密级：公开 中图分类号：O212.1

浙江工商大学

硕士学位论文

论文题目：有限混合模型的参数估计及定阶研究

作者姓名：许耿鑫

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：2018 年 12 月

有限混合模型的参数估计及定阶研究

摘　要

大数据背景下, 庞大复杂的样本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的有力建模工具. 对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, 其意义重大.

本文对有限混合模型的两个问题进行研究. 第一个问题是在给定混合模型的成分个数的条件下, 对模型参数进行估计. 另一个问题则是在成分个数未知情况下的模型选择, 即混合模型的定阶.

针对第一个问题, 文章提出了三种改进的期望最大化算法 (expectation-maximization algorithm, EM 算法), 用于对有限混合模型进行参数估计. 具体地, 在 E 步给定观测数据及当前参数估计值, 计算隐变量的条件期望时, 对混合比例进行不同形式的修改. 改进的算法保持了经典 EM 算法的收敛性. 基于模拟数据以及真实数据的实验结果均表明, 三种改进 EM 算法的收敛速度比经典 EM 算法更快, 同时对有限混合模型的参数估计效果更加准确和稳定.

针对第二个问题, 本文在惩罚对数似然函数中引入 MCP 罚函数 (minimax concave penalty), 提出了 MMCP 方法. 具体地, 该方法在对数似然函数的基础上引入两个惩罚项, 分别对混合比例和成分参数的距离进行惩罚. MMCP 法能够同时实现混合模型的定阶及其它参数的估计. 数值实验结果表明, 相比 MSCAD 方法, MMCP 方法对混合模型的定阶准确率更高.

关键词: 有限混合模型；EM 算法；惩罚似然方法；MCP 罚函数

PARAMETER ESTIMATION AND ORDER SELECTION IN FINITE MIXTURE MODELS

ABSTRACT

Under the circumstances of big data age, the large and complex data set always comes from different species or groups. Finite mixture models form a flexible tool to deal with the heterogeneity. As the top priority of application, it is a significant step to study the mathematical problems.

In this thesis, we investigate two problems in finite mixture models. One is the problem of parameter estimation given the number of components of the model. Another one is to do order selection, that is, to determine the number of components.

As to the first problem, we propose three modified EM algorithms for parameter estimation in finite mixture models. The proposed methods replace the mixing proportions with other estimates while calculating the conditional expectations for the hidden labels given observations and the current estimates for parameters in the E step. We also discuss the convergence properties of the new procedures. Simulation studies show that the new acceleration methods perform much better than the classical EM algorithm in convergence rate and estimation accuracy. A real-data example is examined to illustrate their performance.

Regarding the second problem, a penalized likelihood method with MCP penalty is proposed, called MMCP. To be specific, based on likelihood function, the proposed method introduces two penalty functions on the mixing proportions and the distance between component parameters. The new method can do order selection and parameter estimation simultaneously. Numerical studies show that the new method perform better than MSCAD.

KEYWORDS: Finite mixture models; EM algorithm; Penalized likelihood approach; Minimax concave penalty

目录

第 1 章　引言

1.1　选题背景与研究意义

1.1.1　选题背景

随着计算机、互联网的高速发展, 有限混合模型的应用潜力得到了广泛认可. 混合模型作为模式识别、聚类判别及生存分析等统计方法的支撑, 已经被成功地应用到社会科学、经济学、营销学、天文学、精神病学、遗传学、生物学等领域. 这些行业大都出现了海量、复杂的数据, 普通的单一分布无法在其中挖掘出有用的模式和信息, 而混合模型则有效地解决了这个问题. 纵观当前的众多研究, 不同类型的混合模型被应用到不同的领域当中: 泊松混合模型被应用到医学领域中; 指数混合模型被应用到工程领域当中; 而高斯混合模型则被应用地最广, 这是由于许许多多的随机现象在样本量足够大的情况下, 都可以用高斯分布去逼近.

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处理起来较为复杂. 极大似然估计作为参数估计的经典方法, 由于在混合模型中没有解析解, 因此人们对于混合模型参数估计的早期研究方法基本局限在矩方法上. EM 算法 (Dempster 等, 1977) 的提出, 加上计算机技术的发展, 使得人们能够更高效地处理复杂数据, 这也进一步为有限混合模型的推广应用奠定了基础. EM 算法通过重复 E 步和 M 步, 直到所得的参数迭代序列收敛, 方法的实际操作相对简单, 但其本身也具有不足之处, 如受初始值的影响较大、收敛速度慢且只能达到目标函数的局部最大值, 因此, 对 EM 算法进行改进的意义重大.

关于混合模型的数理研究还有另一个问题, 模型选择, 即确定混合模型的成分个数. 传统 EM 算法在有限混合模型参数估计中的应用是基于成分数已知的情况下进行的, 这符合一些背景知识中给定子群体个数的情形; 然而在实际操作中, 样本数据往往无法直观地呈现成分个数的具体信息. 根据机器学习中的偏差-方差权衡 (bias-variance tradeoff) 这一原则, 一个良好的模型在对给定数据进行拟合时, 需权衡模型自身复杂度, 这有利于模型的进一步推广. 因此, 模型选择的问题在过去几十年中受到了大量的关注.

基于上述背景, 本文拟对有限混合模型的参数估计及定阶进行研究. 一方面, 给出改进的 EM 算法, 加快其在混合模型参数估计中的迭代速度, 同时提升参数估计的准确性和稳定性; 另一方面, 在混合成分未知的情况下, 我们将在前人的工作

基础上引进新的惩罚函数, 利用惩罚似然方法对混合模型进行定阶.

1.1.2 理论意义

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.

1.1.3 实际意义

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, 在如今的大数据背景下, 庞大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地降低模型的复杂度. 因此, 本文研究内容具有明显的应用价值, 能进一步促进有限混合模型在许多科学领域的发展.

1.2 基本概念

混合模型是指一个总体中包含多个子群体的概率模型. 具体地说, 混合模型的总体能够被分成 K 个子群体, 其中每个子群体都能用一个参数模型来表示. 下面, 我们给出有限混合模型的定义.

定义 1.1 令 $\mathcal{F} = \{f(x;\theta); \theta \in \Theta\}$ 为一族参数已知的概率密度函数, 其中 $\theta \in \Theta \subseteq \boldsymbol{R}^m$, $m \geq 1$. 对于具有如下密度函数的随机变量 x, 称其服从混合分布

$$g(x;G) = \int_{\Theta} f(x;\theta) d\Psi(\theta), \tag{1.1}$$

其中 Ψ 是 Θ 上的分布函数, 被称为混合率.

若混合率 Ψ 具有有限个支撑点 $\theta_1, \theta_2, \cdots, \theta_K$, 相对应的权重为 $\pi_1, \pi_2, \cdots, \pi_K$,

即

$$\Psi(\theta) = \sum_{k=1}^{K} \pi_k I(\theta_k \leq \theta),$$

那么式 (??) 可变为

$$g(x; G) = \sum_{k=1}^{K} \pi_k f(x; \theta_k), \tag{1.2}$$

其中参数 $G = (K, \pi_1, \pi_2, \cdots, \pi_{K-1}, \theta_1, \theta_2, \cdots, \theta_K)$, 且 $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$ $(k = 1, 2, \ldots, K)$.

注意 G 中包含了三类参数: K 代表子群体的个数, 即混合模型中的成分个数 (阶数), 我们称估计成分个数 K 为定阶; π_k 被称为混合比例; θ_k 则是成分参数. 子群体的密度函数 $f(x; \theta_k)$ 可以是高斯分布、泊松分布、二项分布等. 根据式 (??) 中 K 是否给定, 对混合模型的参数估计可以分为两大类型, 文章第二章和第三章将分别对此进行研究.

混合模型早在 100 多年前就被数学家所应用. Pearson (1894) 在对 1000 只螃蟹头部到其躯干长度的数据进行分析时, 发现单一的高斯分布无法很好地刻画数据. 在被告知原因可能是样本数据包含了两类螃蟹后, Pearson 提出用两成分高斯混合模型来拟合数据, 而两个权重分别表示两类螃蟹各自所占的比例, 即

$$g(x; G) = \sum_{k=1}^{2} \pi_k f(x; \mu_k, \sigma_k),$$

其中 $G = (\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$, $0 < \pi_1 < 1$ 且 $\pi_1 + \pi_2 = 1$,

$$f(x; \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\{-\frac{(x - \mu_k)^2}{2\sigma_k^2}\}.$$

自 Pearson 提出混合模型之后, 相关的理论知识受到了深入研究, 同时也被应用到天文学、工程学、社会科学、生物学及医药学等领域.

在对有限混合模型进行参数估计之前, 需要假定问题是可识别的. 下面我们看一个例子: 给定两成分均匀混合分布的独立样本 $\{x_1, x_2, \cdots, x_n\}$, 我们的目的是根据这些样本对其总体分布

$$g(x) = \pi_1 U(m_{11}, m_{12}) + (1 - \pi_1) U(m_{21}, m_{22})$$

中参数 $(\pi_1, m_{11}, m_{12}, m_{21}, m_{22})$ 进行估计. 假定该总体的真实模型为

$$g^*(x) = \frac{a}{a+b}U(-a,a) + \frac{b}{a+b}U(-b,b).$$

其中 $a, b > 0$ 且 $a \neq b$. 那么, 这是一个不可识别的问题. 因为改变其中的参数, 得到混合模型

$$g^{**}(x) = \frac{1}{2}U(-b,a) + \frac{1}{2}U(-a,b),$$

可观察到

$$g^*(x) \equiv g^{**}(x).$$

这时我们无法通过给定样本来判断总体的确切模型. 换言之, 我们无法识别各个子群体的分布. 因此, 统计问题的可识别性是非常重要. 下面给出可识别性的定义:

定义 1.2 考虑有限混合模型

$$g(x;G) = \sum_{k=1}^{K} \pi_k f(x;\theta_k),$$

其中 $G = (K, \pi_1, \pi_2, \cdots, \pi_{K-1}, \theta_1, \theta_2, \cdots, \theta_K) \in \boldsymbol{\Omega}$,

$$\boldsymbol{\Omega} = \{G; \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0, \theta_k \in \Theta \subseteq \boldsymbol{R}^m, k = 1, 2, ..., K; K \geq 1\}.$$

对于任意 G^* 及 G^{**}, 若

$$g(x;G^*) \equiv g(x;G^{**})$$

当且仅当 $G^* = G^{**}$ 时成立, 那么称有限混合模型类 $\mathcal{G} = \{g(x;G); G \in \boldsymbol{\Omega}\}$ 是可识别的.

有限指数分布族混合模型, 包括单变量的高斯混合模型、泊松混合模型、伽马混合模型等, 都是可识别的 (Teicher, 1960, 1963). 依据 Holzmann 等 (2004)、Holzmann 等 (2006) 及 Ahmad 等 (2010) 的结论, 可得到有限逻辑混合模型、多变量有限 t 混合模型、有限 Gumbel 混合模型均是可识别的. 此外, 更多与有限混合模型可识别性相关的内容, 可参见以上文献.

1.3 文献综述

1.3.1 有限混合模型的参数估计

在计算机发展之前, 学者们对有限混合模型的参数估计普遍采用矩方法. 早在 1894 年, Pearson 利用 5 个矩方程求解出两成分高斯混合分布的 5 个参数. 而 Day (1969) 和 Cohen (1967) 同样讨论了利用矩方法估计两成分高斯混合分布参数的问题. 矩方法需要求解非线性方程, 并且随着模型成分的增加, 方程数量也随之增加, 这给求解带来很大的困难.

极大似然估计相比矩估计, 有很多更好的性质 (如渐进性质), 但其计算难度并不简单. 伴随着计算机的发展, 极大似然方法在有限混合模型的参数估计中越来越受关注. Hasselblad (1966, 1969) 是第一个将极大似然估计法应用到指数分布族混合模型的学者. 1970 年, Behboodian 进一步细化了高斯混合分布中的参数估计问题. Böhning (1995) 对混合模型中极大似然估计的算法进行了总结, 其中最常规的有两种: Newton-Raphson 方法以及 Dempster 等 (1977) 提出的 EM 算法. 由于 EM 算法的程序相对简洁, 其在相关研究中更受欢迎. 本文中, 我们将重点研究 EM 算法并对其进行改进.

EM 算法的提出, 大大加快了有限混合模型的相关理论研究. 如 Bilmes (1998) 用 EM 算法对高斯混合模型进行参数估计; 此后, Figueiredo 和 Jain (2002) 对高斯混合模型参数估计中的 EM 算法进行改进; 而对于有限混合模型的假设检验问题, Chen 等 (2001) 对经典的似然比检验进行了修改, Chen 和 Chen (2001) 则具体研究了高斯混合分布似然比检验中, 统计量的性质及其大样本分布; Garel (2005) 深入研究了混合模型中似然比检验的经典渐进理论. 此外, McLachlan (2008) 给出了利用 EM 算法对有限高斯混合模型进行参数估计的实例.

