密级：公开　　中图分类号：O212.1

浙江工商大學

硕士学位论文

论文题目：基于DA-EM改进算法的有限混合模型参数估计

作者姓名： 杨圣杰

学科专业： 统计学

研究方向： 数理统计

指导教师： 王伟刚

提交日期：2020 年1月

Dissertation Submitted to Zhejiang Gongshang University

for Master's Degree of Science

Parameter Estimation Of Finite Mixture Model Based On Modified DA-EM Algorithm

Author: **Yang Shengjie**

Major: **Statistics**

Supervisor: **Prof. Wang Weigang**

Jan. 2020

School of Statistics and Mathematics

Zhejiang Gongshang University

Hangzhou, 310018, P. R. China

摘　要

大数据背景下, 很多数据往往来自不同的分组. 由于能较好刻画这种异质性,有限混合模型受到了统计学界和计算机界的重视. 对有限混合模型中的数理问题进行研究, 是很多应用展开的首要任务.其中, 应用最广泛的是高斯混合模型. 本文则主要研究高斯混合模型的参数估计问题.

相较于经典EM算法, 确定性退火期望最大化算法(Deterministic annealing expectation-maximization algorithm, DA-EM算法)引入了退火参数, 有效解决了隐变量的后验概率估计过度依赖参数初始值的问题. 然而使用过小的退火参数初始值极易导致DA-EM 算法收敛到稳定点而使迭代过早停止. 对于此问题, 有学者提出了基于Jacobian矩阵分析选择退火参数理论下界的方法, 但该方法仍不能完全避免模型均值向量的重合.

针对上述问题, 本文提出一种改进的DA-EM算法. 在给定退火参数理论下界的DA-EM算法基础上, 该算法通过引入两个惩罚项, 分别对混合比例和模型参数的距离施以惩罚来修改对数似然函数. 模拟数据和真实数据的实验结果均表明: 在一定条件下, 改进的DA-EM算法对高斯混合模型的参数估计效果更加准确和稳定, 并且其收敛速度也快于DA-EM算法.

关键词：高斯混合模型；DA-EM算法；退火参数；惩罚似然方法；SCAD罚函数

Abstract

In the context of big data, many data often come from different groups. The finite mixture model is highly regarded by the statistics and computer industry because of describing this heterogeneity very well. There are lots of researches on mathematical problems in finite mixture models. The first task of application deployment. Gaussian mixture model is the most widely used among them. Naturally, this paper mainly studies the parameter estimation problem of Gaussian mixture model.

Compared with the classical EM algorithm, the Deterministic annealing expectation-maximization algorithm (DA-EM algorithm) introduces annealing parameters. However, using the initial value of the annealing parameter too small can easily cause the DA-EM algorithm to converge to a stable point and cause the iteration to stop prematurely. To solve this problem, some scholars proposed a method based on Jacobian matrix analysis to select the theoretically lower bound of the annealing parameter, but this method still cannot completely avoid the coincidence of the model mean vector.

Aiming at the above problems, this paper proposes an modified DA-EM algorithm. On the basis of the DA-EM algorithm given the theoretically lower bound of the annealing parameter, the method modifies the log likelihood function by introducing two penalty terms and penalizing the mixture ratio and the distance of the model parameters. At the same time, the method maintains the convergence of the original algorithm. The experimental results of the simulated data and the real data show that: Under certain conditions, the modified DA-EM algorithm estimates the parameters of the Gaussian mixture model. The effect is more accurate and stable, and its convergence speed is faster than DA-EM algorithm.

Keywords: Gaussian mixture model; DA-EM algorithm; Annealing parameter; Penalized likelihood approach; SCAD penalty function

目录

第 1 章　引言

1.1　研究背景及现状

1.1.1　选题背景

随着电子计算机及互联网的迅速发展, 有限混合模型的潜力和应用价值得到了各界的广泛认可. 混合模型已经被应用到社会科学、经济学、天文学、遗传学、生物学等领域, 是生存分析、模式识别、聚类判别、潜在变量分析等统计方法的研究土壤. 这些行业都出现了海量、复杂、高纬度的数据, 一般的单一分布无法在其中挖掘出有用的信息, 而混合模型则能有效地解决这个问题. 有限混合模型作为一种便利的, 半参数方法具有很强的灵活性. 不同类型的混合模型适用于不同的领域: 指数混合模型适用于工程领域; 泊松混合模型适用于医学领域; 而高斯混合模型的适用范围最广. 因为在样本量足够大的前提下, 任意一个分布都可以用若干个高斯分布去逼近.

然而, 不同于传统的单一分布模型, 有限混合模型的参数估计问题的处理方式较为复杂. 极大似然估计(MLE)作为经典的参数估计方法, 在混合模型里不能直接得到解析解. 因此, 人们对混合模型参数估计的早期研究方法还停留在矩方法上, 但由于矩方法的计算过于复杂(在高维数据上尤为明显), 这导致混合模型的研究进展十分缓慢. 而计算机的出现打破了这一现状. 1972年, Tan和Robertson开始使用极大似然法(ML)做高斯混合模型的参数估计, 并从理论上证明了该方法优于矩方法. 紧接着, Dempster(1977)提出了EM算法用于计算有限混合模型的极大似然估计, 大幅降低了参数估计的计算量, 这也进一步为有限混合模型的推广和应用奠定了基础.

有限混合模型自19世纪末被提出以来, 经过数学家, 统计学家等对其理论基础、应用算法的深入研究, 已成为有效拟合复杂密度的建模工具. 混合模型的基本结构简单, 但其分布相当灵活. 参数估计和模型选择作为基础工作, 却是很多研究展开的首要任务, 因此具有重要的理论意义. 尽管EM算法解决了极大似然方法没有解析解的问题, 且相关定阶方法在一定程度上防止了过拟合, 但依旧存在收敛速度过慢、效果不稳定、数值求解算法复杂等问题. 本文对有限混合模型的参数估计及定阶进行研究, 能进一步丰富该领域数理研究的理论体系.

有限混合模型由于能够对异质数据进行建模, 在时下热门的图像处理、人工智能、模式识别等领域受到大量学者的关注. 特别地, 在如今的大数据环境下, 海量的样本数据往往来自很多种类或分组. 有限混合模型正是刻画这种异质性(heterogeneity)的有力建

模工具. 由于有限混合模型的参数估计和定阶为聚类判别、生存分析等统计方法提供了支撑, 提高相关算法的计算速度、准确性和稳定性愈显重要; 同样地, 挑选出一个恰当又简洁模型能有效降低模型的复杂度. 因此, 本文研究内容具有明显的实际应用价值, 能进一步促进有限混合模型在许多科学领域的发展.

1.1.2 研究现状

EM算法的出现为高斯混合模型带来了长足发展, 其自身也因为简单实用、易于计算机编程的实现、应用广泛等特点, 吸引很多学者对其进行大量研究, 为它的良好性能提供了理论支撑, 并给出了很多著名的EM衍生算法. Cheng(1987)给出了EM算法在曲指数簇下的收敛成果; Xu和Jordan(1995)指出了EM算法具有良好的收敛性, 并给出了收敛速度; Ma和Fu(2005)指出当高斯混合模型各个成分的重叠部分足够小时, EM算法可以估计模型参数的真值; Tanner(1990)对于EM算法不易求得期望表达式的情况, 提出了采用近似计算的MCEM算法, 但是该方法依赖抽样, 估计值会随着迭代次数的增加在真值上下波动; 为了加快EM算的收敛速度, Louis(1982)提出了Atiken加速方法, Lange(1995)提出了拟牛顿加速方法; Meng和Rudin向EM算法中加入随机的思想, 提出了随机EM(SEM)算法(1991), 并通过分开估计参数的方式提出了期望条件最大化(ECM)算法(1993); Ma和Ge(2007)通过研究混合模型的几何性质, 给出了用于线状模式发现的广义EM算法.

使用极大似然法估计高斯混合模型参数的前提是给定模型的分支数, 在模型分支数未知的情况下, 想要继续使用极大似然法估计模型参数, 则需要寻找其他途径解决模型选择的问题. 除了一些经典的模型选择方法, 如: AIC(Akaike, 1973), BIC(Schwarz, 1978). 还有一些专门作用于有限混合模型的模型选择方法. Leroux(1992)讨论了经典有限混合模型中基于AIC 和BIC的定阶问题. Chen和Kalbfleisch(1996)提出的惩罚最小距离法, 这种方法通过引入关于混合比率的惩罚项修改对数似然函数, 来防止混合比率取值过小的过拟合类型. 第三类是假设检验类, 有C(α)检验(Neyman和Scott, 1966)、似然比检验(Ghosh和Sen, 1985; Chen和Chen, 2001)、改进似然比检验(Chen等, 2001, 2004; Fu等, 2008)、D-检验(Charnigo和Sun, 2004)和EM-检验(Li和Chen, 2004; Chen等, 2012; Chen和Li, 2012; Chen等, 2017); Chen和Khalili(2008)提出的惩罚似然方法通过对混合率和成分参数距离进行惩罚, 能够同时实现模型定阶及参数估计. Kim和Seo(2014)讨论了在多个局部极大值存在的情况下, 高斯混合模型中成分个数的估计. 此外, 还有Peng(2016)提出的期望水平准则(desirability level criterion), Grimm(2017)提出的k-折交叉验证方法等, 都给有限混合模型的定阶问题提供了新的解决方案.

虽然EM算法性能良好, 但它本身也存在着一些缺陷, 比如: 对参数初始值非常敏感、不能确保收敛到全局最优、需要提前给定隐变量的所有可能的取值等. 针对这些缺陷, 很多学者对EM算法进行了改进. 由于EM算法的估计效果在很大程度上依赖参数的初始值, 合适的参数初始值不仅能加快算法的收敛速度, 还能使得算法避免收敛到局部最优. Figueiredo(2002)发现EM算法对初值敏感以及参数估计值可能位于空间边界等缺点, 改进了EM算法同时证明了EM算法具备自退火性质; Pal(2002)等提出了一种随机EM算法来处理参数初始化问题; 解决EM 算法易于收敛到局部极值的另一类方法就是全局性算法与EM算法相结合的形式, 如粒子群算法(Hametner和Jakubek, 2010)、遗传算法(Santos et al, 2017)、退火算法(Shah et al, 2014)等使得EM算法尽量避免收敛到局部最优. 由Ueda和Nakano(1998)提出的确定性退火算法(DA-EM)进一步实现了全局性算法与经典EM算法结合的基础. Yu等(2018)通过Jacobian矩阵指出必须存在一个退火参数的理论下界, DA-EM算法才能有效避免收敛到稳定点.

1.2 本文的创新点

针对DAEM算法存在退火参数过小极易收敛到稳定点的缺陷, Yu等(2018)通过分析DAEM算法差分方程的Jacobian矩阵提出了一种选择DAEM 算法退火参数理论下界的方法. 然而, 此方法仍存在一些缺陷, 如混合比例位于空间边界的第一类过拟合问题, 无法完全避免各模型均值点的重合问题. 为此, 本文提出一种改进的DA-EM算法(MDA-EM算法), 该方法在对数似然函数上, 引入了两个惩罚项分别对混合比例和模型参数的距离进行惩罚有效解决了这两个缺陷. 通过模拟和对UCI数据集iris进行实证分析表明, 相比于DA-EM算法, 本文提出的算法获得的参数估计更加准确和稳定, 其迭代次数也相对较少.

1.3 本文组织框架

本文主要内容分为5章, 具体组织框架如下:

第1章引言, 主要介绍了有限混合模型和EM算法的研究背景, 研究现状, 并简要阐述了各种经典方法的优缺点和本文的创新点.

第2章准备知识, 不仅介绍了高斯混合模型、EM算法、最大熵原理以及确定性退火算法的基础知识, 而且给出了EM算法和DA-EM算法估计高斯混合模型参数的详细步骤, 并简要推导了退火参数下界的求解过程.

第3章首先介绍了惩罚似然方法及SCAD罚函数, 然后对DAEM算法进行改进提出了MDA-EM算法, 并给出了MDA-EM算法的具体步骤和详细的数值求解过程.

第4章从模拟和实证两个角度,通过对比三种算法(CEM, DA-EM, MDA-EM)的参数估计效果, 验证了MDA-EM算法的优越性.

第5章是总结与展望, 总结概括本文所研究的主要问题、方法及成果, 并对未来的拓展性工作进行安排.

第 2 章　准备知识

2.1　高斯混合模型

高斯混合模型(Gaussian Mixture Model, GMM)是一种常用的基本分布为高斯分布的混合密度模型. 它相当于多个高斯概率密度函数的加权平均, 每个高斯密度函数被称为一个成分(或者称为分模型), 各个成分的参数相互独立. 当分模型的个数足够大时, 高斯混合模型可以以很高的精度去逼近任意一个连续分布, 因此它能较好地刻画数据的空间分布及其特性.

下面, 给出高斯混合模型的详细定义.

在空间$\mathbb{R}^p$中, 满足高斯混合模型的随机变量X的概率密度函数为

$$p(x|\theta) = \sum_{k=1}^{K} \alpha_k \phi\left(x|\mu_k, \Sigma_k\right), \tag{2-1}$$

其中$\theta = \left(\left(\alpha_1, \mu_1, \Sigma_1\right), \cdots, \left(\alpha_K, \mu_K, \Sigma_K\right)\right)$, α_k是每个成分的权重, 称为"混合比例", 且满足条件

$$\forall k = 1, 2, \ldots, K, \alpha_k > 0, \sum_{k=1}^{K} \alpha_k = 1,$$

$\phi\left(x|\mu_k, \Sigma_k\right)$是第k个分模型. 满足高斯概率密度函数

$$\phi\left(x|\mu_k, \Sigma_k\right) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right).$$

假设样本数据集$D = \{x_1, x_2, \cdots, x_n\}$由高斯混合模型(2-1)生成, 可以认为$\forall j = 1, 2, ..., n$, 样本$x_i$的产生过程为: 首先依据概率$\alpha_k$选择第$k$个成分, 然后通过第$k$个成分的概率密度函数$\phi\left(x|\mu_k, \Sigma_k\right)$生成样本$x_i$, 我们可以直接观测到样本$x_i$, 但无法判断该样本具体是由哪个分模型产生的$x_i$, 因此, 可以把样本数据隶属的高斯混合成分称为隐变量. 该隐变量的所有可能取值$\{1, 2, \cdots, c\}$, 作为样本x_i的隐含类别标记.

显然, 我们可以考虑采用极大似然法估计高斯混合模型的参数. 但是, 所观测到的样本数据并不包含高斯混合成分的信息, 待估参数对应的对数似然函数也没有解析解, 所以直接进行极大似然参数估计非常困难. 因此, 对于高斯混合模型问题的求解常使用期望最大(Expectation Maximum, EM)算法通过不断求解下界的极大化逼近对数似然函数极大化的方式对参数进行估计.

2.2 EM算法

当概率模型中既包含观测变量(可以直接观测的变量)又包含隐变量(观测变量对应的潜在变量)时, 对于给定的样本, 我们已经不能再直接使用极大似然法或者贝叶斯法估计模型参数. Dempster(1977)提出一种迭代算法–EM算法, 主要用于存在隐变量的概率模型参数的极大似然估计. 其主要思想: 把极大化似然函数的过程分解成多步迭代过程, 通过条件不断求解下界的极大化逼近对数似然函数极大化的方法估计参数. EM算法的每轮迭代可以分为两个步骤: E步, 利用上一轮迭代所获的估计参数, 重新计算完全对数似然函数(包含观测变量和隐变量的对数似然函数)关于隐变量的条件期望; M步, 极大化完全对数似然函数的条件期望, 更新参数的估计值.

