009
- [4] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977, **39**(1): 1–38
- [5] Ikeda S. Acceleration of the EM algorithm. *Systems Computers in Japan*, 2015, **31**(2): 10–18
- [6] Kuroda M, Sakakihara M. Accelerating the convergence of the EM algorithm using the vector  $\varepsilon$  algorithm. *Comput Stat Data Anal*, 2006, **51**(3): 1549–1561
- [7] Liu C, Rubin D B. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 1994, **81**(4): 633–648
- [8] 李航. 统计学习方法. 北京: 清华大学出版社, 2012: 156–169  
Li H. Statistical Learning Methods. Beijing: Tsinghua University Press, 2012: 156–169
- [9] Li Y, Chen Y. Research on Initialization on EM algorithm based on Gaussian mixture model. *Journal of Applied Mathematics & Physics*, 2018, **06**(1): 11–17
- [10] McLachlan G, Peel D. Finite Mixture Models. New York: Springer-Verlag, 2000
- [11] Shi P, Tsai C. Regression model selection a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2002, **64**(2): 237–252
- [12] Tan Q H, Jiang H J, Ding Y M. Model selection method based on maximal information coefficient of residuals. *Acta Mathematica Scientia*, 2014, **34**(2): 579–592
- [13] Wu C. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 1983, **11**(1): 95–103
- [14] Wei G G, Tanner M. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Publications of the American Statistical Association*, 1990, **85**(411): 699–704
- [15] Xu L, Jordan M I. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 1995, **8**(1): 129–151
- [16] Yu J, Chaomuriligea C, Yang M S. On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures. *Pattern Recognition*, 2018, **77**(1): 188–203

- [17] Yao W. A note on EM algorithm for mixture models. *Statistics & Probability Letters*, 2013, **83**(2): 519–526
- [18] Zhai D H, Yu J, Gao F, et al. K-means text clustering algorithm based on initial cluster centers selection according maximum distance. *Application Research of Computers*, 2014, **31**(3): 713–719

## Research on Resolution Based on EM Algorithm

Lu Nana Yu Jinghu

(*Department of Mathematics, School of Science, Wuhan University of Technology, Wuhan 430070*)

**Abstract:** Parameter resolution is a criterion for measuring whether two adjacent signals can be distinguished under the given noise conditions, it provides an evaluation of the “ruler” for the measurement of sensitive parameters, effective precision and accuracy. This paper proposes a definition of the parameter resolution of EM algorithm, which is based on the EM algorithm, and the idea of Fisher linear discriminant criterion, two-component Gaussian mixed model is taken as an example to verify it. Experiments show that when two normal distributions with a variance of 0.1 have a mean distance greater than 0.206, the EM algorithm can tell the differences between the two distributions under a confidence of 90%, by constructing the connection between experimental results and theoretical derivation, the scale factor graphs with different confidence levels are obtained. The proposed resolution of the parameters provides a quantitative indicator for the accuracy measurement and also provides a new solution for the differentiation of similar signals.

**Key words:** EM algorithm; Resolution; Criterion; Coefficient of variation.

**MR(2010) Subject Classification:** 62P99; 94A12