对于 EM 算法本身, 国内外学者也做了大量研究. 如 Louis (1982) 研究了在不完全数据问题中用 EM 算法求最大似然估计时, 观测信息矩阵的提取方法; Wu (1983) 对一般情形的 EM 算法收敛性质进行了详细研究. 针对 EM 算法收敛速度慢、受初始值影响大等问题, 人们开始对 EM 算法的进行改进, 并提出新的算法, 如 Wei 和 Tanner (1990) 提出的蒙特卡洛 EM (Monte Carlo EM, MCEM) 算法、Meng 和 Rubin (1993) 的条件期望最大化 (expectation conditional maximization, ECM) 算法、Liu 和 Rubin (1994) 提出的 ECME (expectation conditional maximization either) 算法等. Biernacki 等人 (2003) 则对 EM 算法及以上几种改进方法在高斯混合模型中的初始值选择问题进行了研究.

近几年来, 有限混合模型的参数估计问题依旧受到广泛关注. Yin 等人 (2012) 提出了通过频繁更新方法来加速 EM 算法, 并将其应用到高斯混合模型中; Melnykov 和 Melnykov (2012) 针对成分未知的多元高斯混合模型, 提出了 EM 算法初始值选取的新方法; 由于 EM 算法对初始值很敏感, 且无法避免陷入局部最优, Ueda 和 Nakano(1998 年) 提出了确定性退火算法 (deterministic annealing EM algorithm, DA-EM), Yu 等人 (2018) 进一步研究了 DA-EM 算法的收敛性及参数选择问题.

混合模型的发展及应用, 与 EM 算法的推广, 有着相辅相成、密不可分的关系. 有意思的是, 关于混合模型中的参数估计这一最基本的统计问题, 一直未能得到圆满的解决. 尽管 EM 算法得到不断改进, 也仍存在一些不足, 例如: 受初值影响大、算法收敛速度慢、不能保证收敛到全局最大值等, 这些都值得深入研究的问题.

1.3.2 有限混合模型的定阶

在过去的几十年里, 不少学者致力于研究有限混合模型的定阶问题. 所提出的统计方法大致可以分为如下五类: 首先是经典的信息准则类, 如 Akaike 提出的 AIC (1973) 及 Schwarz 提出的 BIC (1978), Leroux (1992) 讨论了经典有限混合模型中基于 AIC 和 BIC 的定阶问题; 其次是由 Chen 和 Kalbfleisch (1996) 提出的惩罚最小距离法, 这种方法通过关于混合比率的惩罚函数, 达到防止其取值过小的过拟合情形. James 等 (2001) 则提出最小化真实密度函数非参数估计与高斯密度函数卷积的 Kullback–Leibler 距离的半参数方法; 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, 2017); 再者, Ishwaran 等 (2001) 基于混合分布空间上的先验分布, 通过分解所得的边际密度函数, 提出了加权贝叶斯因子方法, 以实现对有限混合模型成分数的估计; 最后一类方法是 Chen 和 Khalili (2008) 提出的惩罚似然方法, 该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.

近几年来, 学者们还对混合分布模型选择的相关问题进行研究. 2014 年, Kim 和 Seo 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. Saraiva 等 (2014) 利用一种新的后验分裂合并 MCMC 算法 (posterior split–merge MCMC algorithm, PSM) 来进行定阶及参数估计, 并在模拟数据集和真实数据集上验证算法的性能. 此外, 还有 Peng (2016) 提出的期望水平准则 (desirability level criterion)、Štepánová 和 Vavrečk (2016) 提出的对贪婪算法和合并算法进行结合的

新算法、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.

以上研究方法中, 惩罚似然方法是解决有限混合模型定阶问题较为新颖的方法, 它把回归模型中变量选择的正则化思想引入到混合模型中. 然而, 惩罚对数似然函数的求解需要有效且高效的算法. 参考回归模型中罚函数的改进, 对有限混合模型中惩罚对数似然函数所使用的罚函数进行改进, 以及对现有求解算法进行改进, 都是很有意义的研究方向.

1.4 本文的创新点

第一, 本文提出了三个改进的 EM 算法, 对混合模型进行参数估计. 作为一种优良的数值求解方法, EM 算法在很大程度上促进了混合模型的应用. 然而, 其收敛速度过慢的问题也广受关注. 以往的研究主要集中在初始值的选取、M 步的简化方面, 本文创新性地对 E 步进行改进. 具体地, 每一次迭代中, 新方法都会重新利用一次样本信息, 这使得新算法的收敛速度更快. 同时, 相比传统的 EM 算法, 三个改进的 EM 算法在复杂混合模型中的参数估计效果更加准确和稳定.

第二, 本文将 MSCAD 方法中的 SCAD 罚函数修改为 MCP 罚函数, 提出了 MMCP 方法. 一方面, 由于该方法在对数似然函数的基础上, 分别对混合比例和成分参数的距离进行惩罚, 因此, MMCP 法能够同时实现混合模型的定阶及其它参数的估计, 比 AIC、BIC 等信息准则方法更加高效; 另一方面, 由于罚函数的改进, MMCP 方法能够更好地处理复杂混合模型的模型选择问题. 数值实验结果表明, 相比 MSCAD 方法, MMCP 方法对混合模型的定阶准确率更高.

1.5 基本框架

本文内容分为 4 章, 各部分的主要内容介绍如下.

第 1 章是引言部分, 介绍有限混合模型的知识背景及研究意义、一些基本概念及所研究问题的前提假设, 并阐述各种经典方法的优缺点及本文的创新点.

第 2 章研究了本文的第一个问题: 在成分个数已知的情形下, 有限混合模型的参数估计. 首先介绍经典 EM 算法并对其不足之处进行讨论; 接着, 对 E 步中隐变量条件期望的计算方法进行不同形式的修改, 提出三种改进的 EM 算法, 并阐述其数值求解方法; 最后, 通过模拟数据与实际数据对改进的 EM 算法及经典算法的效果进行比较.

第 3 章研究了本文的第二个问题: 有限混合模型的定阶问题. 从信息准则方法和惩罚最小距离法入手, 阐述其计算效率低及容易导致过拟合的问题; 接着, 我们基于以上工作, 对惩罚似然方法进行改进, 提出 MMCP 方法及其数值求解方法, 并对新方法进行实验模拟与总结.

第 4 章是总结与展望, 总括本文所研究的主要问题、方法及结果, 并对未来的拓展性工作进行安排.

第 2 章　基于改进 EM 算法的有限混合模型参数估计

本章将研究在成分个数已知的情形下, 混合模型的参数估计问题. 具体结构如下: 首先, 本章第 1 节中, 我们对有限混合模型参数估计问题及经典 EM 算法进行介绍, 同时展示其不足之处; 其次, 在第 2 节中, 我们将提出三种改进的 EM 算法, 并介绍其数值求解方法; 最后, 在第 3 节中, 通过模拟数据与实际数据, 对改进的 EM 算法和经典 EM 算法的收敛速度、参数估计效果等进行比较.

2.1　有限混合模型参数估计问题及 EM 算法

我们描述所研究的问题: 对于具有如下概率密度函数的有限混合模型

$$g(x;G)=\int_{\Theta}f(x;\theta)d\Psi(\theta), \tag{2.1}$$

其中参数 $G=(\pi_1,\pi_2,\cdots,\pi_K,\theta_1,\theta_2,\cdots,\theta_K)$, 混合率 $\Psi(\theta)=\sum_{k=1}^{K}\pi_k I(\theta_k\leq\theta)$, 且 $\sum_{k=1}^{K}\pi_k=1$, $\pi_k\geq 0$ $(k=1,2,\ldots,K)$, 给定随机独立样本 $\{x_1,x_2,\cdots,x_n\}$, 在成分个数 K 已知的情形下, 估计混合比例 π_k 以及成分参数 θ_k.

根据所给样本, 我们可以得到参数 G 的对数似然函数

$$l_n(G)=\sum_{i=1}^{n}\log\{\sum_{k=1}^{K}\pi_k f(x_i;\theta_k)\}. \tag{2.2}$$

通过最大化 $l_n(G)$, 即可得到所求参数的最大似然估计. 由于式 (??) 的对数函数内包含求和, 这造成了计算上的困难. 下面介绍如何应用经典的 EM 算法有效解决极大化该对数似然函数的计算问题.

首先, 针对模型 (??), 引入 K 维二值隐变量 $\boldsymbol{z}=(z_1,z_2,\cdots,z_K)^T$ 指示变量 x 所属的成分, 即 $z_k\in\{0,1\}$ 且 $\sum_{k=1}^{K}z_k=1$, 同时, 根据混合比例 π_k 对 z 的边缘概率分布进行赋值, 即

$$p(z_k=1)=\pi_k. \tag{2.3}$$

由于 K 维二值随机变量 $\boldsymbol{z}$ 仅有一项为 1, 其余为 0, 故有

$$p(\boldsymbol{z})=\prod_{k=1}^{K}\pi_k^{z_k}. \tag{2.4}$$

再者, 对 $\boldsymbol{z}$ 给定某个值, x 的条件概率分布为

$$p(x|z_k = 1) = f(x; \theta_k), \tag{2.5}$$

该式亦可写成

$$p(x|\boldsymbol{z}) = \prod_{k=1}^{K} f^{z_k}(x; \theta_k). \tag{2.6}$$

根据乘法公式, 可得完整数据 $(x, \boldsymbol{z})$ 的概率分布为

$$p(x, \boldsymbol{z}) = p(\boldsymbol{z})p(x|\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} f^{z_k}(x; \theta_k) = \prod_{k=1}^{K} [\pi_k f(x; \theta_k)]^{z_k}. \tag{2.7}$$

这里, 我们对基于完整数据 $(x, \boldsymbol{z})$ 所求出的 x 的边缘概率分布进行验证. 根据全概率公式, 有

$$p(x) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(x|\boldsymbol{z}) = \sum_{k=1}^{K} p(z_k = 1)p(x|z_k = 1) = \sum_{k=1}^{K} \pi_k f(x; \theta_k), \tag{2.8}$$

这正好就是式 (**??**) 所展示的混合密度函数. 因此, 我们找到了混合分布的等价公式 (Bishop, 2006).

给定样本 $\{x_1, x_2, \cdots, x_n\}$, 那么相应地, 有隐变量 $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$, 指示各个样本所属的成分, 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 $\sum_{k=1}^{K} z_{ik} = 1$. 根据所给样本及其属性 (组成的完整数据), 我们可以得到参数 G 的似然函数

$$L^c(G) = \prod_{i=1}^{n} p(x_i, \boldsymbol{z}_i) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f(x_i; \theta_k)]^{z_{ik}}.$$

上式第二个等号我们应用了公式 (**??**). 取对数, 可得到参数 G 的完整对数似然函数 (Lehmann, 1998)

$$l_n^c(G) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} [\log \pi_k + \log\{f(x_i; \theta_k)\}]. \tag{2.9}$$

虽然通过引入隐变量, 使得极大化式 (**??**) 比极大化式 (**??**) 的问题容易解决. 然而由于完整数据中包含虚拟数据, 故不可能从极大化 $l_n^c(G)$ 得到混合比例 π_k 以及成分参数 θ_k 的极大似然估计. 正因为 $l_n(G)$ 和 $l_n^c(G)$ 都不能作为计算参数的极

大似然估计的依据 (用 $l_n(G)$ 计算时会遭遇计算上的困难, 而 $l_n^c(G)$ 则是包含了虚拟数据的对数似然函数), 经典的 EM 算法, 提出了第三种对数似然函数, 该函数是经过迭代计算来产生的. 算法具体如下: 先任意地给定一个参数初始值, $G^{(0)}$. 当 $G^{(0)}$ 确定以后, 计算 $l_n^c(G)$ 关于 z_{ik} 的条件期望 (expectation step, E 步)

$$Q(G;G^{(0)}) \coloneqq E[l_n^c(G);G^{(0)}] = \sum_{i=1}^{n}\sum_{k=1}^{K} E[z_{ik}|x_i;G^{(0)}][\log \pi_k + \log\{f(x_i;\theta_k)\}], \quad (2.10)$$

其中

$$\begin{aligned} E[z_{ik}|x_i;G^{(0)}] &= P(z_{ik}=1|x_i;G^{(0)}) \\ &= \frac{P(z_{ik}=1;G^{(0)})P(x_i|z_{ik}=1;G^{(0)})}{\sum_{j=1}^{K} P(z_{ij}=1;G^{(0)})P(x_i|z_{ij}=1;G^{(0)})} \\ &= \frac{\pi_k^{(0)} f(x_i;\theta_k^{(0)})}{\sum_{j=1}^{K} \pi_j^{(0)} f(x_i;\theta_j^{(0)})}. \end{aligned} \quad (2.11)$$

我们称 $Q(G;G^{(0)})$ 为期望对数似然函数 (Lehmann, 1998). 通过将 $Q(G;G^{(0)})$ 极大化 (maximization step, M 步) 得到 $G^{(1)}$ 为

$$G^{(1)} = \arg\max_{G} Q(G;G^{(0)}). \quad (2.12)$$

便从 $G^{(0)}$ 得到 $G^{(1)}$, 依据上述迭代方法, 重复 E 步、M 步两个步骤, 得到 EM 估计序列 $\{G^{(1)}, G^{(2)}, \cdots\}$. 在实际计算中, 当所估计的参数序列收敛的时候, 迭代即可终止.