假设空间$\mathbb{R}^d$中有观测变量X, 其概率密度函数为$p(x|\theta)$, 其中θ为模型待估参数; 空间$\mathbb{R}^{d'}$中有观测变量X对应的不被观测到的变量Z, $p(x,z|\theta)$为X和Z的联合概率密度函数. 具体来说, 在高斯混合模型中, 随机变量Z可以作为样本x 隶属的高斯混合成分, 为样本x的隐含标记变量.

X的样本观测数据集为$D=\{x_1,x_2,\cdots,x_n\}$, 称为"不完全数据", 其对应的对数似然函数为$L(\theta)=\sum_{i=1}^{n}\ln P\left(X=x_i|\theta\right)$. X对应的隐变量Z的样本数据集为$\{Z_1,Z_2,\cdots,Z_n\}$, 其中$\forall i=1,2,\ldots,n$, $Z_i=\{1,2,\ldots,K\}$是一个不能被直接观测的变量. 如果隐变量Z也被观测到了, 那么$\{(x_1,z_1),\cdots,(x_n,z_n)\}$ 称为"完全数据".

对于含有离散隐变量的概率模型,隐变量Z的所有可能取值为$\{1,2,\ldots,K\}$. 我们通过极大化观测数据D关于参数θ的对数似然函数来估计成分参数, 即极大化

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{n}\ln P\left(X=x_i|\theta\right) \\ &= \sum_{i=1}^{n}\ln\sum_{k=1}^{K}P\left(X=x_i,Z_i=k|\theta\right) \\ &= \sum_{i=1}^{n}\ln\left(\sum_{k=1}^{K}P\left(X=x_i|Z_i=k,\theta\right)P\left(Z_i=k|\theta\right)\right). \end{aligned} \tag{2-2}$$

由于对数似然函数(2-2)中包含了未观测样本以及和的对数, 使得极大化过程变得非常复杂, 因此考虑到了迭代的方式, 逐步增加对数似然函数$L(\theta)$以估计参数, 这便是EM算法的本质思想.

假设第t轮迭代后的参数估计值为$\theta^{(t)}$, 为了确保对数似然函数不断增加, 所得的新的

参数估计值θ需要满足$L(\theta) > L\left(\theta^{(t)}\right)$. 因此考虑两者之差

$$
\begin{aligned}
& L(\theta) - L\left(\theta^{(t)}\right) \\
=& \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right) \right) - \sum_{i=1}^{n} \ln P\left(X = x_i | \theta^{(t)}\right) \\
=& \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \frac{P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right)}{P\left(Z_i = k | x_i, \theta^{(t)}\right)} \right) - \sum_{i=1}^{n} \ln P\left(X = x_i | \theta^{(t)}\right),
\end{aligned}
$$

利用Jensen不等式可得

$$
\begin{aligned}
& L(\theta) - L\left(\theta^{(t)}\right) \\
\geq& \sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln \frac{P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right)}{P\left(Z_i = k | x_i, \theta^{(t)}\right)} - \sum_{i=1}^{n} \ln P\left(X = x_i | \theta^{(t)}\right) \\
=& \sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln \frac{P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right)}{P\left(Z_i = k | x_i, \theta^{(t)}\right) P\left(X = x_i | \theta^{(t)}\right)},
\end{aligned}
$$

所以, 极大化$L(\theta)$, 只需极大化

$$
\sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln \frac{P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right)}{P\left(Z_i = k | x_i, \theta^{(t)}\right) P\left(X = x_i | \theta^{(t)}\right)},
$$

从而可以得出

$$
\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln \frac{P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right)}{P\left(Z_i = k | x_i, \theta^{(t)}\right) P\left(X = x_i | \theta^{(t)}\right)} \right) \\
&= \arg\max_{\theta} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln P\left(X = x_i | Z_i = k, \theta\right) P\left(Z_i = k | \theta\right) \right) \\
&= \arg\max_{\theta} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln P\left(X = x_i, Z_i = k | \theta\right) \right),
\end{aligned}
$$

令

$$
Q\left(\theta, \theta^{(t)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} P\left(Z_i = k | x_i, \theta^{(t)}\right) \ln P\left(X = x_i, Z_i = k | \theta\right).
$$

下面给出EM算法的具体步骤.

Step1: 给定参数初始值$\theta^{(0)}$;

Step2(E步): 在第t轮迭代得到的参数估计值$\theta^{(t)}$的基础上, 计算$Q\left(\theta, \theta^{(t)}\right)$函数 (简称,

Q函数)

$$Q\left(\theta,\theta^{(t)}\right)=\sum_{i=1}^{n}E_Z\left[\ln p\left(x_i,z|\theta\right)|x_i,\theta^{(t)}\right].$$

Step3(M步): 通过最大化$Q\left(\theta,\theta^{(t)}\right)$, 获得第$t+1$轮迭代的参数估计值$\theta^{(i+t)}$;

Step4: 令$t:=t+1$, 重复Step2-3, 直到Q函数收敛;

这里, 需要给出EM算法的几点解释说明:

(1) EM算法的迭代过程依赖参数的初始值(迭代最初的参数值), 因此在很大程度上, 参数初始值决定了EM算法参数估计的效果、迭代路径和次数.

(2) EM算法中的每一轮迭代都会使$L\left(\theta\right)$增加或者达到局部极值, 也就是说EM算法不能保证找到全局最优值;

(3) Q函数收敛一般是指, 对于较小的正常数ε,若Q函数满足

$$\left|Q\left(\theta^{(t+1)},\theta^{(t)}\right)-Q\left(\theta^{(t)},\theta^{(t)}\right)\right|<\varepsilon.$$

则停止迭代, 算法达到收敛.

另外, EM算法的收敛性由以下定理提供支持和保证.

Theorem 2.1 *假设$L(\theta)$为观测数据的对数似然函数, $\theta^{(t)}(t=1,2,\dots)$为EM算法得到的参数估计序列, $L\left(\theta^{(t)}\right)(t=1,2,\dots)$为对应的对数似然函数序列.*

(1) 若对数似然函数$L(\theta)$存在上界, 那么$L(\theta^{(t)})$收敛到某一值L^*;

(2) 在函数$Q\left(\theta,\theta^{(t)}\right)$和$L(\theta)$满足一定的条件下,通过EM算法得到的参数估计序列$\theta^{(t)}(t=1,2,\dots)$的收敛值$\theta^*$是对数似然函数$L(\theta)$的稳定点.

定理2.1中的(1)体现了EM算法关于对数似然函数序列$L(\theta^{(t)})$的收敛性,(2)体现看EM算法关于参数估计序列$\theta^{(t)}(t=1,2,\dots)$的收敛性,前者并不蕴含后者.除此之外, 定理只能保证参数估计序列收敛到对数似然函数序列的稳定点, 并不能保证收敛到最大值点. 换句话说, EM算法迭代得到的参数估计值θ^*只能保证是局部最优的, 并不能确保是全局最优. 因此在应用过程中, 一般选取几个不同的初始值进行迭代, 然后通过比较对应的对数似然函数的大小选择最优的参数估计值.

2.2.1 基于EM算法的高斯混合模型参数估计

本节将EM算法应用到高斯混合模型参数估计问题中. 沿用前文的符号定义, 我们假

设数据集D中的样本为独立同分布样本, 且都是从同一个高斯混合模型中产生

$$\begin{aligned} p(x|\theta) &= \sum_{k=1}^{K} \alpha_k \cdot \phi\left(x|\mu_k, \Sigma_k\right) \\ &= \sum_{k=1}^{K} \alpha_k \cdot \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right). \end{aligned}$$

首先, 对于高斯混合模型, 我们使用极大似然法构造对数似然函数

$$L(\theta) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \left(\alpha_k \cdot \phi\left(x_i|\mu_k, \Sigma_k\right)\right), \tag{2-3}$$

然后再使用EM算法估计高斯混合模型的参数.

Step1: 给定参数初始值$\theta^{(0)} = \left(\left(\alpha_1^{(0)}, \mu_1^{(0)}, \Sigma_1^{(0)}\right), \cdots, \left(\alpha_K^{(0)}, \mu_K^{(0)}, \Sigma_K^{(0)}\right)\right)$.

Step2(E步): 令θ^t为经过第t轮迭代得到的参数估计值, 第$t+1$迭代过程中Q函数$Q\left(\theta, \theta^{(t)}\right)$为

$$Q\left(\theta, \theta^{(t)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} \ln\left(\alpha_k \cdot \phi\left(x_i|\mu_k, \Sigma_k\right)\right) \cdot P\left(Z_i = k|x_i, \theta^{(t)}\right), \tag{2-4}$$

令后验概率$\omega_{ik}^{(t+1)} = P\left(Z_i = k|x_i, \theta^{(t)}\right)$, 通过贝叶斯公式计算得出

$$\begin{aligned} \omega_{ik}^{(t+1)} &= P\left(Z_i = k|x_i, \theta^{(t)}\right) \\ &= \frac{P\left(Z_i = k, X = x_i|\theta^{(t)}\right)}{\sum_{k'=1}^{K} P\left(Z_i = k', X = x_i|\theta^{(t)}\right)} \\ &= \frac{P\left(X = x_i|Z_i = k, \theta^{(t)}\right) \cdot P\left(Z_i = k|\theta^{(t)}\right)}{\sum_{k'=1}^{K} P\left(X = x_i|Z_i = k', \theta^{(t)}\right) \cdot P\left(Z_i = k'|\theta^{(t)}\right)} \\ &= \frac{\phi\left(x_i|\mu_k^{(t)}, \Sigma_k^{(t)}\right) \cdot \alpha_k^{(t)}}{\sum_{k'=1}^{K} \phi\left(x_i|\mu_{k'}^{(t)}\Sigma_{k'}^{(t)}\right) \cdot \alpha_{k'}^{(t)}}. \end{aligned} \tag{2-5}$$

Step3(M步): 最大化式(2-4), 得到第$t+1$轮迭代的参数估计值$\theta^{(t+1)}$的更新公式.

(1) 求解$\mu_k^{(t+1)}$的更新公式.

对式(2-4)关于μ_k求偏导, 并且令偏导为0,得到

$$\frac{\partial Q\left(\theta, \theta^{(t)}\right)}{\partial \mu_k} = \sum_{i=1}^{n} \omega_{ik}^{(t+1)} \Sigma_k^{-1}\left(x_i - \mu_k\right) = 0, \tag{2-6}$$

求解式(2-6)可以得出$\mu_k^{(t+1)}$的更新公式

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n \omega_{ik}^{(t)} x_i}{\sum_{i=1}^n \omega_{ik}^{(t+1)}}. \tag{2-7}$$

(2) 求解$\Sigma_k^{(t+1)}$的更新公式.

对式(2-4)关于Σ_k求偏导, 并且令偏导为0, 得到

$$\frac{\partial Q\left(\theta, \theta^{(t)}\right)}{\partial \Sigma_k} = \frac{1}{2}\sum_{i=1}^n \omega_{ik}^{(t+1)} \left(\Sigma_k^{-1}\left(x_i - \mu_k\right)\left(x_i - \mu_k\right)^T \Sigma_k^{-1} - \Sigma_k^{-1}\right) = 0, \tag{2-8}$$

将式(2-7)代入式(2-8), 并计算得出$\Sigma_k^{(t+1)}$ 的更新公式

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^n \omega_{ik}^{(t)} \left(x_i - \mu_k^{(t+1)}\right)\left(x_i - \mu_k^{(t+1)}\right)^T}{\sum_{i=1}^n \omega_{ik}^{(t)}}. \tag{2-9}$$

(3) 求解$\alpha_k^{(t+1)}$的更新公式.

由于α_k满足条件$\sum_{k=1}^K \alpha_k = 1$, 可以利用拉格朗日乘子法, 引入拉格朗日乘子$\lambda$, 对Q函数(2-4)构造拉格朗日函数

$$L(\theta, \lambda) = Q\left(\theta, \theta^{(t)}\right) + \lambda\left(\sum_{k=1}^K \alpha_k - 1\right), \tag{2-10}$$

对式(2-10)关于α_k求偏导为0, 可以得出

$$\alpha_k = -\frac{\sum_{i=1}^n \omega_{ik}^{(t+1)}}{\lambda}, \tag{2-11}$$

将式(2-11)代入$\sum_{k=1}^K \alpha_k = 1$, 可以得到

$$\sum_{k=1}^K \alpha_k = -\frac{\sum_{k=1}^K \sum_{i=1}^n \omega_{ik}^{(t+1)}}{\lambda} = -\frac{n}{\lambda} = 1 \Rightarrow \lambda = -n, \tag{2-12}$$

将式(2-12)代入式(2-11), 得到$\alpha_k^{(t+1)}$ 的更新公式

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^n \omega_{ik}^{(t+1)}}{n}. \tag{2-13}$$

Step4: 令$i := i + 1$, 重复步骤Step2-3, 使用式(2-13)、(2-7)、(2-9)更新参数估计值,

直到算法收敛,即Q函数(2-4) 收敛.

2.3 确定性退火EM算法

EM算法是一种迭代方法, 迭代的结果较大程度上依赖于参数初始值, 并且只能保证收敛到局部最优值. 为了改善EM算法的这一系统性缺陷, Ueda等(1998)提出了确定性退火EM算法(DA-EM). DA-EM算法在经典EM算法中引入了退火机制, DA-EM算法在每个退火参数值下都可以找到局部最优值, 并在退火参数缓慢增加的过程中, 使用上一次迭代得到的局部最优的参数值作为新的算法迭代的初始值进行新一轮的迭代. 相较经典EM算法, 在一定条件下, DA-EM算法可以以更高的概率收敛到全局最优值.

2.3.1 最大熵原理

十九世纪中叶, Clausius首次提出了"熵"这个概念, 并将其定义为物理系的内部状态的参量, 来表示系统内部的混乱程度. 熵越大, 则表示系统内部越混乱.在卡诺定理基础上, Clausius于1850年提出了热力学第二定律(熵增定律), 它指出一个独立系统的混乱程度会逐渐增大, 直到它达到最混乱(熵最大)的状态. 此时系统内部达到平衡态.

过了将近一个世纪, Shannon(1948)提出了信息论, 引入了熵的概念, 从概率论的角度定义了"信息熵", 并给出了计算公式. 设连续型随机变量X的概率密度函数为$f(x)$, 作用区间为$[a,b]$, 则信息熵的计算公式为

$$H(X) = -k\int_a^b f(x)\ln f(x)dx,$$

其中, k为正常数.

假设X为离散型随机变量, 其所有可能的取值为$\{x_1, x_2, \cdots, x_K\}$, 令$X$取值为$x_i$的概率$P(X = x_j)$ 为$p(x_i)$, 则熵的计算公式为

$$H(X) = -k\sum_{i=1}^{c} p(x_i)\ln p(x_i).$$

作为度量随机事件不确定性的标准: 信息熵, 其值越大表示信息量越少. 当随机变量服从均匀分布时, 信息熵达到最大. 这让熵的概念不再局限于热力学而是可以与概率统计学关联起来, 应用到数学各领域之中.