实际计算中, EM 算法的收敛速度很慢. 以高斯混合模型为例, 对某七成分高斯混合模型 (表 **??** 中的模型 7) 进行估计时, EM 算法达到收敛所需迭代次数为 2000 多次; 而当总体分布中的某些子总体相近 (表 **??** 中的模型 8、模型 9 和模型 10) 时, EM 算法则需经过 5000 多次迭代才达到收敛. 章节 **??** 将在混合模型、样本量、成分数等设定不同的情况, 对算法的收敛速度及估计效果进行详细讨论.

2.2 改进 EM 算法的理论与求解

2.2.1 改进 EM 算法的理论

章节 **??** 所描述的经典 EM 算法, 有效地解决了有限混合模型中极大化对数似然函数的计算问题, 但该迭代方法的收敛速度很慢, 特别是在混合成分较多的复杂

混合模型中, 需要经过大量的迭代次数, 参数估计序列才达到收敛. 下面, 我们针对有限混合模型的参数估计问题, 提出三种改进的 EM 算法.

注意到 EM 算法的 E 步中, 在给定观测样本和当前参数估计值 $G^{(m)}$ 后, z_{ik} 的条件期望为

$$E[z_{ik}|x_i;G^{(m)}]=\frac{\pi_k^{(m)}f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K}\pi_j^{(m)}f(x_i;\theta_j^{(m)})}. \tag{2.13}$$

上式中, 混合比例 π_k 被设定为上一次迭代的参数估计结果 $\pi_k^{(m)}$, 并且, 对于任意样本 x_i, 该参数的先验信息设定一致. 现在, 我们将该混合比例的取值做出不同修改.

第一种改进的算法中, 我们对不同样本的混合比例, 给定不同的先验信息, 该先验信息是根据上一次迭代的参数估计结果, 计算出任意样本 x_i "属于" 各个成分的可能性, 即

$$\tilde{\pi}_{ik}^{(m+1)}=\frac{f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K}f(x_i;\theta_j^{(m)})}. \tag{2.14}$$

于是, z_{ik} 的条件期望为可被替换为

$$\tilde{E}[z_{ik}|x_i;G^{(m)}]=\frac{\tilde{\pi}_{ik}^{(m+1)}f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K}\tilde{\pi}_{ij}^{(m+1)}f(x_i;\theta_j^{(m)})}. \tag{2.15}$$

第二种改进的算法受启发于有限混合分布的抽样过程. 众所周知, 有限混合分布的抽样可以分为两步: 第一步是根据设定的混合比例 $\pi_k(k=1,2,\cdots,K)$, 从多项分布 $p_n(\pi_1,\pi_2,\cdots,\pi_K)$ 中随机抽取一次样本 $(n_1,n_2,\cdots,n_K)$; 第二步是在每一个成分 $f(x;\theta_k)(k=1,2,\cdots,K)$ 中, 分别抽取样本容量为 $n_k(k=1,2,\cdots,K)$ 的随机独立样本. 这样得到的所有 n 个样本即为有限混合模型的 n 个随机独立观测值. 第二种改进的方法, 就是根据样本在各个成分密度函数上的概率值, 将其分配到 "最可能" 的成分类别中, 再将式 (**??**) 中的混合比例 π_k 用第 k 个成分的样本量所占比率代替, 即对于第 k 个成分, 有

$$\tilde{\pi}_k^{(m+1)}=card\{k|k\in A\}/n, \tag{2.16}$$

其中,

$$A=\{k_i|k_i=\underset{j\in\{1,2,\cdots,K\}}{\arg\max}\ f(x_i;\theta_j^{(m)});i\in\{1,2,\cdots,n\}\}.$$

于是, z_{ik} 的条件期望为可替换为

$$\tilde{E}[z_{ik}|x_i;G^{(m)}]=\frac{\tilde{\pi}_k^{(m+1)}f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K}\tilde{\pi}_j^{(m+1)}f(x_i;\theta_j^{(m)})}. \tag{2.17}$$

第三种改进的算法是前两种方法的结合. 注意到式 (**??**) 中, $\tilde{\pi}_k^{(m+1)}$ 用第 k 个成分的所分配到的样本量所占比率来计算, 这里我们不直接对样本进行分类, 而是利用第一种改进算法中, 针对不同样本计算出其 "属于" 各个成分的可能性做算术平均, 即

$$\tilde{\pi}_k^{(m+1)} = \sum_{i=1}^{n} \tilde{\pi}_{ik}^{(m+1)}/n, \tag{2.18}$$

其中, $\tilde{\pi}_{ik}^{(m+1)}$ 根据式 (**??**) 计算所得. 接着, 按照式 (**??**) 替换 z_{ik} 的条件期望.

现在, 针对有限混合模型的参数估计问题, 我们将三种改进的 EM 算法步骤总结如下.

Algorithm 1 新方法 1

Input: 数据集 $\{x_1, x_2, \cdots, x_n\}$, 成分数 K;
Output: 参数 G 的估计值;

1: 参数初始值 $G^{(0)}$;
2: **repeat**
3: 计算 $\tilde{\pi}_{ik}^{(m+1)} = \frac{f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K} f(x_i;\theta_j^{(m)})}$;
4: 计算 $\tilde{E}[z_{ik}|x_i; G^{(m)}] = \frac{\tilde{\pi}_{ik}^{(m+1)} f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K} \tilde{\pi}_{ij}^{(m+1)} f(x_i;\theta_j^{(m)})}$;
5: 计算 $G^{(m+1)} = \arg\max_G Q(G; G^{(m)})$, 其中

$$Q(G; G^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{E}[z_{ik}|x_i; G^{(m)}][\log \pi_k + \log\{f(x_i;\theta_k)\}]; \tag{2.19}$$

6: **until** 收敛.

Algorithm 2 新方法 2

Input: 数据集 $\{x_1, x_2, \cdots, x_n\}$, 成分数 K;
Output: 参数 G 的估计值;

1: 参数初始值 $G^{(0)}$;
2: **repeat**
3: 计算 $\tilde{\pi}_k^{(m+1)} = card\{k|k \in A\}/n$, 其中

$$A = \{k_i | k_i = \underset{j\in\{1,2,\cdots,K\}}{\arg\max} f(x_i;\theta_j^{(m)}); i \in \{1, 2, \cdots, n\}\};$$

4: 计算 $\tilde{E}[z_{ik}|x_i; G^{(m)}] = \frac{\tilde{\pi}_k^{(m+1)} f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K} \tilde{\pi}_j^{(m+1)} f(x_i;\theta_j^{(m)})}$;
5: 计算 $G^{(m+1)} = \arg\max_G Q(G; G^{(m)})$, 其中 $Q(G; G^{(m)})$ 如式 (**??**) 所示;
6: **until** 收敛.

值得一提的是, 如同经典的 EM 算法, 新算法只是收敛到局部最大化, 而非全局最大化. 同时, 依据迭代方法, 上述三种改进的 EM 算法均能够计算出参数估计

Algorithm 3 新方法 3

Input: 数据集 $\{x_1, x_2, \cdots, x_n\}$, 成分数 K;
Output: 参数 G 的估计值;
1: 参数初始值 $G^{(0)}$;
2: **repeat**
3: 计算 $\tilde{\pi}_k^{(m+1)} = \sum_{i=1}^n \tilde{\pi}_{ik}^{(m+1)}/n$, 其中 $\tilde{\pi}_{ik}^{(m+1)} = \frac{f(x_i;\theta_k^{(m)})}{\sum_{j=1}^K f(x_i;\theta_j^{(m)})}$;
4: 计算 $\tilde{E}[z_{ik}|x_i; G^{(m)}] = \frac{\tilde{\pi}_k^{(m+1)} f(x_i;\theta_k^{(m)})}{\sum_{j=1}^K \tilde{\pi}_j^{(m+1)} f(x_i;\theta_j^{(m)})}$;
5: 计算 $G^{(m+1)} = \arg\max_G Q(G; G^{(m)})$, 其中 $Q(G; G^{(m)})$ 如式 (**??**) 所示;
6: **until** 收敛.

序列. 实际计算中, 我们会设定终止迭代的条件: 一般地, 对于很小的正数 ε_1 及 ε_2, 如果满足

$$\left\|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}\right\| < \varepsilon_1$$

或者

$$\left\|Q(G^{(m+1)}; G^{(m)}) - Q(G^{(m)}; G^{(m-1)})\right\| < \varepsilon_2$$

则终止算法, 认为序列已达到收敛.

接下来, 我们给出算法的收敛性定理.

定理 2.1 设 $\boldsymbol{X}$ 服从混合分布 (**??**), 并有随机独立观察值 $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$, 记参数 G 的对数似然函数为 $l_n(G|\boldsymbol{x})$. 又设 $(\boldsymbol{X}, \boldsymbol{Z})$ 为构造的完整数据, 相应的对数似然函数为 $l_n^c(G|\boldsymbol{x}, \boldsymbol{z})$. 任意给定的参数初始值 $G^{(0)}$, 由此产生的估计序列 $\{G^{(1)}, G^{(2)}, \cdots\}$ 满足

$$l_n(G^{(m+1)}|\boldsymbol{x}) \geq l_n(G^{(m)}|\boldsymbol{x}), m = 1, 2, \cdots, \tag{2.20}$$

上式等号成立的充要条件为 $Q(G^{(m+1)}|\boldsymbol{x}; G^{(m)}) = Q(G^{(m)}|\boldsymbol{x}; G^{(m)})$, $m = 1, 2, \cdots$, 其中 Q 如式 (**??**) 所示.

证明 记 $(\boldsymbol{X}, \boldsymbol{Z})$ 的联合密度函数为 $f(\boldsymbol{x}, \boldsymbol{z}; G)$, 则

$$f(\boldsymbol{z}|\boldsymbol{x}; G) = \frac{f(\boldsymbol{x}, \boldsymbol{z}; G)}{g(\boldsymbol{x}; G)}$$

为 $\boldsymbol{Z}$ 的条件密度函数, 那么, 相应的对数似然函数有如下关系:

$$l_n(G|\boldsymbol{x}) = l_n^c(G|\boldsymbol{x}, \boldsymbol{z}) - \log f(\boldsymbol{z}|\boldsymbol{x}; G). \tag{2.21}$$

对于给定的参数初始值 $G^{(0)}$, 定义

$$Q(G|\boldsymbol{x};G^{(0)}) = \int \log(f(\boldsymbol{x},\boldsymbol{z};G))f(\boldsymbol{z}|\boldsymbol{x};G^{(0)})d\boldsymbol{z},$$

$$H(G|\boldsymbol{x};G^{(0)}) = \int \log(f(\boldsymbol{z}|\boldsymbol{x};G))f(\boldsymbol{z}|\boldsymbol{x};G^{(0)})d\boldsymbol{z},$$

其中 $Q(G|\boldsymbol{x};G^{(0)})$ 即由式 (**??**) 所示的期望对数似然函数. 由于式 (**??**) 中 $l_n(G|\boldsymbol{x})$ 并不依赖于 $\boldsymbol{z}$, 故其关于密度函数 $f(\boldsymbol{z}|\boldsymbol{x};G^{(0)})$ 的期望仍为 $l_n(G|\boldsymbol{x})$. 于是, 式 (**??**) 两边对密度函数 $f(\boldsymbol{z}|\boldsymbol{x};G^{(0)})$ 求期望, 可得

$$l_n(G|\boldsymbol{x}) = Q(G|\boldsymbol{x};G^{(0)}) - H(G|\boldsymbol{x};G^{(0)}). \tag{2.22}$$

上式中, G 和 $G^{(0)}$ 可以是任何值. 取 $G^{(0)} = G^{(m)}$, $G = G^{(m+1)}$, 可得 $l_n(G^{(m+1)})$ 与 $l_n(G^{(m)})$ 由于 (**??**) 可得, (**??**) 两边之差为

$$\begin{aligned} l_n(G^{(m+1)}|\boldsymbol{x}) - l_n(G^{(m)}|\boldsymbol{x}) &= [Q(G^{(m+1)}|\boldsymbol{x};G^{(m)}) - Q(G^{(m)}|\boldsymbol{x};G^{(m)})] \\ &\quad -[H(G^{(m+1)}|\boldsymbol{x};G^{(m)}) - H(G^{(m)}|\boldsymbol{x};G^{(m)})]. \end{aligned}$$

根据 $G^{(m+1)}$ 的定义, 上式等号右边第一个方括号内的值非负, 故只需证明第二个方括号内的值非正, 即

$$\int [\log(f(\boldsymbol{z}|\boldsymbol{x};G^{(m+1)})) - \log(f(\boldsymbol{z}|\boldsymbol{x};G^{(m)}))]f(\boldsymbol{z}|\boldsymbol{x};G^{(m)})d\boldsymbol{z} \leq 0.$$

由 Jensen 不等式可得

$$\int [\log \frac{f(\boldsymbol{z}|\boldsymbol{x};G^{(m+1)})}{f(\boldsymbol{z}|\boldsymbol{x};G^{(m)})}]f(\boldsymbol{z}|\boldsymbol{x};G^{(m)})d\boldsymbol{z} \leq \log \int f(\boldsymbol{z}|\boldsymbol{x};G^{(m+1)})d\boldsymbol{z} = 0.$$

2.2.2 改进 EM 算法的数值求解

下面, 我们以高斯混合模型和泊松混合模型为例, 说明改进 EM 算法的数值求解.