随机变量的概率分布可以计算熵, 那么, 由随机变量的熵, 是否可以来求其概率分布函数呢? 根据此想法以及热力学的熵增定律, Jaynes(1957)提出了最大熵原理, 并为正态

分布、指数分布、几何分布等多种常用概率密度函数提供了理论基础和过程推导, 使熵原理走出了热力学, 走进了统计的大门.

最大熵原理的核心思想: 在推断随机变量的概率分布时, 充分考虑已知信息(通常为随机变量函数的期望), 且不对任何未知信息作人为假设. 信息的增加会引起信息熵的减少. 因此, 在已有的信息条件约束下, 令熵最大的概率分布所包含的未知信息的假设最少, 此概率分布也就最符合实际. 基于最大熵原理得到的概率分布, 是在满足已知信息基础上, 对所以未知信息都做了均匀假设的条件, 得到的唯一"无偏"分布.

接下来沿用上文定义的符号, 下面给出关于离散型随机变量X的最大熵原理的数学表示, 连续型随机变量自行类推. 假设随机变量X 服从m个约束条件: $\varphi_i\left(g_i(X)\right) = C_i$,$i = 1, 2, \ldots, m$. 其中, $g_i(X)$为随机变量X的函数, $\varphi_i\left(g_i(X)\right)$为$g_i(X)$的期望, C_i 为常数. 则基于这些约束条件的最大熵原理为

$$\max \quad H(X) = -\sum_{i=1}^{K} p\left(x_i\right) \ln p\left(x_i\right),$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{K} p\left(x_i\right) = 1 \\ \varphi_i\left(g_i(X)\right) = C_i, \quad i = 1, 2, \ldots, m \end{cases}$$

2.3.2 确定性退火算法

液体粒子排列无序, 冷却物体时随着温度的缓慢下降, 粒子的热运动减少, 粒子运动逐渐趋于有序. 当温度降至凝结温度时, 粒子运动变为围绕晶体格点的微小震动, 达到较有序的结晶态.

根据Boltzman有序性原理可知, 当温度降低时, 封闭系统在每个温度下达到平衡态的过程遵循热力学定律——对于与周围环境交换热量而温度保持不变的封闭系统, 系统状态总是朝着自由能减少的方向自发进行. 当自由能达到最小值时, 系统达到平衡. 系统的自由能函数可以表示为$F = E - TS$, 其中E表示系统能量, T表示系统温度, S表示系统的熵. 所以在某一固定温度下, 系统状态是系统能量和系统熵相互竞争的结果. 在温度较高时, 系统熵S的权重较大, 系统趋于熵增大的方向, 粒子较为无序. 当温度较低时, 系统熵S的权重变小而使得系统能量E的权重增大, 系统状态趋于能量减小的方向, 粒子整体较为有序.

根据统计力学, 令Z表示系统状态变量, E表示系统能量变量, 封闭系统处于状态k, 系统能量为E_k时的概率分布为Gibbs分布

$$P\left(E=E_k\right)=P(Z=k)=\frac{1}{\gamma}\exp\left(-\frac{E_k}{\alpha T}\right),\quad k=1,2,\ldots,K, \tag{2-14}$$

其中,K为系统状态数, $\gamma=\sum_{k=1}^{K}\exp\left(-\frac{E_k}{\alpha T}\right)$为分配函数, α为Boltzman常数, T为系统温度.

当$T\to\infty$时, 系统处于所有状态的概率相同, Gibbs分布退化成均匀分布; 当$T\in(0,\infty)$时, 系统所处状态的能量越低, 其处于该状态的概率越大, 即系统处于能量较低状态的概率较高; 当$T\to 0$时, 系统处于能量最低时的状态.

根据退火过程, 确定性退火技术将需要求解的极小值问题转化为系统能量$E=E(\theta)$, 其中θ为极小值问题的带估参数. 构造系统能量E对应地满足以下条件的自由能函数$\min F(\theta,T)$:

(1)当$T\to\infty$时, $\min F(\theta,T)$关于θ的全局极小值容易求出;

(2)当$T=0$时, $F(\theta,T)=E(\theta)$.

确定性退火技术利用问题求解过程与物理退火过程的相似性, 将求解极小值的问题转化为自由能函数极小化问题. 确定性退火技术, 以系统在上一温度下自由能函数的极小点作为初值点, 通过求解下一温度(更低温度)的自由能函数极小值在模拟系统达到平衡态的过程.

根据上述分析, 下面给出确定性退火算法的具体步骤:

Step1: 令$i=0$, 取初始温度T_0足够大, 在此温度下, 自由能函数$\min F(\theta,T)$对应的参数估计值$\theta(T_0)$容易求出;

Step2: 降低温度, 令$T_{i+1}=\delta T_i$, 其中$0<\delta<1$, 在得到$\theta(T_i)$的前提下, 求解自由能函数$\min F(\theta,T)$, 得到$\theta(T_{i+1})$, 然后令$i:=i+1$;

Step3:重复步骤Step2, 直到$T_{i+1}\to 0$, 且参数估计值收敛.

2.3.3 基于DA-EM算法的高斯混合模型参数估计

在EM算法中, 后验概率密度函数W_{ik}在参数估计中起到了关键作用, 但是它在迭代的初期是不可靠的, 尤其是给定的参数初始值大幅偏离真实值时. 因此EM算法容易收敛到局部最优, 而DA-EM算法结合EM算法与确定性退火技术, 以改善EM算法依赖初始值的缺陷.

DA-EM算法中引入退火参数β来模拟退火过程中温度的变化. 一般而言, 温度等于退火参数的倒数, 即$T=\frac{1}{\beta}$. 温度缓慢降低, 相应的, 在DA-EM算法中, 从一个较小的退

Algorithm 1 基于DA-EM框架的高斯混合模型参数估计算法

Input: 观测数据$X = \{x_1, \ldots, x_n\}$, 高斯混合模型,$\epsilon_1, \epsilon_2 > 0$, 初始退火参数$\beta \leftarrow \beta^0$, $t = 0$

Output: 高斯混合模型参数$\theta = ((\alpha_1, \mu_1, \Sigma_1), \cdots, (\alpha_K, \mu_K, \Sigma_K))$

1: 参数初始值$\theta^{(0)} = \left(\left(\alpha_1^{(0)}, \mu_1^{(0)}, \Sigma_1^{(0)}\right), \cdots, \left(\alpha_K^{(0)}, \mu_K^{(0)}, \Sigma_K^{(0)}\right)\right)$

2: **while** $\beta \leq 1$ **do**

3: 根据式(2-15)计算后验概率ω_{ik}, 即每个分模型对观测数据的响应度;

4: 根据式(2-16), (2-17), (2-18)计算当前混合模型的参数;

5: 判断两次迭代的最大似然估计参数不在变化, $\left\|\Theta^{(t+1)} - \Theta^{(t)}\right\| < \epsilon_1$, 或者Q函数不再变化, $\left\|Q\left(\theta^{(t+1)}, \theta^{(t)}\right) - Q\left(\theta^{(t)}, \theta^{(t)}\right)\right\| < \epsilon_2$. 否则, $t = t + 1$, 重复步骤3,4;

6: **end while**

火参数β开始, 逐渐增大参数值.为了让β缓慢增长,这里取β 的增长倍数为1.01. 当$\beta = 1$时, DA-EM算法则退化成EM算法, 因此EM算法可以被看作是DA-EM算法的一个特例($\beta = 1$).不同于EM算法, DA-EM算法在后验概率的计算上引入了退火参数β

$$\omega_{ik} = \frac{\left(\alpha_k \phi\left(x_i | u_k, \Sigma_k\right)\right)^\beta}{\sum_{k=1}^K \left(\alpha_k \phi\left(x_i | u_k, \Sigma_k\right)\right)^\beta}. \tag{2-15}$$

对于高斯混合模型而言, 如果ω_{ik}已知, 那么 $\theta = ((\alpha_1, \mu_1, \Sigma_1), \cdots, (\alpha_K, \mu_K, \Sigma_K))$ 的最大似然估计与EM算法相同

$$\alpha_k = \frac{\sum_{i=1}^n \omega_{ik}}{n}, \tag{2-16}$$

$$\mu_k = \frac{\sum_{i=1}^n \omega_{ik} x_i}{\sum_{i=1}^n \omega_{ik}}, \tag{2-17}$$

$$\Sigma_k = \frac{\sum_{i=1}^n \omega_{ik} \left(x_i - \mu_k\right)\left(x_i - \mu_k\right)^T}{\sum_{i=1}^n \omega_{ik}}. \tag{2-18}$$

在DA-EM算法中, 我们将最大化对数似然函数问题转化为求解最小化自由能函数. 算法从一个较高的温度, 也就是从较小的退火参数β^0开始. 但是, 若选取的退火参数β^0过小, $\beta^0 \to 0$, 则计算得出的后验概率$\omega_{ik} \to \frac{1}{c}, \forall i, k$. 换句话说, 所有的样本隶属于所有类的概率都相同. 在下次迭代中, 因样本信息而引起的概率变动不足以改变带了退火参数ω的后验概率(2-15), 仍将得到上一次的迭代结果, 从而导致迭代过程提前结束. 因此, 我们在算法初始化阶段, 选择一个合适的退火参数β初始值尤为关键. 然而, Ueda等(1998)的文献中并没有理论工作是围绕退火参数初始化展开的.

2.4 基于Jacobian矩阵的退火参数下界求解

本节简要介绍Yu等(2018)基于雅各比矩阵的DA-EM算法分析, 由于篇幅限制, 我们

将直接给出相关定理结论, 省略其冗长的推导过程.

首先, 先介绍Jacobian矩阵. Jacobian矩阵是函数的一阶偏导数以一定的形式排列成的矩阵, 其行列式称为Jacobian行列式. 我们考虑参数空间为

$$\Psi = \left\{ \omega = [\omega_{ik}]_{n\times K} \,|1 \le i \le n,\ 1 \le k \le K, \omega_{ik} \ge 0, \sum_{i=1}^{K} \omega_{ik} = 1 \right\},$$

下面求解DA-EM算法的Jacobian矩阵.

对于固定参数β, 将DA-EM算法的迭代过程重新写成参数映射

$$\omega_{ik}^{(t+1)} = \theta_{ik}^{(\beta)}(\omega) = \frac{\left(d_{ik}^{(t+1)}\right)^{\beta}}{\sum_{j=1}^{K}\left(d_{ij}^{(t+1)}\right)^{\beta}}, \tag{2-19}$$

其中,

$$d_{ik}^{(t+1)} = \alpha_{ik}^{(t+1)} \left(\det\left(\Sigma_k^{(t+1)}\right)\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(x_i - \mu_k^{(t+1)}\right)^T \left(\Sigma_k^{(t+1)}\right)^{-1}\left(x_i - \mu_k^{(t+1)}\right)\right), \tag{2-20}$$

且$\alpha_i^{(t+1)}$, $\mu_i^{(t+1)}$, $\Sigma_i^{(t+1)}$与上文相同, 分别为高斯混合模型的混合比例、均值向量、协方差矩阵参数, t表示迭代次数. 如果算法收敛至$\omega^* \in \Psi$, 那么ω^*必须满足

$$\omega^* = \theta^{(\beta)}(\omega^*). \tag{2-21}$$

当$\omega^* = [K^{-1}]_{n\times K}$, 根据式(2-16),(2-17),(2-18)可得$\mu_k^* = \sum_{i=1}^{n} x_i/n = \bar{x}, \forall k, \alpha_k^* = \frac{1}{K}, \forall k$. 那么, DA-EM算法的唯一聚类中心就是样本的均值向量. 而在下一次迭代中, 仍会得到相同结果 (算法仍收敛于$\omega^* = [K^{-1}]_{n\times K}$) , DA-EM算法因其迭代限制提前结束, 从而导致聚类失败. 显然, 在实际应用过程中我们应该避免得到这样的聚类结果. 如果在算法的初始阶段令$\beta \leftarrow 0$, 那么DA-EM 算法将不可避免地收敛于$\omega^* = [K^{-1}]_{n\times K}$.

考虑到文章的完整性及可读性, 我们给出以下定理. 定理的证明参见Yu等(2018)第190-194页.

Theorem 2.2 *$i = 1, \cdots, n, j = 1, \cdots, n, k = 1, \cdots, K, r = 1, \cdots, K-1$, 那么DA-EM算法Jacobian矩阵的元素等于*

$$\frac{\partial \theta_{ik}^{(\beta)}}{\partial \omega_{jr}} = -\beta\frac{\omega_{ik}\omega_{ir}}{2n\alpha_r}(H_r)_{ij} + \beta\frac{\omega_{ik}\omega_{iK}}{2n\alpha_K}(H_K)_{ij} + \beta\frac{\delta_{kr}\omega_{ik}}{2n\alpha_k}(H_k)_{ij}, \tag{2-22}$$

其中

$$(H_r)_{ij} = -\left(x_j-\mu_r\right)^T\Sigma_r^{-1}\left(x_j-\mu_r\right)+s+1-\left(x_i-\mu_r\right)^T\Sigma_r^{-1}\left(x_i-\mu_r\right) + \left(\left(x_j-\mu_r\right)^T\Sigma_r^{-1}\left(x_i-\mu_r\right)+1\right)^2,$$

$$\delta_{kr}=\begin{cases}0 & if\, k\neq r\\ 1 & if\, k=r\end{cases}$$

定理2.2给出了DA-EM算法的Jacobian矩阵的通用形式. Jacobian矩阵在某一点的谱半径能反映该点的性质: 若Jacobian矩阵在该点的谱半径$\rho<1$, 则说明算法在该点收敛. 根据式(2.25)可以看出Jacobian矩阵的计算与退火参数β相关. 如果退火参数没有合适的初始值, 那么算法就有可能收敛于$\omega^*=[K^{-1}]_{n\times K}$. 为了分析算法在稳定点的Jacobian矩阵, 我们先定义矩阵操作vec(M).

Definition 2.1 *对于$p\times q$维度的矩阵M, vec(M)就是对矩阵M按列叠加得到的向量. 对于矩阵$M=(m_1,\cdots,m_q)$, 其中$m_1,\cdots,m_q$是矩阵M的列向量,* vec$(M)=\left(m_1^T,\cdots,m_q^T\right)^T$.

Lemma 2.1 *DA-EM算法在点$\omega^*=[K^{-1}]_{n\times K}$的Jacobian矩阵$\frac{\partial\theta^{(\beta)}(\omega)}{\partial\omega}$的形式为*

$$\left.\frac{\partial\theta_{ik}^{(\beta)}}{\partial\omega_{jr}}\right|_{\forall i,k,\omega_k=K^{-1}}=\beta\frac{\delta_{kr}}{2n}(H)_{ij}=\beta\frac{\delta_{kr}}{2n}\left(A^TA\right)_{ij}, \tag{2-23}$$

其中

$$(H_r)_{ij} = -\left(x_j-\bar{x}\right)^T\sigma_x^{-1}\left(x_j-\bar{x}\right)+s+1-\left(x_i-\bar{x}\right)^T\sigma_x^{-1}\left(x_i-\bar{x}\right) + \left[\left(x_j-\bar{x}\right)^T\sigma_x^{-1}\left(x_i-\bar{x}\right)+1\right]^2,$$

$$\bar{x}=n^{-1}\sum_{i=1}^n x_i,\sigma_x=n^{-1}\sum_{i=1}^n\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)^T,$$

$$\delta_{kr}=\begin{cases}0 & k\neq r\\ 1 & k=r\end{cases}$$

$$A=\begin{bmatrix}\sqrt{2} & \cdots & \sqrt{2}\\ \sqrt{2}\sigma_x^{-1/2}\left(x_1-\bar{x}\right) & \cdots & \sqrt{2}\sigma_x^{-1/2}\left(x_n-\bar{x}\right)\\ \text{vec}\left(\sigma_x^{-1/2}\left(x_1-\bar{x}\right)\left(x_1-\bar{x}\right)^T\sigma_x^{-1/2}\right) & \cdots & \text{vec}\left(\sigma_x^{-1/2}\left(x_n-\bar{x}\right)\left(x_n-\bar{x}\right)^T\sigma_x^{-1/2}\right)\end{bmatrix}.$$

下面, 我们根据得到的Jacobian矩阵估计DA-EM的退火参数下界.

显然, DA-EM算法应避免收敛到点$\omega^* = [K^{-1}]_{n\times K}$, 换句话说, 退火参数的初始值$\beta^{(0)}$使得DA-EM算法在点$\omega^* = [K^{-1}]_{n\times K}$的Jacobian矩阵的谱半径大于等于1. 这样DA-EM算法才不会在初始阶段就直接收敛到$\omega^* = [K^{-1}]_{n\times K}$从而使参数估计失效.

由引理2.1得出定理2.3.

Theorem 2.3 *记Jacobian矩阵$\frac{1}{\beta}\frac{\partial\theta^{(\beta)}(z)}{\partial z}$在点$\omega^* = [K^{-1}]_{n\times K}$的谱半径为$\lambda^*_{\max}$,*

$$\left.\frac{\partial\theta_{ik}^{(j)}}{\dot{\theta}z_{jr}}\right|_{\forall i,k,z_{ik}=K^{-1}} = \beta\frac{\delta_{kr}}{2n}(H)_{ij} = \beta\frac{\delta_{kr}}{2n}\left(A^TA\right)_{ij},$$

$\omega^* = [K^{-1}]_{n\times K}$不是DA-EM算法的渐进稳定点的必要条件为$\beta \geq (\lambda^*_{\max})^{-1}$.

至此, 我们已经给出了DA-EM算法的退火参数β初始值下界的具体求法. 在第三章中, 我们都将使用根据Jacobian矩阵分析所得的退火参数初始值下界来推导有关DA-EM的改进算法, 后面篇幅不再赘述.

2.5 小结

本章研究了高斯混合模型的参数估计问题. 首先, 我们介绍了高斯混合模型的理论. 然后介绍了经典EM算法的推导及其在高斯混合模型参数估计中的应用. 由于EM算法易受初始值的影响而导致算法收敛到局部最优的系统性问题, 我们在第三小节描述了最大熵原理以及确定性退火算法, 并介绍了基于DA-EM算法的高斯混合模型参数估计方法. 而又因为退火参数过小而导致DA-EM算法的初始阶段迭代过早收敛的问题, 我们简要介绍了Yu等(2018)关于DA-EM算法的退火参数初始值的理论求解.

第 3 章　基于改进DA-EM算法的有限混合模型参数估计

本章首先介绍了惩罚似然法和SCAD罚函数. 然后, 在模型个数已知的前提下, 本章通过引入两个惩罚项对DA-EM算法中的对数似然函数进行了修改, 从而改进算法提出了MDA-EM算法, 由于新方法涉及到多元线性回归变量选择的思想, 我们在下文先介绍相关的知识.

3.1　相关知识

3.1.1　惩罚最小距离法

Chen和Kalbfeisch(1996)在AIC(Akaike, 1973), BIC(Schwarz, 1978)等信息准则法思想上, 提出了惩罚最小距离法(penalized minimum-distance criterion). 信息准则法直接对模型参数个数进行惩罚, 而惩罚最小距离法则是对混合比例进行惩罚. 惩罚最小距离法中的距离为

$$D\left(F_n(x), C(x;G)\right) = d\left(F_n(x), C(x;G)\right) - C_K \sum_{k=1}^{K} \log \alpha_k, \tag{3-1}$$

其中

$$C(x;\theta) = \sum_{k=1}^{K} \alpha_k \phi\left(x|\mu_k, \Sigma_k\right),$$

为累计分布函数, 记经验分布函数为$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(x_i \le x\right)$; $d\left(F_n(x), C(x;G)\right)$表示$F_n(x)$和$C(x;\theta)$两个函数之间的距离. 关于$d\left(F_1, F_2\right)$的例子包括:

(1) kolmogrov-Smirnov距离

$$d\left(F_1, F_2\right) = \sup_y \left|F_1(y) - F_2(y)\right|;$$

(2) Cramer-VonMises距离

$$d\left(F_1, F_2\right) = \int \left[F_1(y) - F_2(y)\right]^2 d\left\{F_1(y) + F_2(y)\right\}.$$

由式(3-1)可以看出, 如果α_k过小, 那么该距离函数的值就会很大. 因此, 该方法可以防止EM算法迭代的参数达到参数空间的边界. 一般来说, 各个分模型的混合比例应当不

会很小(对于比例很小的数据可以直接当做异常值来处理). 所以, 对于给定真实的K, 在对数似然函数上的比例上做一定的限制(简称比例罚)可以降低EM算法收敛到局部最优的概率. 然而, 该方法只考虑到了高斯混合模型混合比例的估计, 并没有考虑混合模型的分模型参数(μ, Σ). 因此我们在下一小节中具体研究分模型参数的其他约束方法.

3.1.2 回归模型中的变量选择

考虑线性回归模型

$$Y = X\beta + \varepsilon, \tag{3-2}$$

其中$X \in R^{n\times d}$为设计矩阵, $\beta \in R^d$为回归系数, $Y \in R^n$为相应变量, $\varepsilon \in R^n$为误差向量, 且其变量独立同分布$\varepsilon_i \sim N\left(0, \sigma^2\right), i \in \{1, 2, \cdots, n\}$.

Tibshirani(1996)提出了一种多元线性回归模型的变量选择方法, 即LASSO. 该方法以最小化带罚函数的残差平方和来估计回归系数

$$\widehat{\beta}_{LASSO} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \cdot \|\beta\|_1,$$

其中$\lambda \geq 0$为可调参数.

由于LASSO对所有回归系数都采取了相同程度的惩罚, 导致所估计的参数实际是有偏的. 为此, Fan和Li(2001)提出了SCAD方法, 即

$$\widehat{\beta}_{SCAD} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{d} p_{SCAD}\left(\beta_j\right),$$

其中

$$p'_{SCAD}\left(\beta_j\right) = \gamma I\left(|\beta_j| \leq \gamma\right) + \frac{\left(a\gamma - |\beta_j|\right)_+}{a-1} I\left(|\beta_j| > \gamma\right), \gamma \geq 0, a > 2.$$

Chen和Khalili(2008)提出的MSCAD方法, 把多元线性回归模型中SCAD方法的变量选择思想运用到了混合模型的参数估计和定阶问题中. 通过引入一个关于参数距离的惩罚函数, 对参数相近的模型进行合并, 从而得到一个较低且更接近真实的阶数K. 本文根据MSCAD方法在对数似然函数上引入两个惩罚项, 分别对混合比例和参数之间的距离施加惩罚, 对DA-EM算法进行了改进, 提出了MDA-EM算法.

3.2 MDA-EM算法的理论与数值求解

3.2.1 MDA-EM算法的理论

2.4节中基于Jacobian矩阵分析的DA-EM算法是在不改变DA-EM算法求解过程的基础上, 根据Jacobian矩阵的谱半径来判断DA-EM算法在稳定点$\omega^* = [K^{-1}]_{n\times K}$的收敛性, 同时也给出了退火参数的理论下界. 但定理2.3只给出了避免收敛到稳定点的必要条件, 若直接使用该退火参数的理论下界仍有概率导致算法收敛到稳定点. 另外, 当有多个分模型时($K > 2$), DA-EM算法虽然没有收敛到稳定点, 但是仍会出现多个聚类中心重叠的情形.

Chen和Khalili(2008)提出了改进的SCAD方法(modified SCAD, MSCAD): 通过对混合比例和模型参数之间的距离施加惩罚, 以减少模型的复杂度. 具体地, 对于预先设定的K($K \geq k_0$), 将各高斯模型参数按大小进行排序(这里的参数是各个分模型的均值向量)$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$. 令$\eta_k = \|\mu_{k+1} - \mu_k\|_2, k = 1, 2, \ldots, K-1$. 定义如下惩罚似然函数

$$\tilde{l}_n(\theta) = l_n(\theta) + C_K \sum_{k=1}^{K} \log \alpha_k - \sum_{k=1}^{K-1} p_{SCAD}\left(\eta_k\right), \tag{3-3}$$

其中

$$l_n(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \alpha_k \phi\left(x_i | \mu_k, \Sigma_k\right) \right\},$$

$$p'_{SCAD}(\eta) = \gamma\sqrt{n} I(\sqrt{n}\eta \leq \gamma) + \frac{\sqrt{n}(a\gamma - \sqrt{n}\eta)_+}{a-1} I(\sqrt{n}\eta > \gamma),$$

可调参数$C_K > 0$, $\gamma \geq 0$, $a > 2$. 那么, 在空间

$$\Theta = \left\{ \theta : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_K, \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \geq 0 \right\}$$

上极大化惩罚对数似然函数(3-3), 能同时对高斯混合模型进行定价和模型参数估计. 注意到惩罚项$p_{SCAD}(\cdot)$是由Fan和Li(2001)提出, 旨在解决线性回归分析中变量选择问题.

这里, 我们用一个新的惩罚项$p_{NEW}(\cdot)$来代替$p_{SCAD}(\cdot)$, 即

$$p'_{NEW}(\eta) = -\gamma\sqrt{n} I(\sqrt{n}\eta \leq \gamma) - \frac{\sqrt{n}(a\gamma - \sqrt{n}\eta)_+}{a-1} I(\sqrt{n}\eta > \gamma),$$

新的惩罚项(简称NEW罚)与SCAD罚的不同之处在于: SCAD罚能惩罚模型参数之间的距离较大的参数, 当参数距离减小时, 罚似然函数将会增大. 因此, 在极大化过程中, 间距

很小的参数会趋向一致, 并通过合并参数相同的模型以减小模型个数; NEW罚则刚好相反, 当参数间距减小时, 罚似然函数值将会减小, 从而在极大化过程中间距很小的参数会相互远离彼此.

总结所提出的新方法: 我们定义惩罚对数似然函数

$$\tilde{l}_n(\theta) = l_n(\theta) + C_K \sum_{k=1}^{K} \log \alpha_k - \sum_{k=1}^{K-1} p_{NEW}(\eta_k), \tag{3-4}$$

其中$\log \alpha_k$对混合比例进行约束, $p_{NEW}(\eta_k)$对模型参数之间的距离进行约束. 在DA-EM算法基础上, 修改对数似然函数为式(3-4), 即可实现高斯混合模型的参数估计.

3.2.2 数值求解

下面, 我们对DA-EM算法作适当修改, 以极大化惩罚对数似然函数(3-4).

沿用上文的符号定义, $\theta = ((\alpha_1, \mu_1, \Sigma_1), \cdots, (\alpha_K, \mu_K, \Sigma_K))$. 给定随机独立样本$\{x_1, x_2, \cdots, x_n\}$, 对应的隐变量$\{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$表示每个样本所属的分模型, 即对于$i \in \{1, 2, \cdots, n\}$, 有$\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{iK})^T$, 其中$z_{ik} \in \{0, 1\}$且$\sum_{k=1}^{K} z_{ik} = 1$. 结合观测变量和隐变量生成的完全数据 $\{(x_1, \boldsymbol{z}_1), \cdots, (x_n, \boldsymbol{z}_n)\}$, 得到惩罚完全对数似然函数

$$\tilde{l}_n^c(\theta) = l_n^c(\theta) + C_K \sum_{k=1}^{K} \log \alpha_k - \sum_{k=1}^{K-1} p_{NEW}(\eta_k), \tag{3-5}$$

其中

$$l_n^c(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log \alpha_k + \log \left\{\phi(x_i|\mu_k, \Sigma_k)\right\}\right]. \tag{3-6}$$

E步中, 在给定观测样本$\{x_1, x_2, \cdots, x_n\}$和当前参数估计值$\theta^{(t)}$条件下, 计算$\tilde{l}_n^c(\theta)$关于$z_{ik}$的条件期望, 即

$$\begin{aligned} Q\left(\theta; \theta^{(t)}\right) =& \sum_{i=1}^{n} \sum_{k=1}^{K} E\left[z_{ik}|x_i; \theta^{(t)}\right] \log \left\{\phi(x_i|\mu_k, \Sigma_k)\right\} - \sum_{k=1}^{K-1} p_{NEW}(\eta_k) \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{E\left[z_{ik}|x_i; \theta^{(t)}\right] + \frac{C_K}{n}\right\} \log \alpha_k, \end{aligned} \tag{3-7}$$

其中

$$E\left[z_{ik}|x_i; \theta^{(t)}\right] = \omega_{ik}^{(t+1)} = \frac{\left(\alpha_k^{(t)} \phi\left(x_i|u_k^{(t)}, \Sigma_k^{(t)}\right)\right)^{\beta}}{\sum_{k=1}^{K} \left(\alpha_k^{(t)} \phi\left(x_i|u_k^{(t)}, \Sigma_k^{(t)}\right)\right)^{\beta}}, \tag{3-8}$$

这里β为上文提及的退火参数.