首先, 给定高斯混合模型

$$g(x;G) = \sum_{k=1}^{K} \pi_k N(x;\mu_k,\sigma_k^2)$$

的随机独立样本 $\{x_1, x_2, \cdots, x_n\}$, E 步直接由公式 (**??**) 或 (**??**) 计算可得 z_{ik} 条件期望的替换值 $\tilde{E}[z_{ik}|x_i; G^{(m)}]$, M 步求解 $G^{(m+1)} = \arg\max_G Q(G; G^{(m)})$, 其中 $G = \{\pi_k, \mu_k, \sigma_k : k = 1, 2, \cdots, K\}$,

$$Q(G; G^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{E}[z_{ik}|x_i; G^{(m)}][\log \pi_k + \log N(x_i; \mu_k, \sigma_k^2)]. \tag{2.23}$$

注意到限制条件 $\sum_{j=1}^{K} \pi_j = 1$, 于是问题转化为求解方程组

$$\begin{cases} \frac{\partial Q(G;G^{(m)})}{\partial \mu_k} = 0, \\ \frac{\partial Q(G;G^{(m)})}{\partial \sigma_k^2} = 0, \\ \frac{\partial [Q(G;G^{(m)}) + \lambda(\sum_{j=1}^{K} \pi_j - 1)]}{\partial \pi_k} = 0. \end{cases}$$

可得

$$\begin{cases} \hat{\mu}_k = \frac{\sum_{i=1}^{n} \tilde{E}[z_{ik}|x_i;G^{(m)}]x_i}{\sum_{i=1}^{n} \tilde{E}[z_{ik}|x_i;G^{(m)}]}, \\ \hat{\sigma}_k = \sqrt{\frac{\sum_{i=1}^{n} \tilde{E}[z_{ik}|x_i;G^{(m)}](x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{n} \tilde{E}[z_{ik}|x_i;G^{(m)}]}}, \\ \hat{\pi}_k = \frac{\sum_{i=1}^{n} \tilde{E}[z_{ik}|x_i;G^{(m)}]}{n}. \end{cases}$$

再者, 给定泊松混合模型

$$g(x; G) = \sum_{k=1}^{K} \pi_k P(x; \theta_k)$$

的随机独立样本 $\{x_1, x_2, \cdots, x_n\}$, 则 M 步中为求解 $G^{(m+1)} = \arg\max_G Q(G; G^{(m)})$, 其中 $G = \{\pi_k, \theta_k : k = 1, 2, \cdots, K\}$),

$$Q(G; G^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{E}[z_{ik}|x_i; G^{(m)}][\log \pi_k + \log P(x_i; \theta_k)]. \tag{2.24}$$

同样地, 考虑到限制条件 $\sum_{j=1}^{K} \pi_j = 1$, 问题可转化为求解方程组

$$\begin{cases} \frac{\partial Q(G;G^{(m)})}{\partial \theta_k} = 0, \\ \frac{\partial [Q(G;G^{(m)}) + \lambda(\sum_{j=1}^{K} \pi_j - 1)]}{\partial \pi_k} = 0. \end{cases}$$

解得

$$\begin{cases} \hat{\theta}_k = \frac{\sum_{i=1}^n \tilde{E}[z_{ik}|x_i;G^{(m)}]x_i}{\sum_{i=1}^n \tilde{E}[z_{ik}|x_i;G^{(m)}]}, \\ \hat{\pi}_k = \frac{\sum_{i=1}^n \tilde{E}[z_{ik}|x_i;G^{(m)}]}{n}. \end{cases}$$

2.3 数值实验

为了说明本文所提三种改进 EM 算法的性能, 我们将对传统 EM 算法和新方法在模拟和实例数据上进行研究. 具体地, 我们计算四种算法的计算速度, 并评估他们对参数估计的准确度和稳定性.

2.3.1 模拟研究

我们分别从高斯混合模型和泊松混合模型中产生样本数据. 依据模型的复杂度, 高斯混合模型的样本设定 $n = 100, 400$, 泊松混合模型则针对所有模型产生 $n = 100, 500$ 的样本. 每个模型的运行结果均基于 500 次蒙特卡洛模拟.

例 2.1 首先, 我们从如下高斯混合模型的密度函数

$$g(x;G) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\}.$$

中生成样本数据. 10 个模型的参数设定如表?? (Ishwaran 等, 2001) 所示. 我们令所有成分的标准差 $\sigma_k = 1$, 并对于每一个模型的估计, 均考虑 σ 已知和未知两种情形.

表 2-1: 高斯混合模型参数设定.

模型	(π_1, μ_1)	(π_2, μ_2)	(π_3, μ_3)	(π_4, μ_4)	(π_5, μ_5)	(π_6, μ_6)	(π_7, μ_7)
1	(1/3,0)	(2/3,3)					
2	(1/2,0)	(1/2,3)					
3	(1/2,0)	(1/2,1.8)					
4	(1/4,0)	(1/4,3)	(1/4,6)	(1/4,9)			
5	(1/4,0)	(1/4,1.5)	(1/4,3)	(1/4,4.5)			
6	(1/4,0)	(1/4,1.5)	(1/4,3)	(1/4,6)			
7	(1/7,0)	(1/7,3)	(1/7,6)	(1/7,9)	(1/7,12)	(1/7,15)	(1/7,18)
8	(1/7,0)	(1/7,1.5)	(1/7,3)	(1/7,4.5)	(1/7,6)	(1/7,7.5)	(1/7,9)
9	(1/7,0)	(1/7,1.5)	(1/7,3)	(1/7,4.5)	(1/7,6)	(1/7,9.5)	(1/7,12.5)
10	(1/7,0)	(1/7,1.5)	(1/7,3)	(1/7,4.5)	(1/7,9)	(1/7,10.5)	(1/7,12)

图?? 展示所有高斯混合模型的密度函数. 可以看出, 当且仅当相邻两个成分的均值之差大于 2σ 时, 才能在总体密度中直观地区分开来. 参考 Ishwaran 等

(2001), 我们称密度函数中“峰”的个数为众数 (mode). 因此, 众数少于成分数的混合模型的参数估计是相对复杂问题.

参考 Chen 和 Khalili (2008) 的做法, 我们对各参数的初始值设定如下: 混合比例 $\pi_k^{(0)} = 1/K$, 成分均值 $\mu_k^{(0)} = 100(k-1/2)/K\%$ 样本分位数 $(k = 1, \cdots, K)$. 对于 σ 未知的情形, 借鉴 Xu 和 Chen (2015), 成分标准差被统一设定为上、下四分位数之间的样本标准差. 值得一提的是, 仅 EM 算法要求设定混合比例的初始值. 在算法的执行过程中, 当 $\left\|\boldsymbol{\pi}^{(new)} - \boldsymbol{\pi}^{(old)}\right\|_1 < 10^{-5}$, $\left\|\boldsymbol{\mu}^{(new)} - \boldsymbol{\mu}^{(old)}\right\|_1 < 10^{-5}$ 且 $\left\|\boldsymbol{\sigma}^{(new)} - \boldsymbol{\sigma}^{(old)}\right\|_1 < 10^{-5}$(其中 $\|\cdot\|_1$ 是 l_1 范数) 时, 迭代终止, 认为算法已达收敛. 参考 Chen 和 Khalili (2008), 模型 1 至模型 6 采用样本量 $n = 100$, 而模型 7 至模型 10 采用 $n = 400$.

表?? 展示了所有算法 500 次模拟的平均迭代次数及相应的标准差. 从表中可以看出, 无论是收敛速度还是稳定性, 三种改进算法均优于传统的 EM 算法. 特别地, 在那些众数少于成分数的混合模型中, 新方法达到同样收敛准则的迭代次数仅有 EM 算法的 2%-33%. 为了直观地展现该过程, 我们画出了第 10 个高斯混合模型在 σ 未知的情形下, 四种算法的迭代路径, 如图?? 所示. 可以看出, 新方法 1、2、3 分别需要 75、1058、697 次迭代, 而 EM 算法则需要 2738 次迭代才能达到收敛;

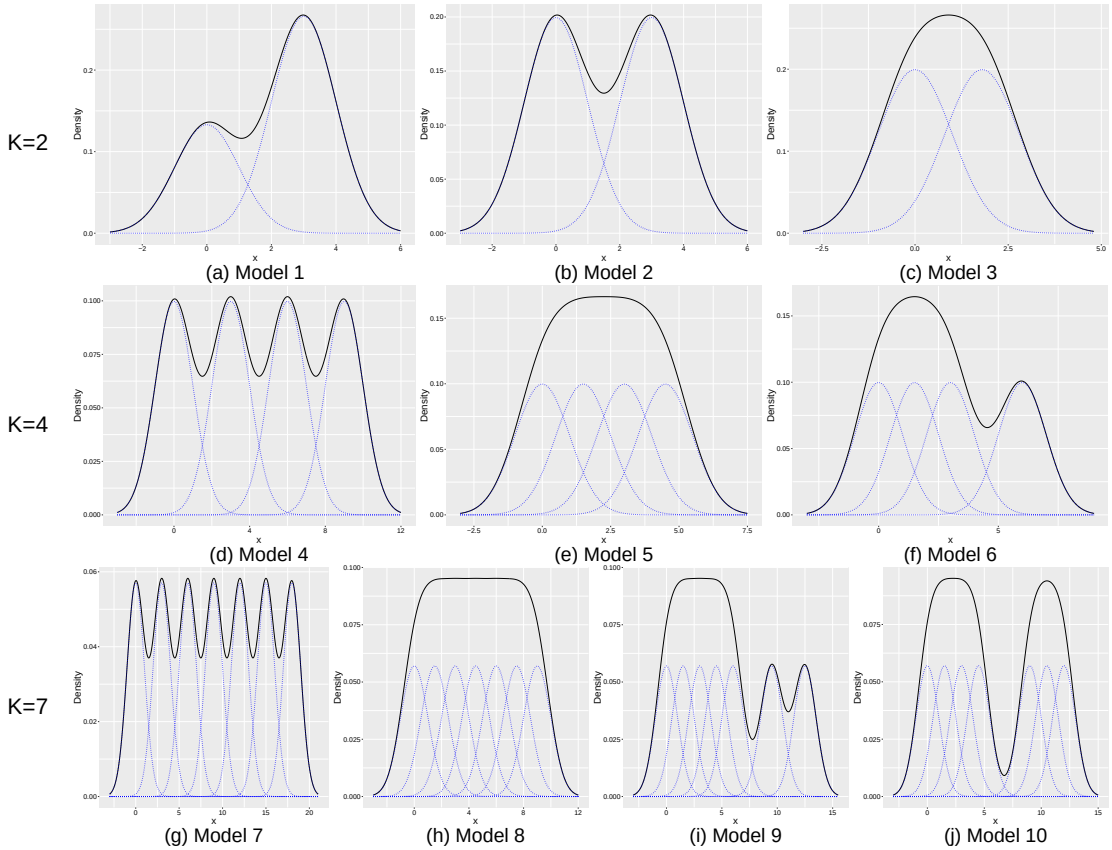


图 2-1: 高斯混合模型的密度函数. 图中实线为混合模型总体的密度函数, 虚线为各成分根据混合比例 π_k 放缩后的密度函数.