M步为求解$\theta^{(t+1)} = \arg\max_\theta Q\left(\theta;\theta^{(t)}\right)$. 在约束条件$\sum_{k=1}^{K}\alpha_k = 1$下, 采用拉格朗日乘子法, 有

$$\frac{\partial\left[Q\left(\alpha;\alpha^{(t)}\right)+\lambda\left(\sum_{k=1}^{K}\alpha_k-1\right)\right]}{\partial\alpha_k}=0,$$

得到

$$\alpha_k^{(t+1)}=\frac{\sum_{i=1}^{n}E\left[z_{ik}|x_i;\theta^{(t)}\right]+C_K}{n+KC_K},k=1,2,\cdots,K, \tag{3-9}$$

由于$p_{NEW}(\eta)$在$\eta=0$处不可导, 而传统的Newton-Raphson方法要求有一阶和二阶导, 因此不能直接应用到最大化式(3-5)中. 然而, 考虑到$\eta=0$是$p_{NEW}(\eta)$唯一的奇异点, 根据Fan和Li(2001)的建议, 将以局部二次逼近在每轮数值迭代中替换罚函数$p_{NEW}(\eta)$

$$\tilde{p}_{NEW}\left(\eta;\eta_k^{(t)}\right)=p_{NEW}\left(\eta_k^{(t)}\right)+\frac{p'_{NEW}\left(\eta_k^{(t)}\right)}{2\eta_k^{(t)}}\left(\eta^2-\eta_k^{(t)2}\right), \tag{3-10}$$

于是, 对于模型参数μ_k的求解等同于求解以下方程组

$$\begin{cases}\sum_{i=1}^{n}\omega_{i1}\dfrac{\partial}{\partial\mu_1}\log\phi\left(x_i|\mu_1,\Sigma_1\right)-\dfrac{\partial\tilde{p}_{NEW}\left(\eta_1;\eta_1^{(t)}\right)}{\partial\mu_1}=0\\ \sum_{i=1}^{n}\omega_{ik}\dfrac{\partial}{\partial\mu_k}\log\phi\left(x_i|\mu_k,\Sigma_k\right)-\dfrac{\partial\tilde{p}_{NEW}\left(\eta_{k-1};\eta_{k-1}^{(t)}\right)}{\partial\mu_k}-\dfrac{\partial\tilde{p}_{NEW}\left(\eta_k;\eta_k^{(t)}\right)}{\partial\mu_k}=0,\\ k=2,3,\cdots,K-1\\ \sum_{i=1}^{n}\omega_{iK}\dfrac{\partial}{\partial\mu_K}\log\phi\left(x_i|\mu_K,\Sigma_K\right)-\dfrac{\partial\tilde{p}_{NEW}\left(\eta_{K-1};\eta_{K-1}^{(t)}\right)}{\partial\mu_K}=0,\end{cases}$$

将式(3-10)代入上述方程组, 得出

$$\begin{cases} \sum_{i=1}^{n} \omega_{i1} \dfrac{\partial}{\partial \mu_1} \log \phi\left(x_i|\mu_1, \Sigma_1\right) + \dfrac{\partial}{\partial \mu_1} \left\{ \dfrac{p'_{NEW}\left(\eta_1^{(t)}\right)}{2\eta_k^{(t)}} \left(\mu_2 - \mu_1\right)^2 \right\} = 0 \\ \sum_{i=1}^{n} \omega_{ik} \dfrac{\partial}{\partial \mu_k} \log \phi\left(x_i|\mu_k, \Sigma_k\right) - \dfrac{\partial}{\partial \mu_k} \left\{ \dfrac{p'_{NEW}\left(\eta_{k-1}^{(t)}\right)}{2\eta_{k-1}^{(t)}} \left(\mu_k - \mu_{k-1}\right)^2 \right\} \\ + \dfrac{\partial}{\partial \mu_k} \left\{ \dfrac{p'_{NEW}\left(\eta_k^{(t)}\right)}{2\eta_k^{(t)}} \left(\mu_{k+1} - \mu_k\right)^2 \right\} = 0, k = 2, 3, \cdots, K-1 \\ \sum_{i=1}^{n} \omega_{iK} \dfrac{\partial}{\partial \mu_K} \log \phi\left(x_i|\mu_K, \Sigma_K\right) - \dfrac{\partial}{\partial \mu_K} \left\{ \dfrac{p'_{NEW}\left(\eta_{K-1}^{(t)}\right)}{2\eta_{K-1}^{(t)}} \left(\mu_K - \mu_{K-1}\right)^2 \right\} = 0, \end{cases}$$

整理, 可得

$$\begin{cases} \sum_{i=1}^{n} \omega_{i1} \dfrac{\partial}{\partial \mu_1} \log \phi\left(x_i|\mu_1, \Sigma_1\right) - \dfrac{p'_{NEW}\left(\eta_1^{(t)}\right)}{\eta_k^{(t)}} \mu_2 + \dfrac{p'_{NEW}\left(\eta_1^{(t)}\right)}{\eta_k^{(t)}} \mu_1 = 0 \\ \sum_{i=1}^{n} \omega_{ik} \dfrac{\partial}{\partial \mu_k} \log \phi\left(x_i|\mu_k, \Sigma_k\right) - \dfrac{p'_{NEW}\left(\eta_{k-1}^{(t)}\right)}{\eta_{k-1}^{(t)}} \mu_k + \dfrac{p'_{NEW}\left(\eta_{k-1}^{(t)}\right)}{\eta_{k-1}^{(t)}} \mu_{k-1} \\ - \dfrac{p'_{NEW}\left(\eta_k^{(t)}\right)}{\eta_k^{(t)}} \mu_{k+1} + \dfrac{p'_{NEW}\left(\eta_k^{(t)}\right)}{\eta_k^{(t)}} \mu_k = 0, k = 2, 3, \cdots, K-1 \\ \sum_{i=1}^{n} \omega_{iK} \dfrac{\partial}{\partial \mu_K} \log \phi\left(x_i|\mu_K, \Sigma_K\right) - \dfrac{p'_{NEW}\left(\eta_{K-1}^{(t)}\right)}{\eta_{K-1}^{(t)}} \mu_K + \dfrac{p'_{NEW}\left(\eta_{K-1}^{(t)}\right)}{\eta_{K-1}^{(t)}} \mu_{K-1} = 0. \end{cases} \tag{3-11}$$

以上方程组包含K个方程, 可求解K个模型参数$\mu_k^{(t+1)}, k = 1, 2, \cdots, K$. $\Sigma_k^{(t+1)}$的求解与式(2-21)的相同

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} \omega_{ik}^{(t+1)} \left(x_i - \mu_k^{(t+1)}\right) \left(x_i - \mu_k^{(t+1)}\right)^T}{\sum_{i=1}^{n} \omega_{ik}^{(t+1)}}. \tag{3-12}$$

下面我们给出MDA-EM的算法流程

在使用所提出的方法来解决参数估计问题之前, 我们首先需要选择可调参数γ和C_K. Chen和Kalbfleisch(2001)建议如果参数μ_K在区间$[-M, M]$或$[M^{-1}, M]$(M为一个较大的正数)内, 一个合适的取值为$C_K = \log M$. 这里, 我们参考Xu和Chen(2015)的做法, 令$M = \max\left\{\|x_1\|_2, \|x_2\|_2, \cdots, \|x_n\|_2\right\}$, 其中$\|x_i\|_2$表示$x_i$的二范数. 关于$p_{NEW}(\eta)$罚函数

Algorithm 2 基于MDA-EM框架的高斯混合模型参数估计算法

Input: 观测数据$X = \{x_1, \ldots, x_n\}$, 高斯混合模型, $\epsilon_1, \epsilon_2 > 0$, 初始退火参数$\beta \leftarrow \beta^0$, $t = 0$

Output: 高斯混合模型参数$\theta = ((\alpha_1, \mu_1, \Sigma_1), \cdots, (\alpha_K, \mu_K, \Sigma_K))$

1: 参数初始值$\theta^{(0)} = \left(\left(\alpha_1^{(0)}, \mu_1^{(0)}, \Sigma_1^{(0)}\right), \cdots, \left(\alpha_K^{(0)}, \mu_K^{(0)}, \Sigma_K^{(0)}\right)\right)$

2: **while** $\beta \leq 1$ **do**

3: 根据式(2-15)计算后验概率ω_{ik}, 即每个分模型对观测数据的响应度;

4: 根据式(3-9), (3-11), (3-12)计算当前混合模型的参数;

5: 判断两次迭代的最大似然估计参数不在变化, $\left\|\Theta^{(t+1)} - \Theta^{(t)}\right\| < \epsilon_1$, 或者Q函数不再变化, $\left\|Q\left(\theta^{(t+1)}, \theta^{(t)}\right) - Q\left(\theta^{(t)}, \theta^{(t)}\right)\right\| < \epsilon_2$. 否则, $t = t + 1$, 重复步骤3,4;

6: **end while**

中的γ和a. 前者我们采用10-折交叉验证选出最佳参数, 而后者可以根据Brehent和Huang(2011)的建议, 令a=3.

3.3 小结

本章研究了基于改进DA-EM算法的高斯混合模型参数估计问题. 在第一小节中, 我们介绍了与罚似然方法相关的两类方法, 一类是基于信息准则思想而提出的惩罚最小距离法, 另一类是旨在解决多元线性回归变量选择问题的方法LASSO和SCAD. 在第二小节, 因为Yu等(2018)给出关于退火参数的理论下界并不能很好地避免多个聚类中心重叠的情况, 本文根据MSCAD方法在似然函数上对参数间的距离和混合比例施加惩罚的思想, 对DA-EM算法进行了改进, 给出了MDA-EM算法和数值求解过程.

第 4 章　数值实验

本章通过模拟数据和真实数据相结合的方式来验证MDA-EM算法的参数估计效果, 主要采用聚类准确率指标. 需要说明的是, 聚类效果的判断指标也是参差不齐的. 针对同一聚类结果, 不同的指标可能得到不同的判断结果, 并且目前学术界没有公认的统一判断标准. 由于文章中所有模拟和实证使用的样本数据都带有真实类别标签, 因此可以将不同算法得到的样本聚类标签与样本实际标签对比, 计算出算法准确度, 从而较为客观地衡量各个方法的估计效果, 考察各算法参数估计的有效性. 为了减少聚类指标的差异对聚类效果判断的影响, 本文将不含主观因素的聚类准确度作为聚类效果的主要指标. 显然, 聚类的准确度越高, 该算法的参数估计则越有效. 此外, 由于数值模拟和实证分析可以获得参数的真值, 因此本文在该部分还将给出参数真值与参数估计值之间的距离(MSE), 来更好地评估算法性能.

4.1　模拟研究

参考Yu等(2018)的做法,在空间$\mathbb{R}^2$上, 本节通过数值模拟的方式产生了满足以下高斯混合模型的300个模拟数据

$$\begin{aligned} p(x|\theta) &= \sum_{k=1}^{K} \alpha_k \cdot \phi\left(x|\mu_k, \Sigma_k\right) \\ &= \sum_{k=1}^{K} \alpha_k \cdot \frac{1}{(2\pi)^{2/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right), \end{aligned} \tag{4-1}$$

其中

$$\begin{gathered} K = 3 \\ \alpha_1 = 1/3 \quad , \quad \alpha_2 = 1/3 \quad , \quad \alpha_3 = 1/3 \\ \mu_1^T = (-1.3, 0) \quad , \quad \mu_2^T = (0, 0) \quad , \quad \mu_3^T = (1.3, 0) \\ \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix} \end{gathered} \tag{4-2}$$

同时保留每个模拟数据所属的真实分模型(高斯混合成分), 也就是它们的真实类别标记.

下面, 分别使用classical EM(CEM)算法、deterministic annealing EM(DA-EM)算法、以及modified deterministic annealing EM(MDA-EM)算法模拟产生的仿真数据(不含真实标签)估计高斯混合模型的参数, 为了更好地对比三种算法的聚类效果, 我们

用同一聚类中心对算法进行初始化: $\alpha_1^{(0)} = \alpha_2^{(0)} = \alpha_3^{(0)}$, $\mu_1^{(0)} = (-1,0)^T$, $\mu_2^{(0)} = (0,0)^T$, $\mu_3^{(0)} = (1,0)^T$, $\Sigma_1^{(0)} = \Sigma_2^{(0)} = \Sigma_3^{(0)} = I$, 计算不同参数估计值对应的聚类准确率以及其与参数真值的距离, 评估参数估计的准确性.

由于算法的初始聚类中心是从样本随机挑选的, 为了使模拟研究的结果更有说服力, 我们进行了50次蒙特卡洛模拟. 下面给出一系列图表作为解释, 其中我们用红色、绿色、蓝色对应表示CEM算法、DA-EM算法、MDA-EM算法的估计效果.

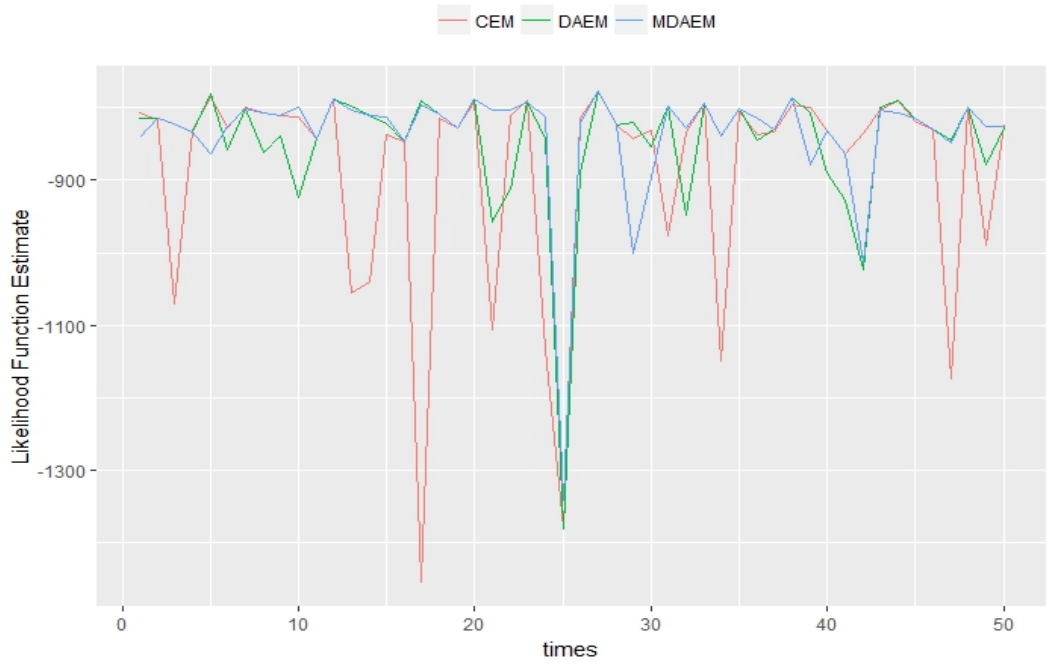

图 4.1: 聚类准确度

首先,我们用红色实线, 绿色实线、蓝色实线分别表示三种算法CEM算法、DA-EM算法、MDA-EM算法的参数估计效果. 从图4.1可以看出CEM算法大概有1/3的实验聚类准确度低于80%, DA-EM算法大约有1/5的的实验聚类准确度低于80%, 而代表MDA-EM算法的蓝色实线基本都在其余两条线之上, 只有6次(大约1/10)的聚类准确度低于80%.

对比图4.2可以看出, 对数似然函数和聚类准确度之间有较强的相关性: 对这三类算法而言, 聚类准确度高的模拟次数对应的对数似然函数值相应较高, 而聚类准确度低的模拟次数对应的对数似然函数值也明显偏低. 就拿CEM算法来说, 第3,13,14,17组的聚类准确度偏低, 所对应的对数似然函数值也明显比其余二者低, 从而从另一个角度体现三种算法在当前模拟条件下参数估计效果的优劣.

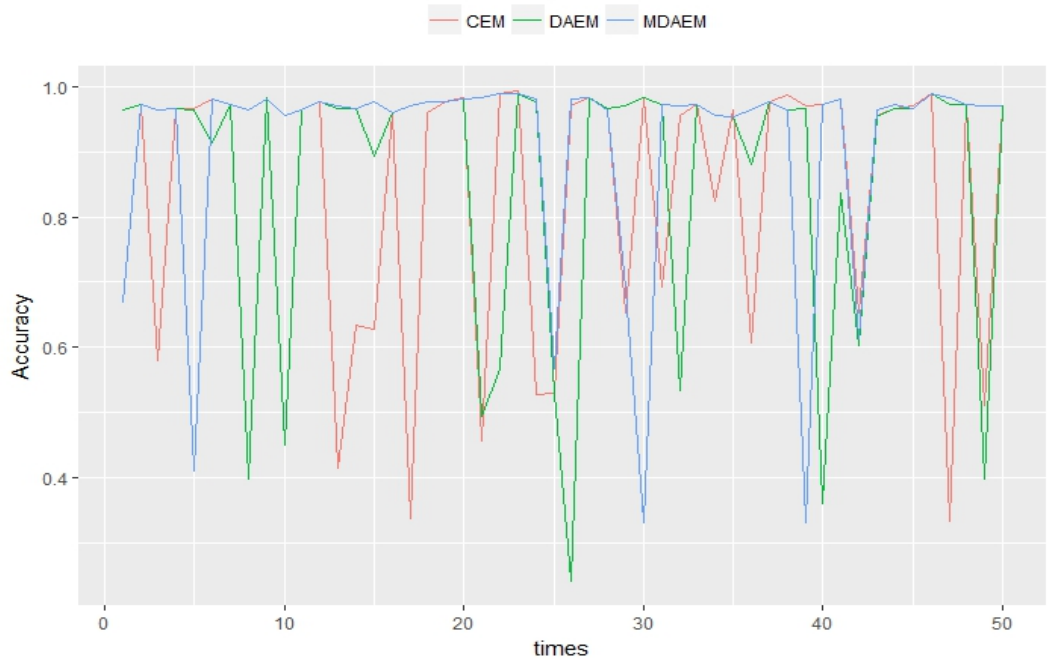

图 4.2: 对数似然函数值

对于参数估计效果, 本文给出另一个评价指标, 模拟所得的三种算法参数估计值与样本参数真值的差距:

(1)关于高斯混合模型的比例估计. 如图4.3所示, 在对应混合比例的三个箱线图上, 我们可以明显看出异常值的个数CEM>DA-EM>MDA-EM, 上下四分位之差也是CEM>DA-EM>MDA-EM. 这些都表明了在混合比例的估计效果上, DA-EM算法估计参数的稳定性比CEM算法好, 而本文提出的MDA-EM算法又优于DA-EM算法.

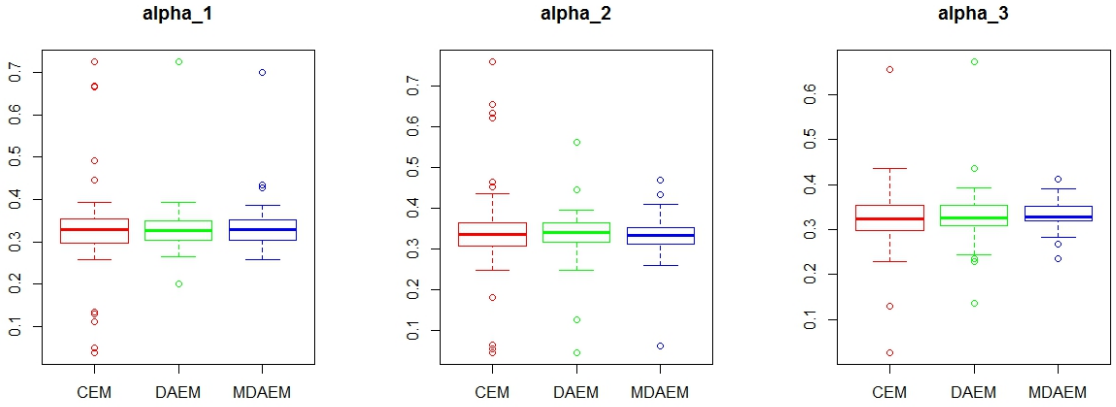

图 4.3: 混合比例箱线图

(2)关于高斯混合模型的均值向量估计, 如图4.4所示为各算法估计的均值向量与模拟均值向量真值之差的2范数, 总体来说DA-EM 算法的均值向量估计是要优于CEM,

而MDA-EM算法的均值估计又是优于DA-EM算法的.

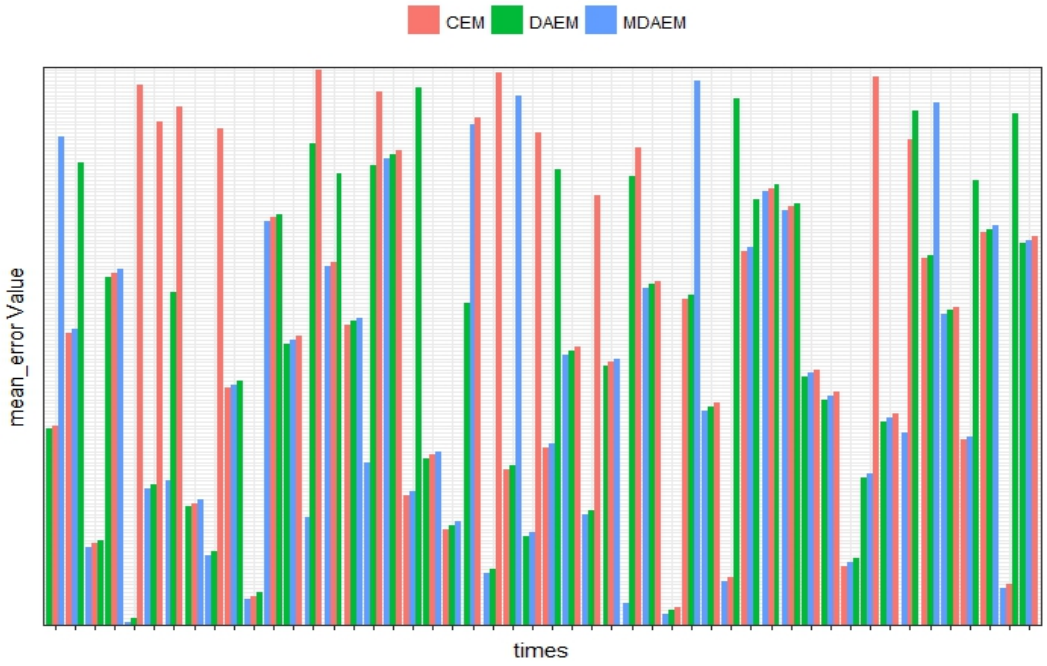

图 4.4: 均值向量估计误差和的2范数

(3)关于高斯混合模型的协方差矩阵估计, 如图4.5所示为各算法估计的协方差矩阵与模拟协方差矩阵真值之差的F范数, 其中F范数是把一个矩阵中每个元素的平方求和开根号. 与均值向量相同, 我们得出结论: 三种算法的协方差矩阵估计效果为MDA-EM>DA-EM>CEM.

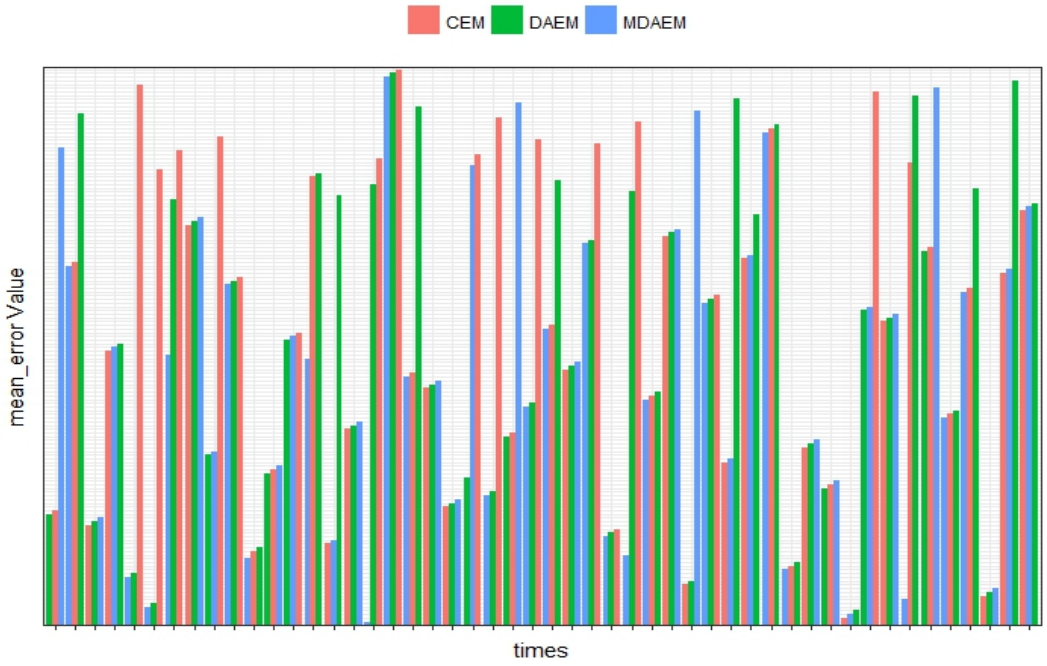

图 4.5: 协方差矩阵估计误差的F范数

表 4.1: 高斯混合模型仿真研究中算法的收敛性和参数估计精度.

	CEM	DA-EM	MDA-EM
ACCURACY	0.8482(0.2049)	0.8609(0.2111)	**0.9085(0.1695)**
STEP	**258.1(198.1)**	1117.6(656.1)	1051.0(2074.5)
INP	-887.2	-847.9	**-837.1**
AIC	-907.3	-867.9	**-857.1**
BIC	-944.3	-904.9	**-894.2**
$MSE(\hat{\alpha})$	0.1297	0.0728	**0.0592**
$MSE(\hat{\mu})$	0.6314	0.4788	**0.3751**
$MSE(\hat{\Sigma})$	0.6167	0.5935	**0.4621**

注释: 括号前的数字表示50次模拟的参数平均值, 括号中的数字表示相应的方差. $MSE(\hat{\theta})$记为各模型参数估计值与参数真值的均方误差. 加粗的数字表明所对应列的算法估计效果最佳.

表4.1的第一列为三种模型的各个评价指标, 依次为: 模型准确度、迭代次数、对数似然函数值、AIC、BIC、$\hat{\alpha}$的均方误差、$\hat{\mu}$的均方误差、$\hat{\Sigma}$的均方误差. 在这50次模拟中, 对于各个指标我们展开详细讨论:

(1) 聚类准确度方面, MDA-EM算法的平均聚类准确度最高, DA-EM算法次之, CEM算法则表现最差.除此之外, MDA-EM算法所得准确度的标准差最小, 所求的估计参数也最稳定;

(2) 迭代次数方面, CEM算法毫无疑问是有优势的, 因为不管是DA-EM算法还是MDA-EM算法, 都是在某个退火参数下进行迭代到收敛, 而每有一次退火参数的更新都会让EM算法进行迭代到收敛. 从某种意义上说, CEM算法与后两者算法之间是没有可比性的. 若是只有后两者间作对比, MDA-EM的平均迭代次数更少一些. 由于在对数似然函数上引入了惩罚项, 个别情况下的迭代次数不稳定, 使得迭代次数的标准差变得很大. 综合来说, MDA-EM算法在迭代次数上优于DA-EM算法.

(3) 对数似然函数值和AIC、BIC等信息准则方面, MDA-EM算法具有明显的优势.

(4) 参数的均方误差方面, MDA-EM算法所估计的参数相比其余二者也是更接近真值的.

模拟数据的真实聚类结果见图4.6.

图4.6中方形点表示第1类样本数据的真实标签, 圆形点表示第2类样本数据的真实标签, 三角形点表示第3类样本数据的真实标签; 蓝色表示聚类后的第1类标签, 红色表示聚类后的第2类标签, 绿色表示聚类后的第3类标签. 而黑色大方形点、黑色大圆形点、黑色大三角形点分别表示第1,2,3类样本数据的真实均值向量. 图中X轴表示样本数据的

第1维度, Y轴表示第2维度.

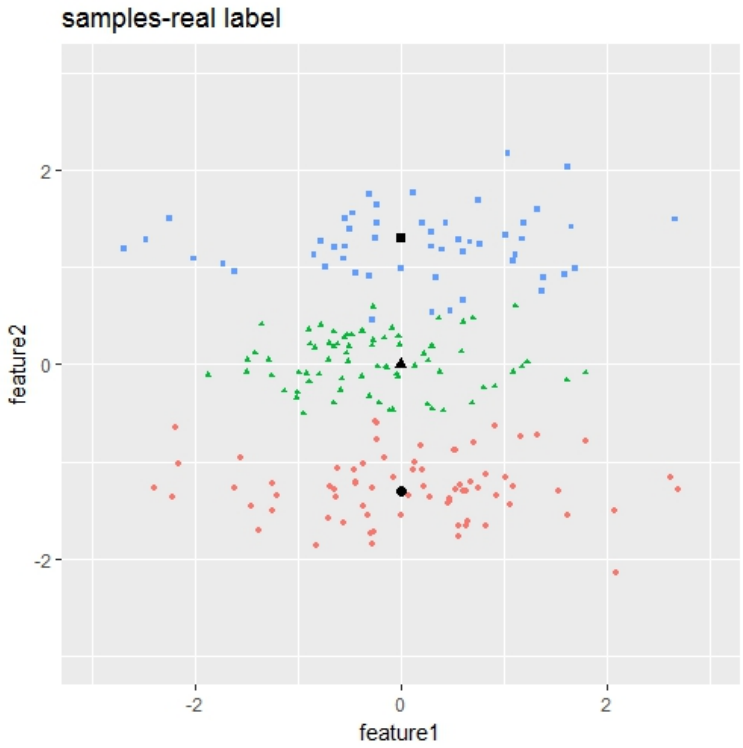

图 4.6: 模拟数据的真实聚类结果

随机选取三种模型参数估计效果存在差异性的三组模拟实验, 并对其做详细分析. 这三组模拟实验对应的聚类准确度, 对数似然函数值见表4.2.

表 4.2: 三组模拟实验对应的各指标值.

实验组号	第14组	第22组	第49组
ACCURACY(CEM)	0.6333	0.9900	0.5100
ACCURACY(DA-EM)	0.9667	0.5667	0.3967
ACCURACY(MDA-EM)	0.9767	0.9900	0.9700
Inp(CEM)	-1040.5	-810.8	-810.8
Inp(DA-EM)	-811.3	-911.7	-803.0
Inp(MDA-EM)	-990.0	-878.0	-826.0

下面, 我们给出这三组模拟实验中三组算法的参数估计对应的聚类效果, 并分别对其做详细分析.

图4.7, 4.8, 4.9展示了第14组模拟实验的聚类效果. 结合表4.2可以看出, CEM算法在该次模拟实验中陷入到了局部最大值点, 并且对数似然函数值相比另外两者低25%, 其

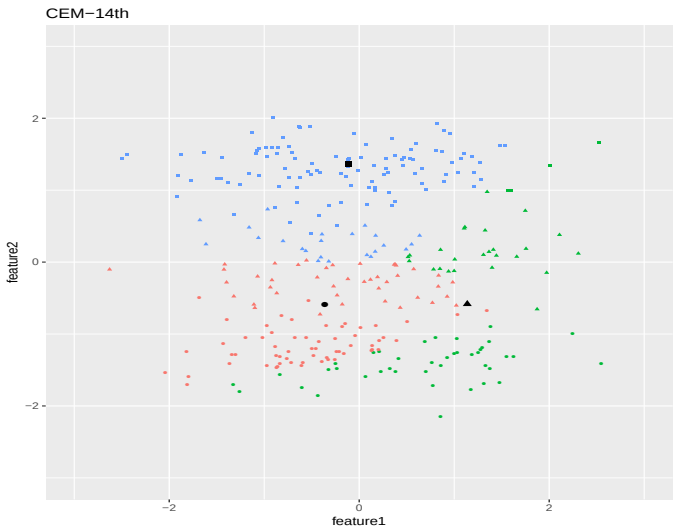

图 4.7: CEM算法对应的聚类效果(14组)

聚类效果也很一般. 而MDA-EM算法则继承了DA-EM算法的全局收敛能力, 并且在对数似然函数值上有微弱的优势.