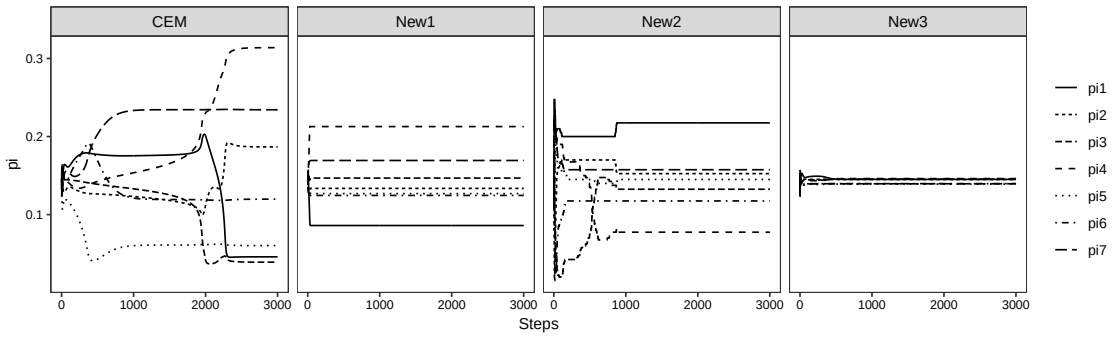
相比传统的 EM 算法, 本文所提出的新方法的收敛路径变化急剧, 并且需要的迭代次数明显小得多.

表 2-2: 各种算法对高斯混合模型进行参所估计的收敛速度.

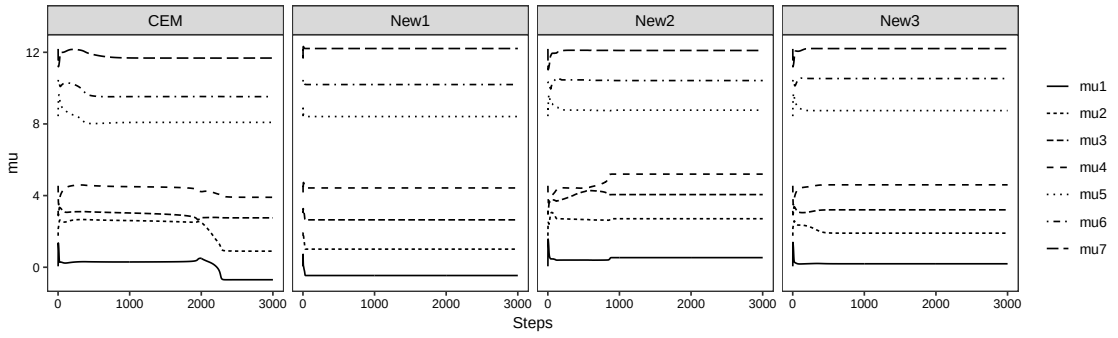
模型	n	成分数	众数	EM 算法	新方法 1	新方法 2	新方法 3
				方差已知情形			
1	100	2	2	26.9(8.7)	15.4(4.9)	15.9(3.5)	20.2(5.7)
2	100	2	2	20.9(6.5)	11.0(2.9)	12.3(2.6)	14.9(3.4)
3	100	2	1	118.7(93.3)	14.8(3.1)	13.1(3.3)	18.8(4.1)
4	100	4	4	65.7(55.9)	31(21.8)	37.6(16.6)	53.0(44.2)
5	100	4	1	829.0(1061.6)	55.5(40.5)	140.4(188.7)	262(355.8)
6	100	4	2	707.7(1137.1)	49.3(36.5)	99.3(86.3)	214.4(357.9)
7	400	7	7	88.4(59.7)	46.6(24.6)	63.5(22.8)	79.2(38.7)
8	400	7	1	2515.6(2802.0)	148.7(109.8)	361.3(592.5)	714.2(659.1)
9	400	7	3	2172.4(2840.9)	110.8(86.7)	201.1(181.2)	426.1(457.5)
10	400	7	2	2105.5(2550.7)	60.6(21.4)	154.7(126.9)	263.5(363.9)
				方差未知情形			
1	100	2	2	176.2(292.4)	30.2(25.8)	35.5(12.4)	72.8(42.9)
2	100	2	2	119.6(121.4)	23.2(13.4)	29.2(8.6)	58.5(25.4)
3	100	2	1	777.7(1313.0)	26.0(18.4)	87.1(324.0)	101.4(263.1)
4	100	4	4	490.7(703.6)	44.1(33.4)	93.7(71.8)	178.7(131.0)
5	100	4	1	1186.0(1453.1)	56.5(38.3)	173.4(209.3)	339.9(377.9)
6	100	4	2	1010.7(1129.3)	61.3(102.4)	178.9(210.3)	294.7(334.8)
7	400	7	7	2615.6(2527.5)	71.5(61.7)	553.9(491.2)	780.4(812.0)
8	400	7	1	6025.9(3791.6)	104.2(71.4)	1347.4(1575.4)	1986.8(1787.7)
9	400	7	3	5401.0(4098.8)	92.6(64.3)	956.9(803.0)	1571.2(1447.2)
10	400	7	2	5242.1(4132.0)	99.9(72.6)	1089.1(1255.2)	1443.1(1526.6)

注: 众数为模型对应的密度函数图像中“峰”的个数. 圆括号前的数代表 500 次模拟的平均迭代次数, 圆括号内为相应的标准差.

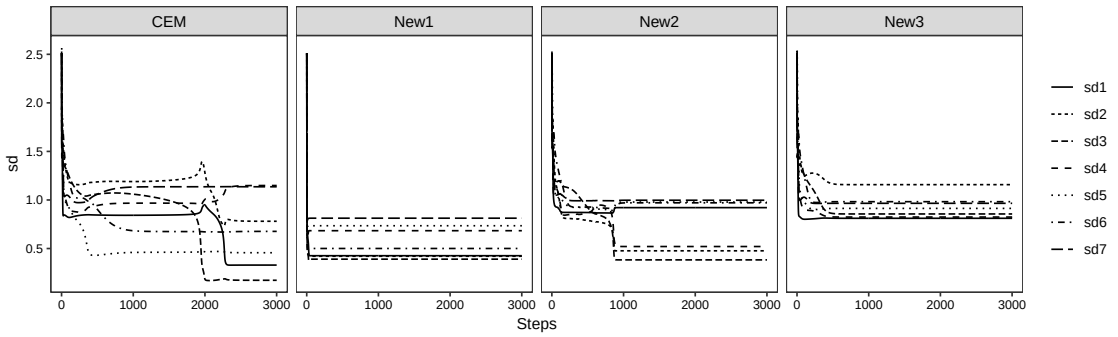
图**??** 至图**??** 展示了各算法对混合比例 π_k 及成分参数 μ_k、σ_k 的估计情况. 在模型 1 中, 传统 EM 算法估计效果最好, 但新方法与真实参数值相差不多. 在模型 2 和模型 3 中, 所有模型都估计得很好, 其中新方法的估计值方差更小、更加稳定. 在四成分的混合模型, 即模型 4 至模型 6 中, 新方法对参数的估计效果明显优于传统 EM 算法, 特别是在模型众数小于 4 的情况 (模型 5 和模型 6) 下, 优势更显然. 同样地, 在模型 7 至模型 10 中, 新方法远远好于 EM 算法, 并且值得一提的是, 如图**??**(s) 和 (t) 所示, 在第 9、第 10 两个高斯混合模型且 σ 未知的情形下, 新方法 2、3 对于成分标准差的估计更靠近真实值, 而 EM 算法则偏向于低估该值. 综上不同高斯混合模型的结果, 可以得出结论: 新方法的性能优于传统的 EM 算法, 特别是在众数小于成分数的复杂混合模型中.



(a) mixing proportions



(b) component means



(c) component standard deviations

图 2-2: 四种算法对七成分高斯混合模型 (表**??**中的模型 10) 进行参数估计的迭代路径. 其中 “CEM”、“New1”、“New2” 及 “New3” 分别指经典的 EM 算法、新方法 1、新方法 2 及新方法 3. 图 (a)、(b)、(c) 分别是混合比例、成分均值及成分标准差, 且横坐标表示算法的迭代步数, 纵坐标表示参数估计值的大小.

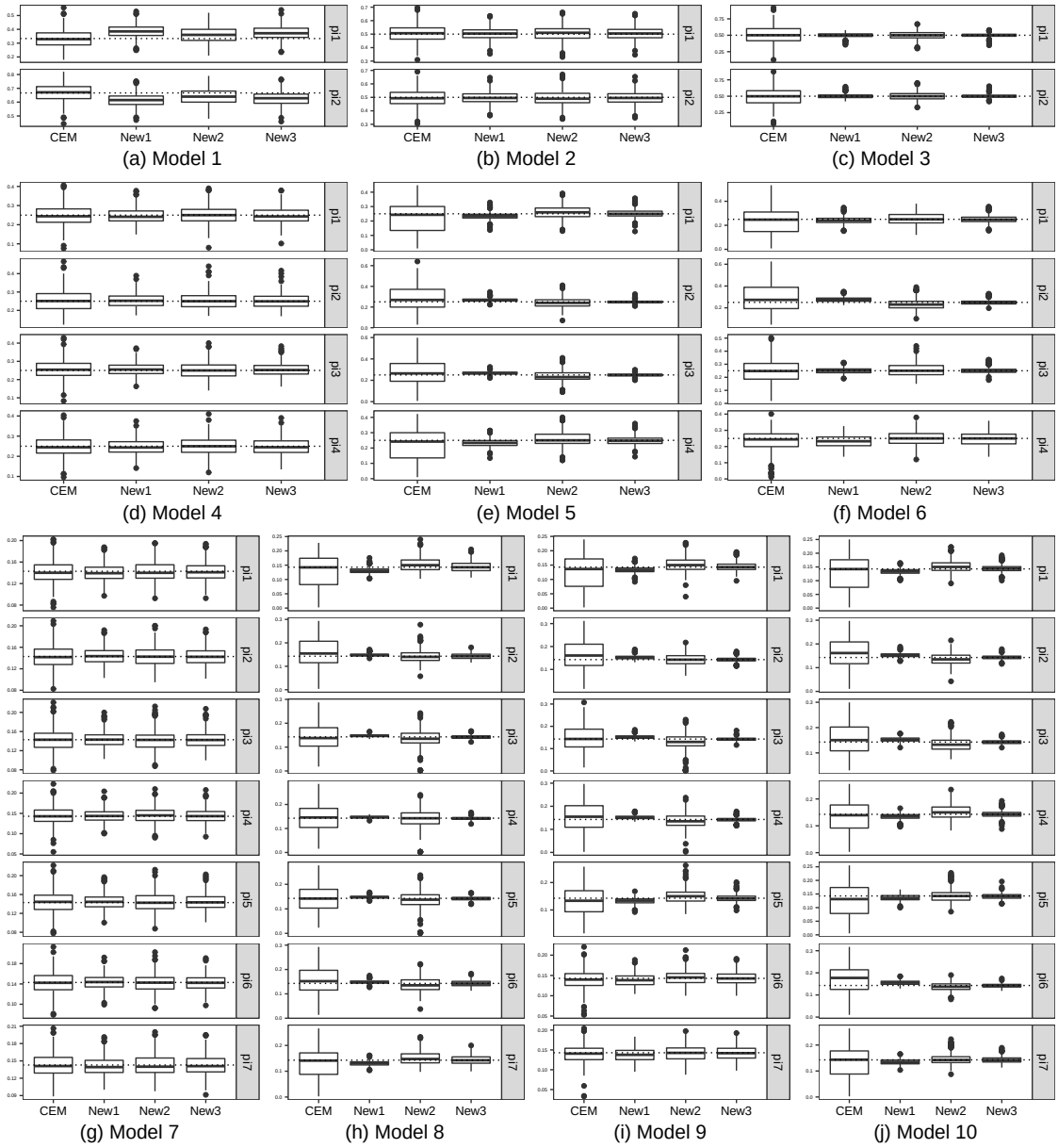


图 2-3: 方差已知情形下, 各算法对高斯混合模型的混合比例 π_k 的估计结果. "CEM"、"New1"、"New2" 及 "New3" 分别指经典的 EM 算法、新方法 1、新方法 2 及新方法 3. 图中虚线表示参数的真实值, 箱形图显示了 500 次蒙特卡洛模拟的参数估计结果.