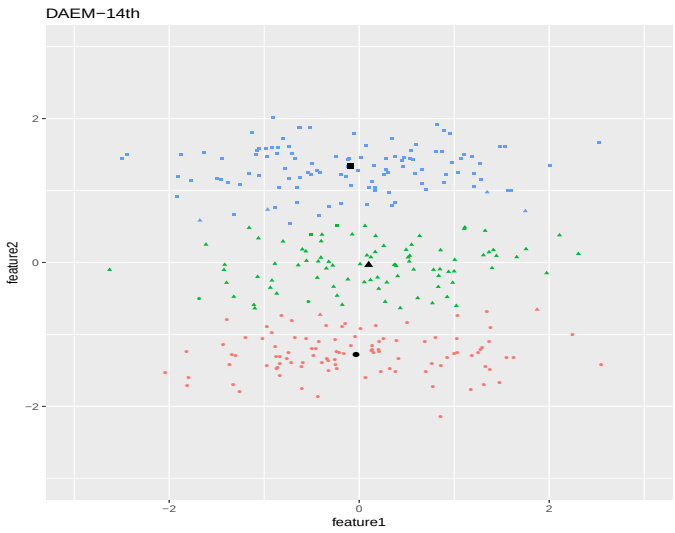

图 4.8: DA-EM算法对应的聚类效果(14组)

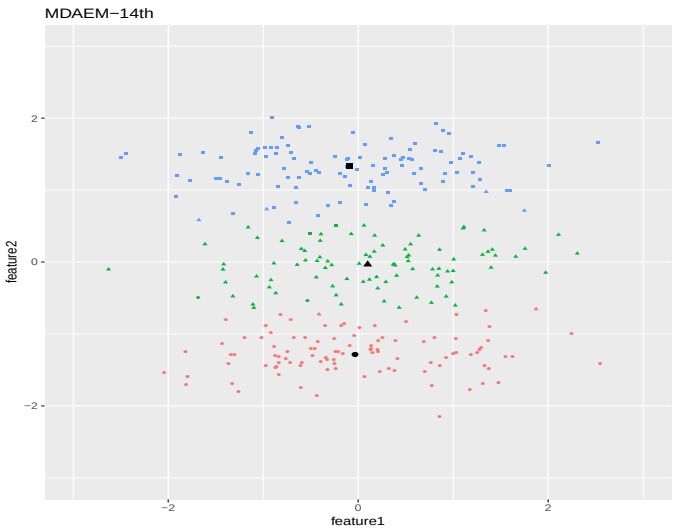

图 4.9: MDA-EM算法对应的聚类效果(14组)

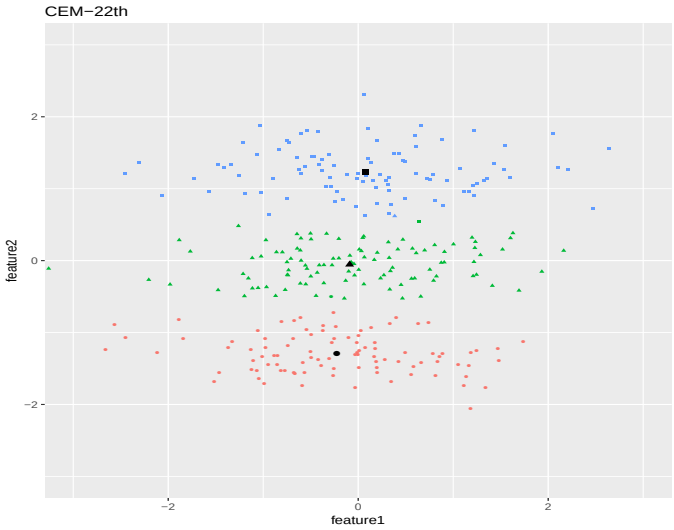

图 4.10: CEM算法对应的聚类效果(22组)

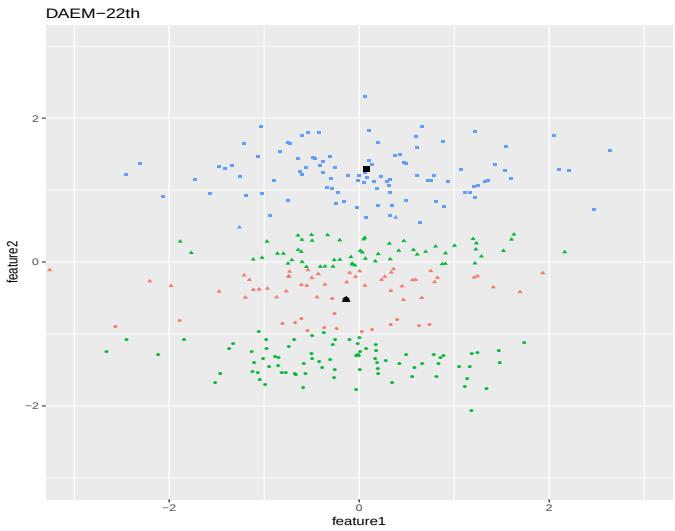

图 4.11: DA-EM算法对应的聚类效果(22组)

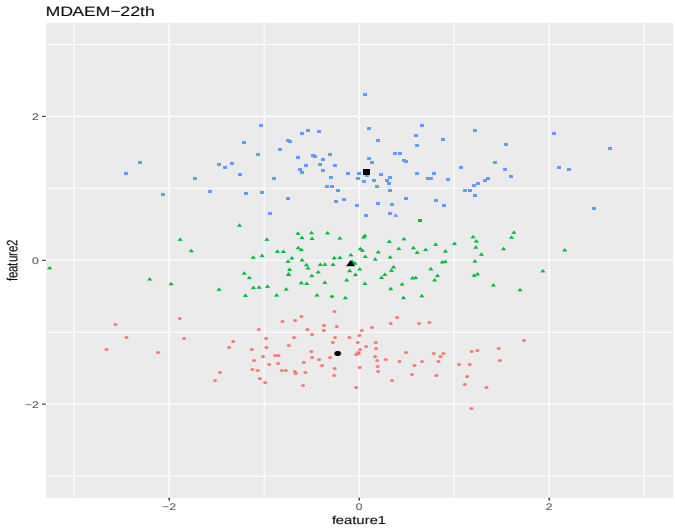

图 4.12: MDA-EM算法对应的聚类效果(22组)

图4.10, 4.11, 4.12为第22组模拟实验的聚类效果. 在第22次模拟中, CEM达到了完美的聚类效果, 但是图4.11 却出现两个聚类中心重叠的现象(红色样本聚类中心和绿色样本聚类中心重合), 导致聚类准确度下降. 由此可以看出, 使用退火参数理论下界的DA-EM算法虽然可以避免算法收敛到稳定点$\omega^* = [K^{-1}]_{n\times K}$, 但仍不能避免存在两个或多个聚类中心聚集的现象. 然而, MDA-EM算法在第22次模拟中仍然能较好的聚类, 其原因上文已经提及不再赘述.

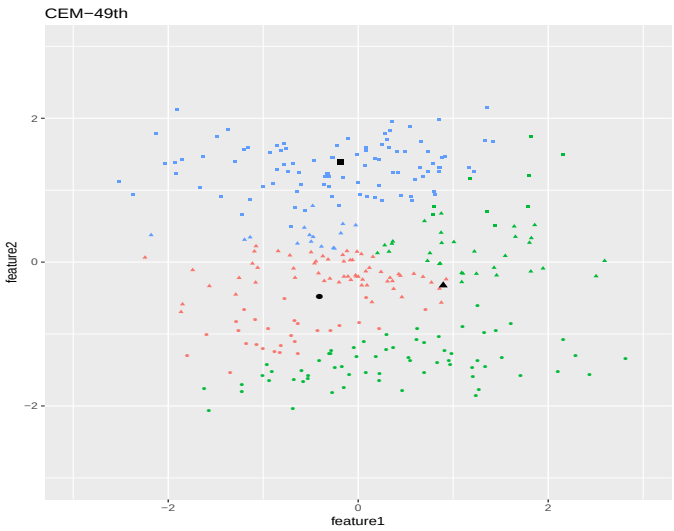

图 4.13: CEM算法对应的聚类效果(49组)

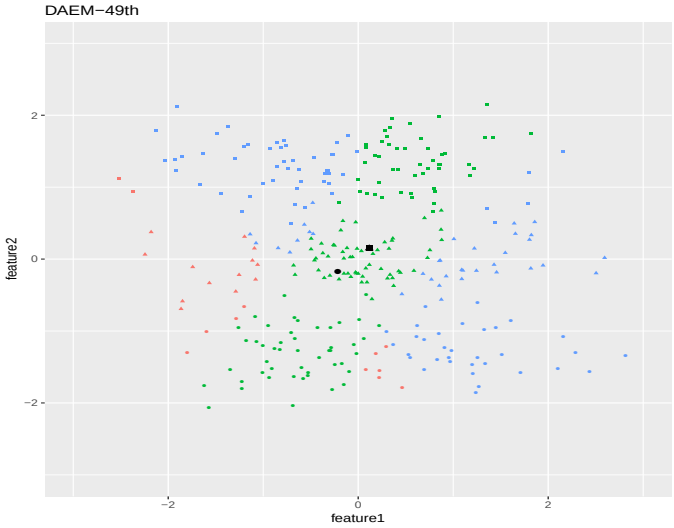

图 4.14: DA-EM算法对应的聚类效果(49组)

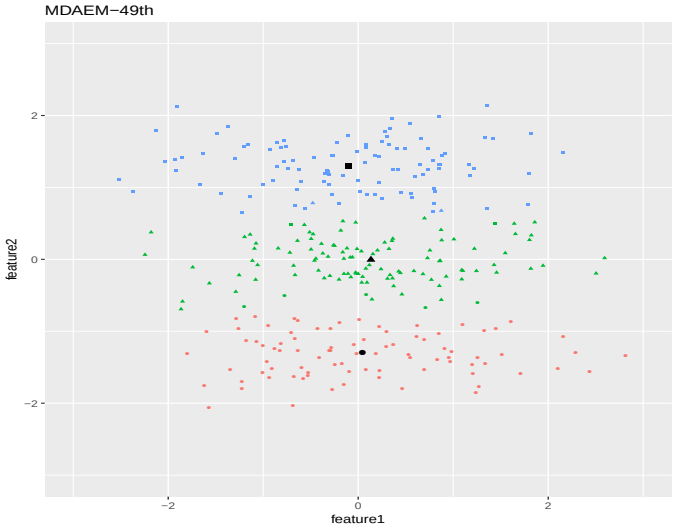

图 4.15: MDA-EM算法对应的聚类效果(49组)

图4.13-4.14以及4.15为第49组模拟实验的聚类效果. 在CEM算法和DA-EM算法的聚类效果都不理想的情形下, 本文提出的MDA-EM算法仍能保持较高的聚类准确度.

结合表4.1-4.2的统计数据, 我们可以得出: 在随机初始参数条件下, CEM算法容易收敛到局部最优, 虽然DA-EM算法的综合性能优于CEM算法. 但又因为其退火参数下界理论的局限性而不能完全避免两个或多个聚类中心重叠的情况. 而本文提出的MDA-EM算法在罚函数的作用下能使得迭代过程能避开聚类中心重叠的情况, 相比于DA-EM算法, 大幅增加了算法收敛到全局最优的概率. 但从本质上讲, MDA-EM算法还是确定性退火EM算法, 若对数似然函数的两个峰的距离过大时, 和DA-EM算法相同, MDA-EM算法仍然只能收敛到局部最优而非全局最优.

下面, 我们使用与上文相同的条件, 只令样本数据$n = 400$, $n = 200$情况下做同样的50次模拟实验, 得到表4.3和表4.4.

我们结合表4.1-4.3以及4.4可以得出: 当样本数据较为稠密时, 三种算法的估计精度差距不大并且都比较高, 这是因为使用越多的数据, 获得的样本信息就越全面, 个别离群样本对总体的影响也就越小, EM算法同时也越不容易收敛到局部极值; 当样本数据较为稀疏时, 三种算法估计精度的差距较大, 同样, 使用越少的数据, 获得的样本信息就越少, 离群样本对总体的影响就越大, EM算法也就容易收敛到局部极值. 值得一提的是, 在样本个数$n = 400$和$n = 200$的情形下, MDA-EM算法达到收敛的迭代次数大约只有DA-EM算法的60%, 三个模拟实验在迭代次数均值上都优于DA-EM算法. 在DA-EM算法基础上, 改进的算法能加快迭代的速率.

表 4.3: 高斯混合模型仿真研究中算法的收敛性和参数估计精度(n=400).

	CEM	DA-EM	MDA-EM
ACCURACY	0.8953(0.1791)	0.8966(0.1795)	**0.9072(0.1726)**
STEP	**243.0(166.1)**	1134.6(781.4)	585.4(632.4)
INP	-1127.3	-1132.2	**-1102.4**
AIC	-1147.3	-1152.2	**-1122.4**
BIC	-1187.3	-1192.1	**-1162.3**
$MSE(\hat{\alpha})$	0.0887	0.0679	**0.0499**
$MSE(\hat{\mu})$	0.4511	0.3818	**0.3747**
$MSE(\hat{\Sigma})$	0.4295	**0.3795**	0.4260

表 4.4: 高斯混合模型仿真研究中算法的收敛性和参数估计精度(n=200).

	CEM	DA-EM	MDA-EM
ACCURACY	0.75(0.2483)	0.8154(0.2099)	**0.8764(0.1724)**
STEP	**258.9(179.9)**	1246.1(737.9)	716.0(909.7)
INP	-607.3	-613.3	**-567.3**
AIC	-627.3	-633.3	**-587.3**
BIC	-660.3	-666.2	**-620.3**
$MSE(\hat{\alpha})$	0.2031	0.1495	**0.1212**
$MSE(\hat{\mu})$	0.9575	0.7046	**0.5156**
$MSE(\hat{\Sigma})$	0.8562	0.7533	**0.6240**

4.2 实证分析

本节使用著名UCI公开数据集iris进行实证分析. 为了方便图像的显示, 我们只采用该数据集的第2维度和第4维度来做三种算法的聚类分析. 同样地, 我们先建立高斯混合模型

$$\begin{aligned} p(x|\theta) &= \sum_{k=1}^{K} \alpha_k \cdot \phi\left(x|\mu_k, \Sigma_k\right) \\ &= \sum_{k=1}^{K} \alpha_k \cdot \frac{1}{(2\pi)^{2/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right), \end{aligned} \tag{4-3}$$

$$K = 3$$

$$\alpha_1 = 1/3 \quad , \quad \alpha_2 = 1/3 \quad , \quad \alpha_3 = 1/3$$

$$\Sigma_1 = \begin{pmatrix} 0.1437 & 0.0093 \\ 0.0093 & 0.0111 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.0985 & 0.0412 \\ 0.0412 & 0.0391 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.1040 & 0.0476 \\ 0.0476 & 0.0754 \end{pmatrix} \tag{4-4}$$

其中, $\Sigma_k, k = 1, 2, 3$为三类样本的真实协方差, 现在令$\mu_k, k = 1, 2, 3$为随机抽取的三个样本的值(初始聚类中心随机化). 重复操作20次, 从而产生20组实验数据. 分别使用三种算法估计高斯混合模型的参数, 计算参数估计值对应的聚类准确度、对数似然值、各参数的均方误差等, 以此评价三种算法的聚类效果及参数估计的有效性. 我们设定迭代收敛条件为$\left\|\Theta^{(t+1)} - \Theta^{(t)}\right\| < 10^{-5}$.

同样, 我们给出20次随机初始聚类中心的聚类效果表4.5.

表 4.5: 关于iris数据集的参数估计效果.

	CEM	DA-EM	MDA-EM
ACCURACY	0.8422(0.2483)	0.9272(0.2099)	**0.9467(0.1724)**
STEP	**99.7(43.3)**	1731.3(580.8)	1202.6(815.7)
INP	-127.3	-124.2	**-123.9**
AIC	-147.3	-144.2	**-143.9**
BIC	-177.4	-174.3	**-174.0**
$MSE(\hat{\alpha})$	0.3483	0.3154	**0.2297**
$MSE(\hat{\mu})$	0.4591	0.4386	**0.3121**
$MSE(\hat{\Sigma})$	**0.0990**	0.1346	0.1129

从表4.5中可以看出, 对于给定随机的初始聚类中心, MDA-EM算法总体表现良好, 估计效果较稳定, 得到的参数估计值也较为准确. 但在少数实验中, MDA-EM算法同样没能避免收敛到局部最优. 但即使MDA-EM算收敛到局部最优, 其所得的参数估计效果也仍然好于CEM算法和DA-EM算法. 相对而言, CEM算法表现平庸, 估计效果不太稳定, 但由于平均迭代次数远远少于其余两种算法, 因此也具有一定的实用性.

随机挑选一次聚类效果较差的实验并将它们的聚类结果与真实聚类做对比, 如图4.16所示.

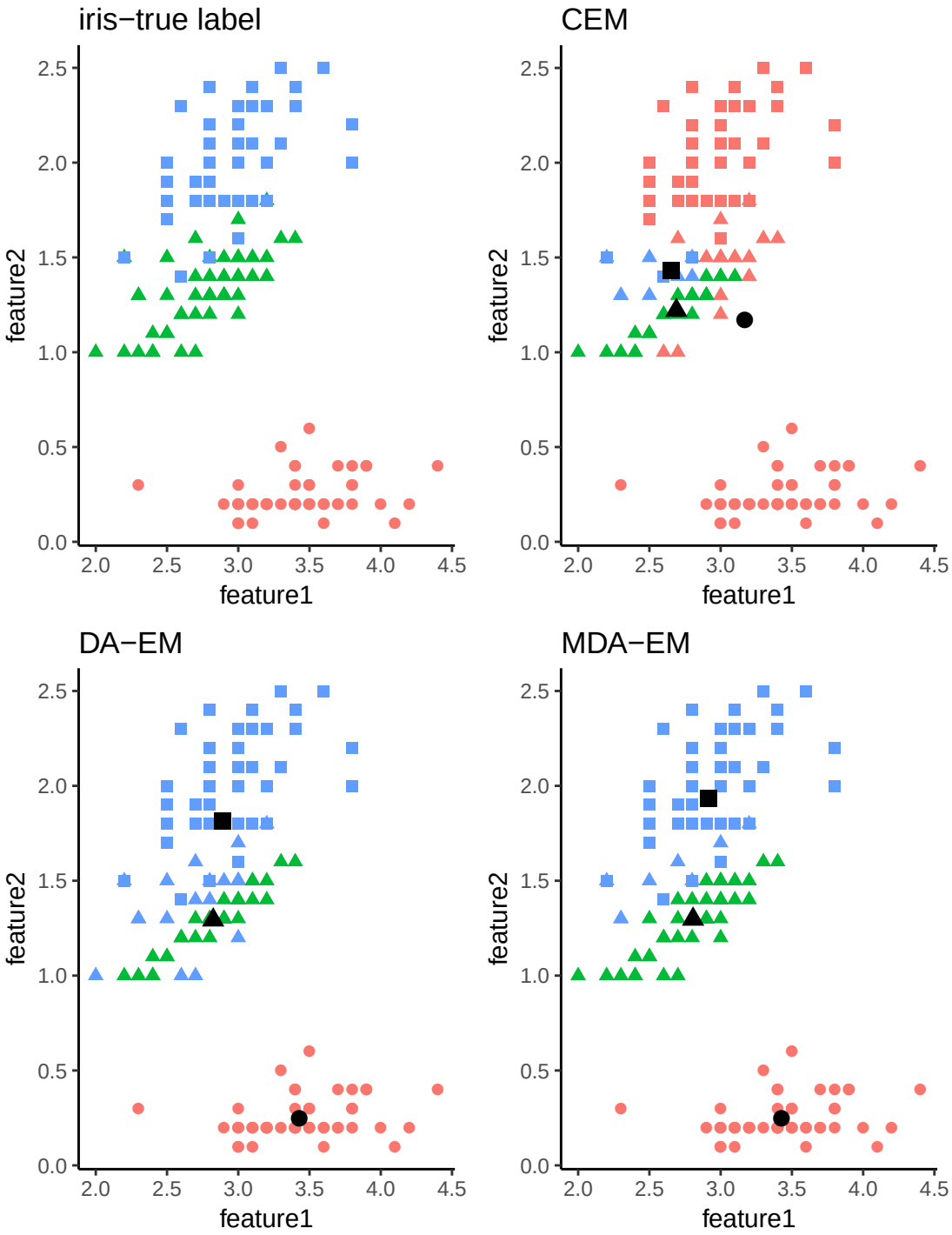

图 4.16: 第11次实验的算法聚类效果与真实聚类

第 5 章　总结与展望

5.1　总结

有限混合模型做为一种解决复杂分布的工具, 可以逼近几乎所有的概率密度, 因此它在理论和应用上均被广泛研究. 而作为自然界最广泛存在的有限混合模型: 高斯混合模型. 作为实际操作中基础又重要的工作, 本文主要对高斯混合模型的参数估计问题进行研究.

由于数据中隐变量的存在使得EM算法是解决此类问题的首要选择, 但是EM算法存在许多局限性. 相较于经典EM算法, DA-EM算法引入了退火参数, 解决了隐变量的后验概率估计过度依赖参数初始值的问题. 然而,使用过小的退火参数初始值将导致DA-EM算法收敛到稳定点而使迭代早停.

本文在Yu等(2018)关于求解退火参数下界的理论基础上, 本文提出了一种改进的DA-EM算法. 在给定模型个数的条件下, MDA-EM算法对参数估计的准确性和稳定性更高, 模拟实验也证明了新方法能有效避免多个高斯混合模型均值向量重叠的情形. 但是该方法的数值求解过程较为复杂, 且其渐进性有待证明, 这些都是值得进一步研究的问题.

5.2　未来的工作

下面展开几个拓展性问题, 即其他的有限混合模型参数估计, 不同成分的多元高斯混合模型及算法优化问题.

5.2.1　其他的有限混合模型参数估计及不同成分的混合模型

由于文章篇幅的局限性, 本文主要研究的混合模型是高斯混合模型, 然而泊松混合模型、二项混合模型也是非常具有研究价值的, 第三章所提出的改进算法的思想依旧是可以应用到这些模型当中的.

此外, 本文所研究的混合模型都是同质性混合模型, 即所有的分模型都是属于一类(高斯模型)的. 更为一般化的问题, 是研究不同成分的混合模型, 至于如何应用是我们未来将进行的工作.

5.2.2　算法优化问题

最值问题一直是被广泛研究的经典问答. 文中第二章提出了一种改进的DA-EM算

法, 新算法加快了收敛速度, 同时在参数估计上比DA-EM 算法更加准确及稳定. 然而, 与DA-EM算法相同, 当对数似然函数的两个极值点相距过远时, 该算法依旧只能收敛到局部最大值, 而不能保证获得全局最大值. 第三章通过修改DA-EM算法最大惩罚对数似然函数, 该算法一方面需要通过交叉验证进行调参, 另一方面在每次迭代的M步都需要求解方程组, 而在迭代的过程中很容易出现矩阵奇异的情形. 前两者都要消耗大量的计算成本. 同样的, 该算法无法保证收敛到目标函数的全局最大值. 因此, 相关算法的优化问题将是我们未来的工作之一.

参考文献

[1] 陈希孺, 倪国熙. 数理统计学教程[M]. 合肥: 中国科学技术大学出版社, 2009.

[2] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[3] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[4] Akaike H. Information theory and an extention of the maximum likelihood principle[C]. 2nd International Symposium on Information Theory, 1973:267-281.

[5] Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection[J]. Annals of Applied Statistics, 2011, 5(1):232-253.

[6] Chen H, Chen J, Kalbfleisch J D. A modifed likelihood ratio test for homogeneity in finite mixture models[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001, 63(1):19-29.

[7] Chen H, Chen J, Kalbfleisch J D. Testing for a finite mixture model with two components[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2004, 66(1):95-115.

[8] Chen H, Chen J. Large sample distribution of the likelihood ratio test for normal mixtures[J]. Statistics & Probability Letters, 2001, 52(2):125-133.

[9] Chen H, Chen J. The likelihood ratio test for homogeneity in finite mixture models[J]. Canadian Journal of Statistics, 2001, 29(2):201-215.

[10] Chen J, Kalbfleisch J D. Penalized minimum distance estimates in finite mixture models[J]. Canadian Journal of Statistics, 1996, 24(2):167-175.

[11] Chen J, Khalili A. Order selection in finite mixture models with a nonsmooth penalty[J]. Journal of the American Statistical Association, 2008, 103(484):1674-1683.

[12] Chen J, Li P, Fu Y. Inference on the order of a normal mixture[J]. Journal of the American Statistical Association, 2012, 107(499):1096-1105.

[13] Chen J. On finite mixture models[J]. Statistical Theory and Related Fields, 2017, 1(1):15-27.

[14] Cheng X X. Convergence of the EM algorithm[J]. Beijing Daxue Xuebao , 1987,(3):1-8.

[15] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the loyal Statistical Society, 1977: 1-38.

[16] Fan J, Li R. Variable selection via nonconvave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456):1348-1360.

[17] Figueiredo M A T, Jain A K. Unsupervised learning of finite mixture models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3):381-396.

[18] Fu Y, Chen J, Li P. Modifed likelihood ratio test for homogeneity in a mixture of von Mises distributions[J]. Journal of Statistical Planning and Inference, 2008, 138(3):667-681.

[19] Ghosh J K, Sen P K. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results[C]. Proceedings of the Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer, Volume 2. eds L. LeCam and Richard A. Olshen, 1985:789 – 806.

[20] Grimm K J, Mazza G L, Davoudzadeh P. Model selection in finite mixture models: a k-fold cross-validation approach[J]. Structural Equation Modeling A Multidisciplinary Journal, 2017, 24(2):1-11.

[21] Hametner C, Jakubek S. Comparison of EM algorithm and partical swarm optimisation for local model network training[C]. Robotics Automation and Mechatronics, 2010: 267-272.

[22] Jaynes E T. Information theory and statistical mechanics[J]. Physical Review, 1957, 106(4):620-630.

[23] Kim D, Seo B. Assessment of the number of components in Gaussian mixture models in the presence of multiple local maximizers[J]. Journal of Multivariate Analysis, 2014, 125(125):100-120.

[24] Lange K. A gradient algorithm locally equivalent to the EM algorithm[J]. Journal of the Royal Statistical Society, 1995, 57(2):425-437.

[25] Leroux B G. Consistent estimation of a mixing distribution[J]. The Annals of Statistics, 1992:1350-1360.

[26] Louis T A. Finding the observed information matrix when using the EM algorithm[J]. Journal of the Royal Statistical Society, 1982, 44(2):226-233.

[27] Ma J H, Fu S. On the correct convergence of the EM algorithm for Gaussian mixtures[J]. Pattern Recognition, 2005, 38(12):2602-2611.

[28] Ma J H, Ge Y. The finite mixture model and its EM algorithm for line-type image patterns[J]. Chinese Journal of Computers, 2007, 30(2): 288-296.

[29] Meng X L, Rubin D B. Maximum likelihood estimation via the ECM algorithm: A general framework[J]. Biometrika,1993,80(2):267-278.

[30] Meng X L, Rubin D B. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm[J]. Publications of the American Statistical Association, 1991, 86(416):899-909.

[31] Neyman J, Scott E L. On the Use of C(α) optimal tests of composite hypothesis[J]. Bulletin of the International Statistical Institute, 1966, 41(1):477 – 497.

[32] Pal S K, Mitra P. Multispectral image segmentation using the: rough-set-initialized EM algorithm[J]. IEEE Transactions on Geoscience and Remote Sensing, 2002, 40(11):2495-2501.

[33] Peng W. Model selection for Gaussian mixture model based on desirability level criterion[J]. Optik - International Journal for Light and Electron Optics, 2016, 130:797-805.

[34] Santos A, Figueiredo E, Sliva M. Genetic-based EM algorithm to improve the robustness of Gassian mixture models for damage detection in bridges[J]. Structural Control and Health Monitoring, 2017, 24(3):1-9.

[35] Schwarz G. Estimating the dimension of a model[J]. The Annals of Statistics, 1978, 6(2):461-464.

[36] Shah N J, Patil H A, Madhavi M C. Deterministic annealing EM algorithm for developing TTS system in Gujarati[C]. International Symposium on Chinese Spoken Language Processing, 2014:526-530.

[37] Shannon C E, Weaver W. A mathematical thery of communication[J]. Bell System Technical Journal, 1948, 27(3):379-423.

[38] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996:267-288.

[39] Ueda N, Nakano R. Deterministic annealing EM algorithm[M]. Elsevier Science Ltd, 1998:271-282.

[40] Vlassis N, Likas A. A greedy EM algorithm for Gaussian mixture learning[J]. Neural Processing Letters, 2002, 15(1):77-87.

[41] Xu C, Chen J. A thresholding algorithm for order selection in finite mixture models[J]. Communications in Statistics - Simulation and Computation, 2015, 44(2):433-453.

[42] Xu L, Jordan M I. ON convergence properties of the EM algorithm for Gaussian mixtures[J]. Neural Computation, 1995, 8(1):129-151.

[43] Yu J, Chaomurilige C, Yang M S. On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures[J]. Pattern Recognition, 2018, 77: 188-203.

致　谢

时间过得飞快, 我的研究生生涯即将迎来了尾声. 感谢浙工商给予我的美好回忆, 在这里我学习了许多, 也成长了许多, 也受到了许多帮助. 在论文完成之际, 我想对所有帮助过我的人说声谢谢.

首先, 我要感谢我的导师王伟刚教授. 正是有了王老师的帮助, 我的论文才能够顺利完成. 在学习上, 王老师总是对我耐心指导, 在我遇到困难时能及时为我指点迷津, 帮助我克服学习路上的一个个难关. 王老师对于学术的严谨以及对生活的乐观态度, 深深影响着我. 在生活上, 王老师也经常关心我, 给我无微不至的关怀, 教我为人处世的道理.

其次, 我要感谢明瑞星老师和董雪梅老师. 他们不仅教授给我专业的理论知识, 还开拓了我在机器学习道路上的视野, 是我能够完成这篇论文的强大助力.

此外我还要感谢许耿鑫师兄在学业上对我的诸多提点和帮助, 同时也预祝师兄顺利博士毕业.

最后, 我要感谢我的父母, 感谢我的女朋友. 感谢他们一直以来对我的照顾和支持, 让我能没有顾虑地在求学路上一直前行.

再次感谢所有帮助过我的人们, 希望你们能被这个世界温柔对待!

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含本人为获得浙江工商大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名:_____ 导师签名:_____

签字日期: 年 月 日 签字日期: 年 月 日

关于论文使用授权的说明

本学位论文作者完全了解浙江工商大学有关保留、使用学位论文的规定：浙江工商大学有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅，可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并且本人电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

学位论文作者签名:_____ 导师签名:_____

签字日期: 年 月 日 签字日期: 年 月 日