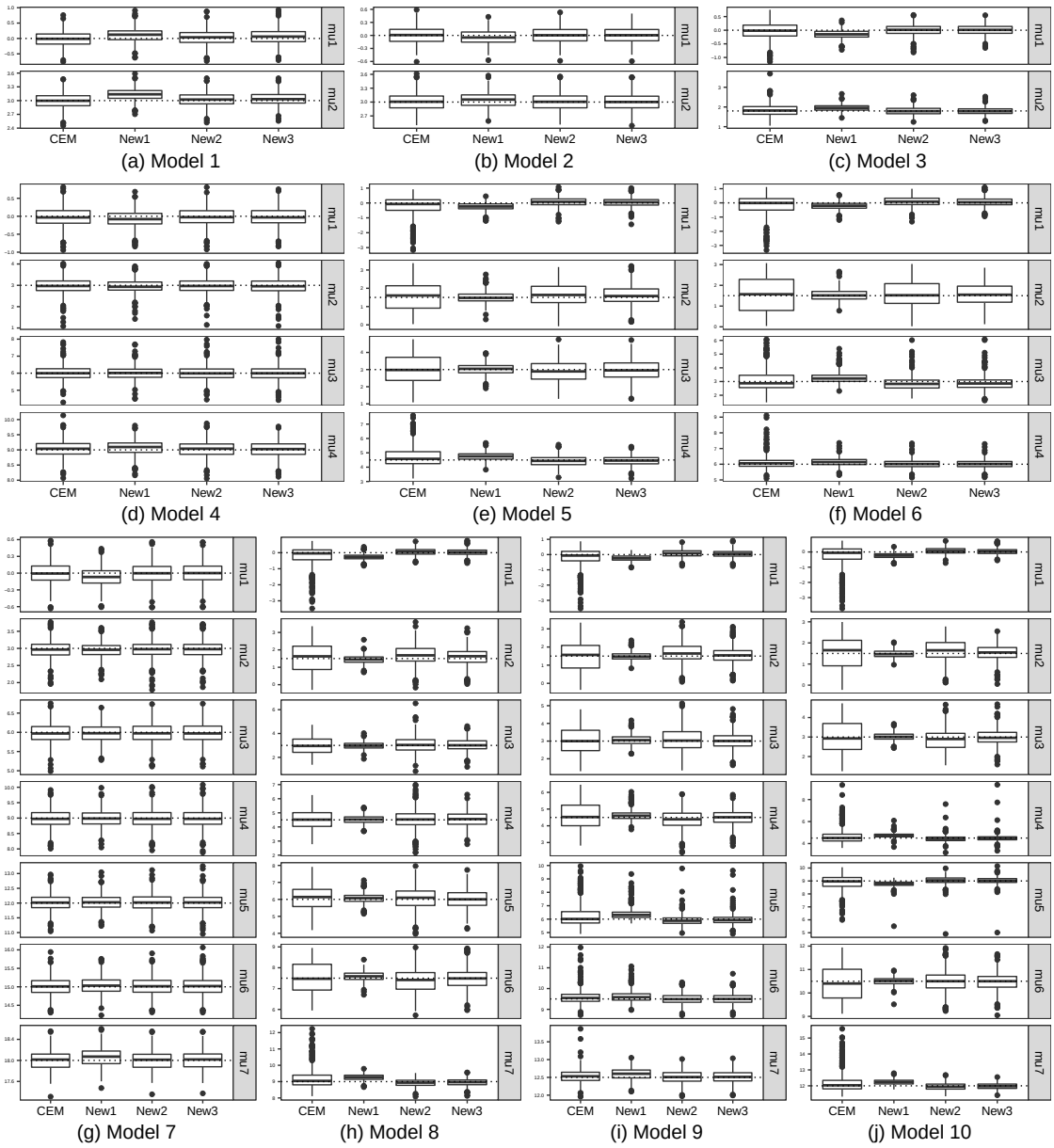


图 2-4: 方差已知情形下, 各算法对高斯混合模型的成分均值 μ_k 的估计结果.

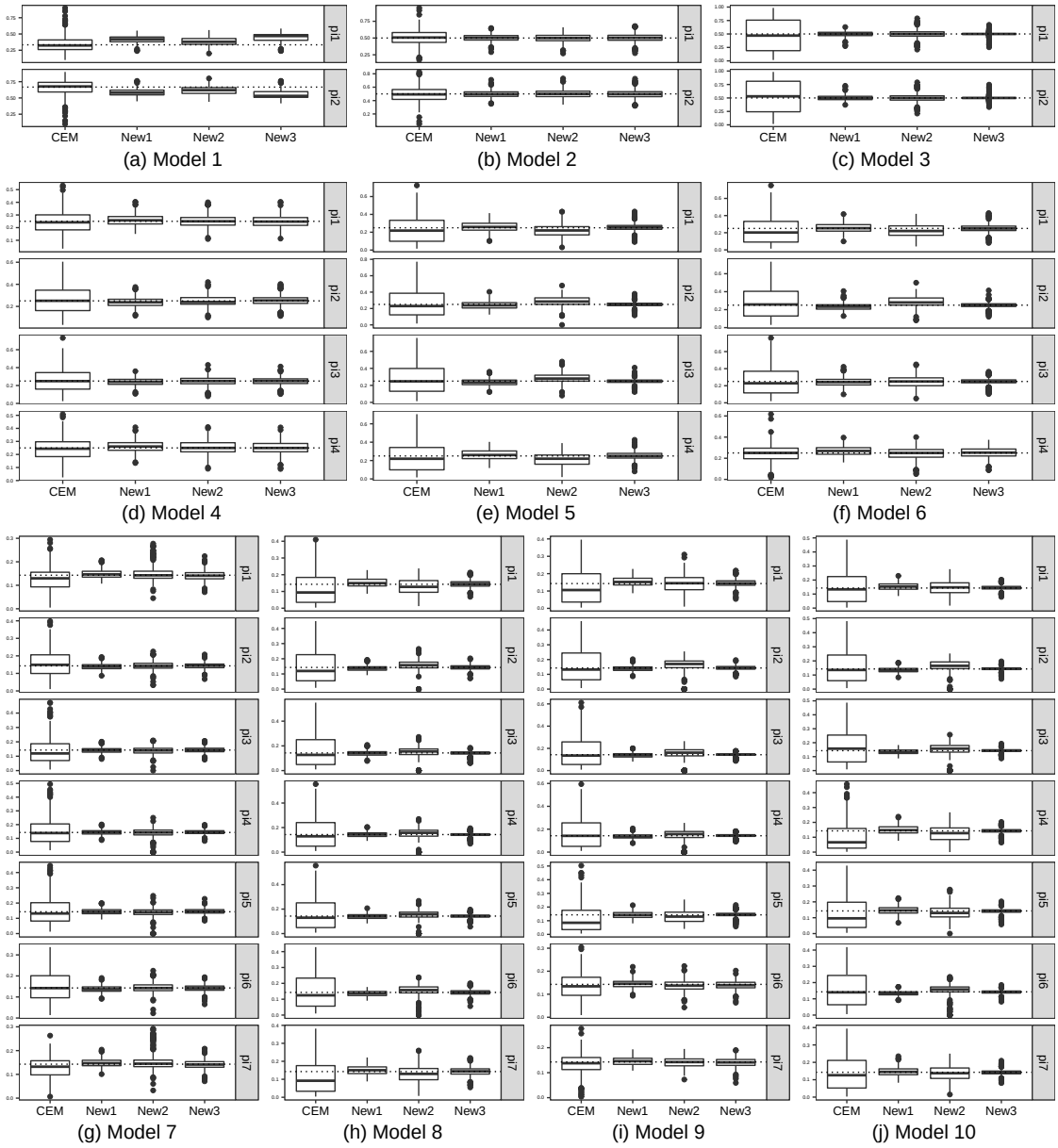


图 2-5: 方差未知情形下, 各算法对高斯混合模型的混合比例 π_k 的估计结果.

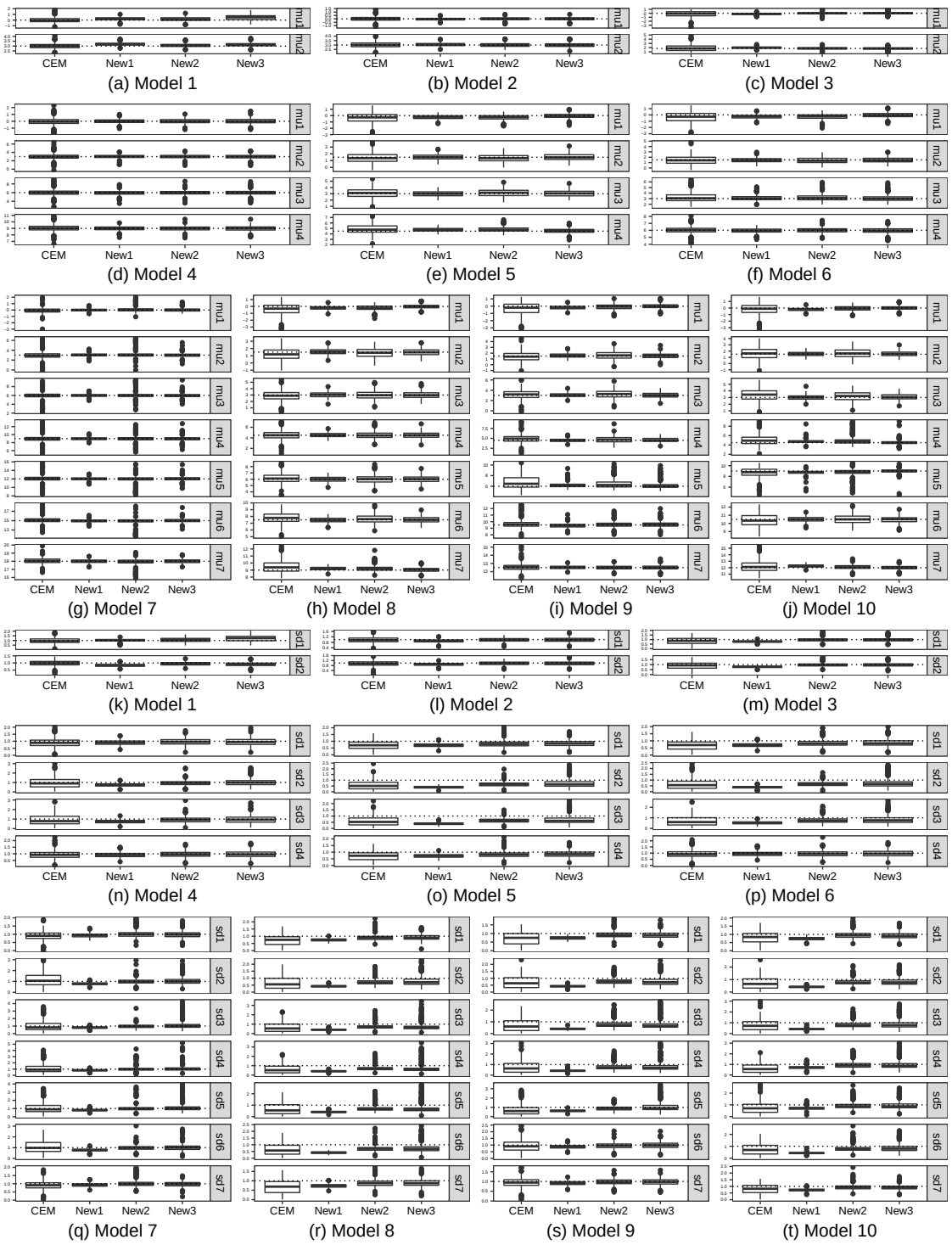


图 2-6: 方差未知情形下, 各算法对高斯混合模型的成分均值 μ_k ((a)-(j)) 及成分标准差 σ_k ((k)-(t)) 的估计结果.

例 2.2 对于泊松混合模型, 其密度函数如下

$$g(x;G)=\sum_{k=1}^{K}\pi_k\frac{\theta_k^x}{x!}\exp(-\theta_k).$$

我们从 Woo 和 Sriram (2007) 中选取了 7 个模型. 相应的参数设定可见表??. 对于每一个模型, 我们均模拟了 $n=100,500$ 两种不同样本量的情形, 模拟次数同样为 500 次.

表 2-3: 泊松混合模型参数设定.

模型	(π_1,θ_1)	(π_2,θ_2)	(π_3,θ_3)	(π_4,θ_4)
1	(0.5,1)	(0.5,9)		
2	(0.8,1)	(0.2,9)		
3	(0.95,1)	(0.05,10)		
4	(0.45,1)	(0.45,5)	(0.1,10)	
5	(1/3,1)	(1/3,5)	(1/3,10)	
6	(0.3,1)	(0.4,5)	(0.25,9)	(0.05,15)
7	(0.25,1)	(0.25,5)	(0.25,10)	(0.25,15)

表?? 展示了各种算法的收敛情况. 从表中可以看出, 随着成分数的增加, 即混合模型越复杂, 新方法的性能较传统 EM 算法的优势更加明显. 同样地, 箱线图?? 和?? 展示了不同方法对混合比例 π_k 和成分参数 θ_k 的估计情况. 在模型 1 至模型 4 中, 所有方法的估计效果均不理想, 而在模型 5 至模型 7 中, 新方法远远好于传统 EM 算法的估计效果.

从以上两种模拟例子可以看出, 相比传统的 EM 算法, 本文所提出的 3 中新方

表 2-4: 各种算法对泊松混合模型进行参所估计的收敛速度.

模型	n	EM 算法	新方法 1	新方法 2	新方法 3
1	100	24.2(6.3)	9.5(1.7)	11.5(1.3)	14.7(2.1)
	500	24.1(2.9)	9.7(0.8)	11.8(0.6)	15.1(1.0)
2	100	5.3(1.3)	2.6(0.6)	4.8(0.9)	4.9(1.0)
	500	5.2(0.5)	2.7(0.5)	4.9(0.4)	5.0(0.3)
3	100	27.5(33.6)	3.8(3.6)	6.0(1.5)	6.9(4.0)
	500	20.3(6.1)	3.0(0.3)	5.3(0.8)	6.1(0.8)
4	100	269.2(358.4)	39.2(24.6)	56.4(106.6)	99.4(196.3)
	500	251.5(193.0)	32.5(13.1)	30.5(9.9)	74.8(46.6)
5	100	346.9(439.3)	43.2(26.2)	53.6(130.6)	95.9(100.8)
	500	259.9(156.1)	40.0(13.9)	36.8(6.8)	86.3(35.1)
6	100	628.9(890.6)	84.6(176.2)	191.4(428.4)	289.1(443.2)
	500	1345.1(1518.5)	56.7(15.1)	147.6(293.0)	360.0(475.4)
7	100	869.3(1197.0)	62.3(34.7)	115.4(306.2)	232.9(390.2)
	500	1580.3(1491.7)	77.5(37.0)	58.0(30.0)	178.2(114.1)

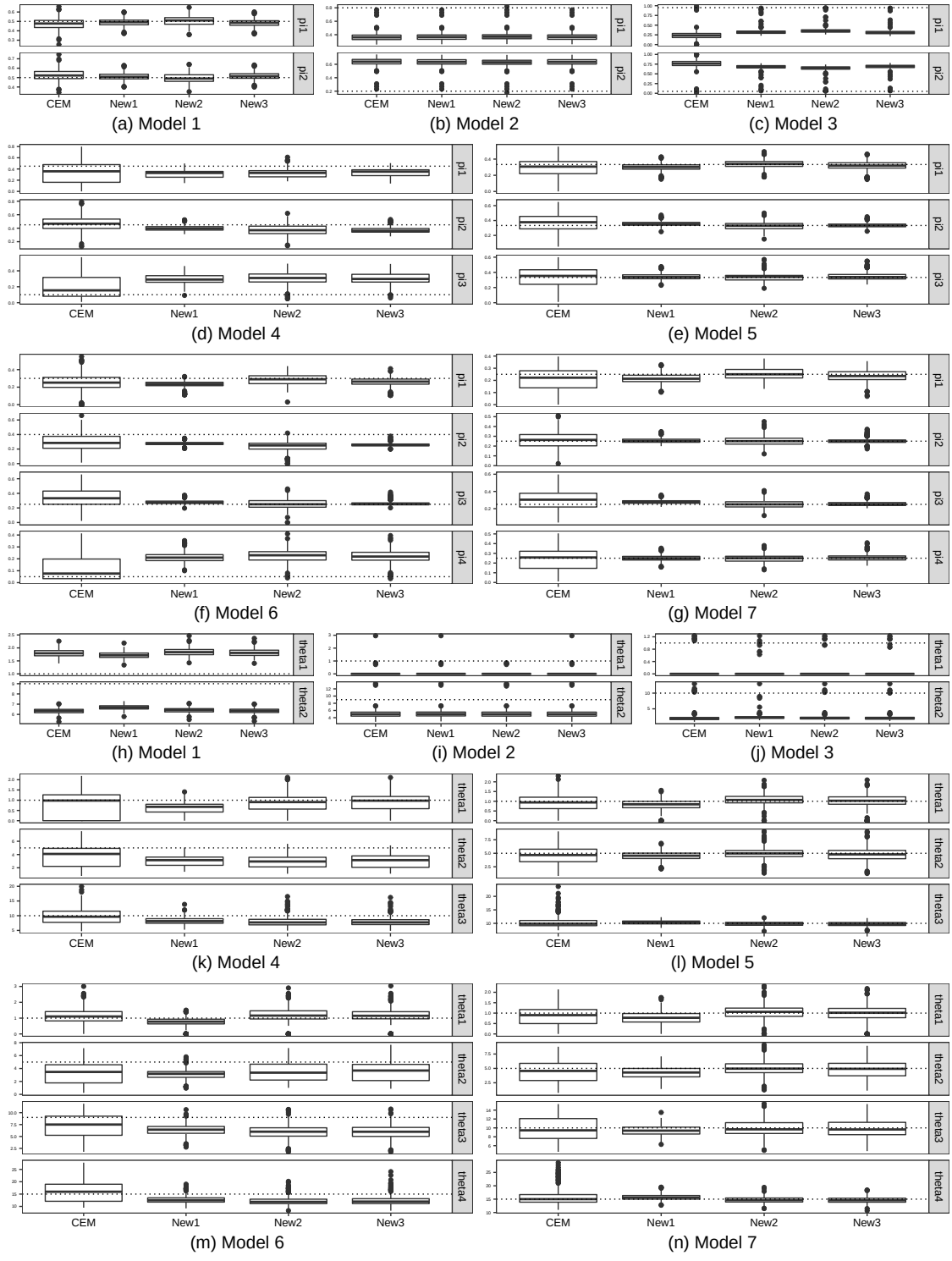


图 2-7: 样本量为 100 时, 各算法对泊松混合模型的混合比例 π_k ((a)-(g)) 及成分参数 θ_k ((h)-(n)) 的估计结果.

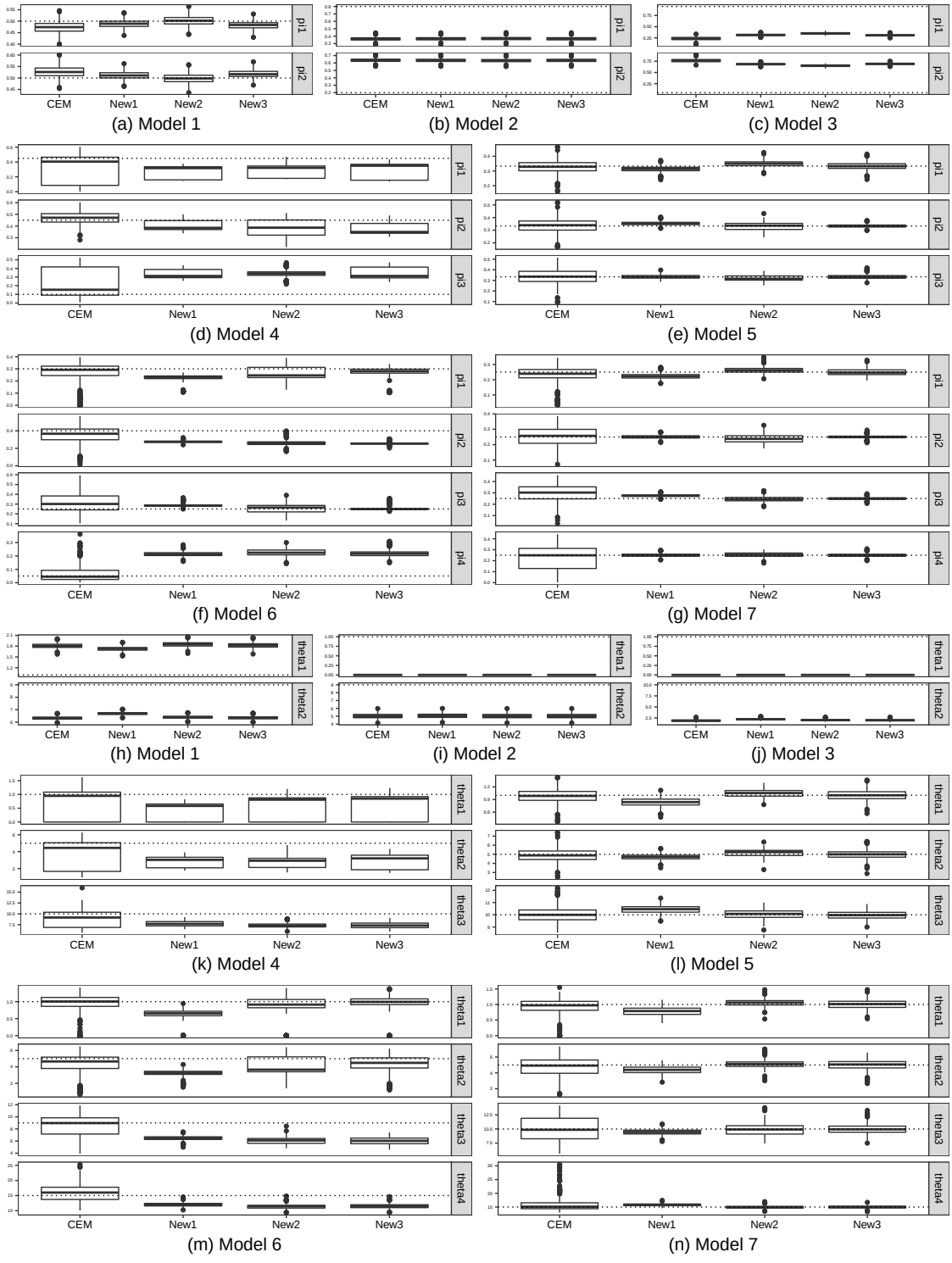


图 2-8: 样本量为 500 时, 各算法对泊松混合模型的混合比例 π_k ((a)-(g)) 及成分参数 θ_k ((h)-(n)) 的估计结果.

法明显减少了收敛所需的迭代次数, 特别是在多成分、少众数的复杂混合模型中. 当然, 新方法也存在改进之处, 如上面的泊松混合模型 1 至模型 4 中, EM 算法和改进的方法均不能取得令人满意的效果.

2.3.2 应用实例

下面我们将改进的 EM 算法应用到实际数据中.

例 2.3 (天文数据集) 我们现研究一个著名的天文数据集. 该数据集包含 82 个远离银河系的星系移动速度数据 (见直方图**??**). 星系移动速度的多模态, 可能意味着由大空隙包围的超星系团的存在, 并且其中的每个模态代表每个星系团以自身的速度移动 (Roeder, 1990). 因此, 我们可以用一个带共同成分方差的有限高斯混合分布来对观测数据进行拟合.

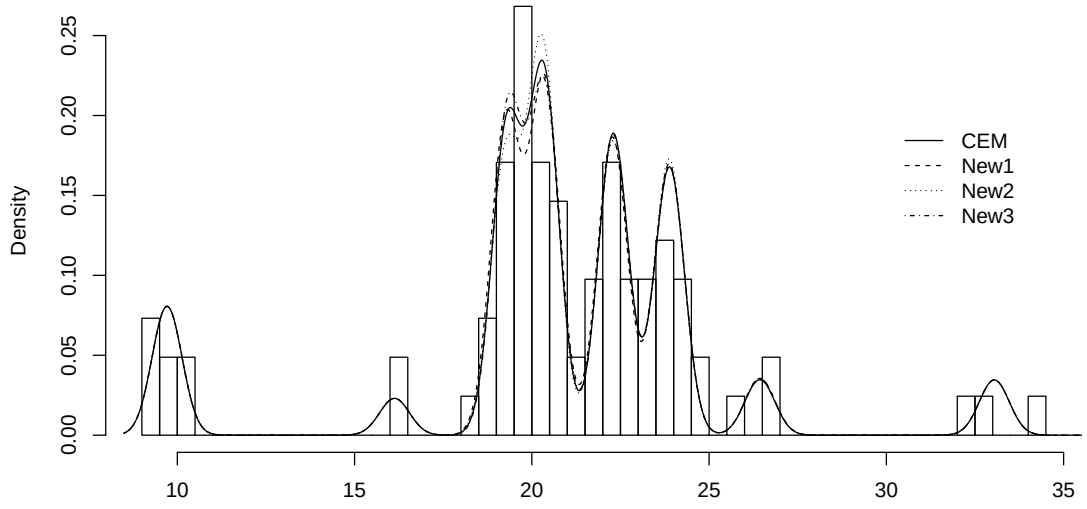


图 2-9: 星系数据直方图及模型拟合结果.

Richardson 和 Green (1997) 应用贝叶斯方法, 得出该混合分布的成分数在 5 到 7 之间; Xu 和 Chen (2015) 应用改进的 SCAD 方法 (modified SCAD, MSCAD), 得出该混合分布的成分数为 7. 他们的工作不仅证实了数据描述的星系上存在着超星系团, 也为接下来我们的信息准则法 (information-based methods) 提供成分数的可选范围.

针对 $K = 5, 6, \ldots, 11$ 的每种情形, 我们均利用传统的 EM 算法及本文所提出的三种改进方法对相应高斯混合分布的参数进行估计, 再利用 AIC 和 BIC 两种信息准则法决定 K 的取值. 算法运行的初始值、收敛准则设定同模拟研究一样.

表 **??** 给出了所有方法的模型结果. 显然, 相比传统的 EM 算法, 三种改进的方法都只需要更少的迭代次数就能达到收敛 (除了方法 2 中 $K = 5$ 的情形以及方

法 3 中 $K=7$ 的情形). 再者, 基于三种改进方法的信息准则均选择了八成分的高斯混合分布作为最终的模型, 模型参数如表 ?? 所示. 图 ?? 显示, 不同方法对该天文数据的模型分布拟合结果区别不大, 因此改进 EM 方法的模型精度是可以接受的.

表 2-5: 各算法对星系数据的估计结果.

K	EM 算法				新方法 1			
	迭代次数	对数似然	AIC	BIC	迭代次数	对数似然	AIC	BIC
5	48	-321.6	-335.6	-352.5	21	-338.9	-352.9	-369.8
6	20	-235.8	-252.8	-273.3	18	-244.9	-261.9	-282.4
7	77	-209.4	-229.4	-253.5	47	-222.7	-242.7	-266.7
8	72	-198.3	-221.3	-249.0	27	-198.4	-221.4	-249.1
9	106	-196.6	-222.6	-253.8	42	-197.3	-223.3	-254.6
10	186	-196.6	-225.6	-260.5	52	-197.0	-226.0	-260.9
11	222	-193.7	-225.7	-264.2	57	-194.5	-226.5	-265.0
K	新方法 2				新方法 3			
	迭代次数	对数似然	AIC	BIC	迭代次数	对数似然	AIC	BIC
5	153	-321.6	-335.6	-352.5	40	-321.6	-335.6	-352.5
6	18	-235.8	-252.8	-273.3	19	-235.8	-252.8	-273.3
7	69	-222.4	-242.4	-266.5	95	-222.5	-242.5	-266.6
8	31	-198.3	-221.3	-249.0	42	-198.3	-221.3	-249.0
9	40	-197.3	-223.3	-254.6	61	-197.3	-223.3	-254.5
10	59	-197.3	-226.3	-261.2	84	-196.7	-225.7	-260.6
11	84	-194.6	-226.6	-265.1	105	-193.9	-225.9	-264.4

表 2-6: 星系数据的参数估计结果. $\hat{\pi}_k$、$\hat{\mu}_k$ 及 $\hat{\sigma}$ 分别表示混合比例、成分均值及成分标准差 $(k=1,2,\ldots,8)$.

方法	$(\hat{\pi}_1,\hat{\mu}_1)$	$(\hat{\pi}_2,\hat{\mu}_2)$	$(\hat{\pi}_3,\hat{\mu}_3)$	$(\hat{\pi}_4,\hat{\mu}_4)$	$(\hat{\pi}_5,\hat{\mu}_5)$
EM 算法	(0.09,9.7)	(0.02,16.1)	(0.20,19.3)	(0.24,20.3)	(0.20,22.3)
新方法 1	(0.09,9.7)	(0.02,16.1)	(0.21,19.3)	(0.23,20.4)	(0.20,22.3)
新方法 2	(0.09,9.7)	(0.02,16.1)	(0.18,19.3)	(0.26,20.3)	(0.20,22.3)
新方法 3	(0.09,9.7)	(0.02,16.1)	(0.21,19.3)	(0.23,20.4)	(0.20,22.3)

方法	$(\hat{\pi}_6,\hat{\mu}_6)$	$(\hat{\pi}_7,\hat{\mu}_7)$	$(\hat{\pi}_8,\hat{\mu}_8)$	$\hat{\sigma}$
EM 算法	(0.18,23.9)	(0.04,26.4)	(0.04,33)	0.42
新方法 1	(0.18,23.9)	(0.04,26.4)	(0.04,33)	0.42
新方法 2	(0.18,23.9)	(0.04,26.4)	(0.04,33)	0.42
新方法 3	(0.18,23.9)	(0.04,26.4)	(0.04,33)	0.42

2.4 小结

本章研究了有限混合模型的参数估计问题. 首先, 我们介绍了经典 EM 算法在混合模型参数估计中的应用, 以及算法收敛速度慢的问题. 接着, 我们对 E 步中隐变量条件期望的计算方法进行修改, 提出了三种改进的 EM 算法: 第一种改进算法视上一次迭代的参数估计结果为先验信息, 将每个样本所属各个成分的可能性作为其混合比例的新估计值; 第二种改进算法的思想来源于有限混合模型的抽样过程. 我们根据每一次迭代后的参数估计值对样本进行分类, 用各个成分的样本比率来替换混合比例; 第三种改进算法将第一种算法中每个样本所属各个成分的可能性作平均, 得到混合比例的新估计值. 新算法和经典 EM 算法一样都具有收敛性, 但由于每一次迭代时, 新算法都会利用样本信息, 因此加快了收敛速度, 这在本章最后的模拟研究及实例应用中都得到了印证. 同时, 数值试验结果也展示了新算法对有限混合模型的参数估计效果更准确和稳定.

值得一提的是, 本章所展示的有限混合分布是一元的, 但无需额外工作量, 便可拓展到多元情形.

第 3 章　基于 MMCP 方法的有限混合模型定阶研究

本章将研究另一个问题, 有限混合模型的定阶. 由于所提出的新方法是在 AIC、BIC 等信息准则方法的基础上进行改动的, 因此我们首先在本章第 1 节中介绍相关工作, 其次在第 2 节中介绍新方法及其数值求解算法, 最后在第 3 节中用实验数据验证所提出 MMCP 方法的效果.

3.1　相关工作

3.1.1　基于信息准则法的有限混合模型定阶

AIC (Akaike, 1973) 和 BIC (Schwarz, 1978) 是最常见的两种模型选择方法. Leroux (1992) 将 AIC 和 BIC 应用到有限混合模型的定阶问题中.

假定有来自某有限混合分布的一组独立随机样本, 记对数似然函数为 $l_n(G)$. AIC 信息被定义为备选模型中参数个数的函数. 在有限混合模型中, 参数的个数与阶 K 成正比, 因此有

$$AIC(K) = l_n(\hat{G}) - a_K,$$

式中 $\hat{G}$ 是 K 成分混合模型的参数 G 的极大似然估计, a_K 是 K 成分混合模型中的参数总个数.

BIC 信息与 AIC 信息的表达式只有稍微不同, 为

$$BIC(K) = l_n(\hat{G}) - \frac{1}{2}(\log n)a_K,$$

式中 n 为样本个数.

在实际中, 我们需要在一个较大的参数空间 (如 $[0, 2K_0]$, 其中 K_0 为真实的成分个数) 中计算每一个 K 取值下的信息量, 最后取 AIC 或 BIC 最大时的 K 作为模型阶的估计. 由于 AIC 及 BIC 不是关于参数个数的单调递增函数, 因此所得到的模型在一定程度上能够防止过拟合. 但由于这两种方法都需要在一个较大的参数空间上进行搜索, 特别是当模型很复杂 (K_0 很大) 时, 需要消耗大量的计算时间.

3.1.2　基于惩罚最小距离法的有限混合模型定阶

Chen 和 Kalbfleisch (1996) 在 AIC 及 BIC 的思想上, 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的个数进

行惩罚, 而惩罚最小距离法则是对混合比例进行惩罚. 给定一有限混合模型的随机独立样本 $\{x_1, x_2, \cdots, x_n\}$, 并假设其累计分布函数为

$$C(x; G) = \sum_{k=1}^{K} \pi_k F(x, \theta_k).$$

记经验分布函数为 $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$; $d(F_n(x), C(x; G))$ 表示 $F_n(x)$ 和 $C(x; G)$ 两个函数之间的距离 (如 kolmogrov-Smirnov 距离、Cramer-VonMises 距离).

惩罚最小距离法中的距离函数为

$$D(F_n(x), C(x; G)) = d(F_n(x), C(x; G)) - C_K \sum_{k=1}^{K} \log \pi_k, \tag{3.1}$$

其中 $C_K > 0$ 为可调参数. 由上式可以看出, 如果某些 π_k 过小, 那么该距离函数的值就会很大, 因此该方法能够防止 K 估计值较大引起的过拟合情况, 然而, 该方法的计算程序和信息准则法一样, 需要在一个较大的参数空间里计算成分数 K 的所有可能取值对应的距离函数值, 再进行比较之后得出结论. 再者, 基于惩罚最小距离法对有限混合模型依旧不能防止另一种过拟合情况: 混合模型中存在相近成分, 即某些子群体的参数差别不大, 我们将在章节**??** 中具体研究这种过拟合情况并提出相应的解决方法.

3.1.3 回归模型中的变量选择

考虑线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

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

Tibshirani (1996) 提出了一种多元线性回归模型的变量选择方法, 即 LASSO. 该方法通过最小化惩罚残差平方和以估计回归系数

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

其中 $\lambda \geq 0$ 为可调参数. LASSO 方法能够同时实现变量选择和参数估计.

由于 LASSO 对所有回归系数都进行了同等程度的惩罚, 导致其估计结果是有

偏的. Fan 和 Li (2001) 对该问题进行了研究并提出了 SCAD 方法, 即

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

其中

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

Zhang (2010) 提出了另一种非凸罚函数 (nonconvex penalty functions), MCP 罚函数. 该方法通过下式求得回归系数的近似无偏估计

$$\widehat{\boldsymbol{\beta}}_{MCP} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_{MCP}(\beta_j),$$

其中

$$p'_{MCP}(\beta_j) = (\gamma - \frac{|\beta_j|}{a})_+, \gamma \geq 0, a > 1.$$

MCP 和 SCAD 对于系数绝对值大于 γa 的回归变量不做压缩, 因此在一定程度上弥补了 LASSO 有偏估计的不足.

以上回归模型中的变量选择方法, 直观上与有限混合模型的定阶问题无关. 而如前文所述, 信息准则法和惩罚最小距离法计算繁琐, 且虽然后者能够有效防止出现混合模型中成分的混合比例过小的情况, 但依旧无法避免某些成分参数相近这一过拟合的情况. Chen 和 Khalili (2008) 提出的 MSCAD 方法, 把回归模型中 SCAD 方法的变量选择思想巧妙地应用到了混合模型的定阶问题中. 通过引入一个关于成分参数距离的惩罚函数, 将相近的成分进行合并, 从而得到一个较低的阶数. 值得一提的是, MSCAD 方法的数值求解算法很复杂, 而且根据 Zhang (2010)、Breheny 和 Huang (2011) 的工作, MCP 对回归系数的估计效果都优于 SCAD, 因此下一章节将对 MSCAD 方法的惩罚项进行改进, 提出 MMCP 方法.

3.2 MMCP 方法的理论与算法

有限混合模型的定阶问题, 即是模型复杂度与拟合效果之间的权衡问题. 经典的 AIC 方法 (Akaike, 1973) 和 BIC 方法 (Schwarz, 1978) 是被使用最广泛的两种方法. 这些方法在对数似然函数的基础上, 考虑模型参数个数对拟合优度的影响. 而惩罚最小距离法 (Chen 和 Kalbfleisch, 1996) 则对混合比例进行惩罚. 在本节中,

我们将对 Chen 和 Khalili (2008) 提出的惩罚似然方法进行改进. 与其他方法不同, 这种方法通过引入惩罚项到似然函数中, 以防止两种类型的过拟合. 特别地, 该方法引入一个非光滑惩罚项来合并有限混合模型中相近的成分, 于是稀疏回归模型的思想被恰到好处地应用于有限混合模型的定阶中. 通过选择恰当的罚函数, 该方法能够同时实现定阶及成分参数的估计.

3.2.1 MMCP 方法的理论

重述一下所研究的问题: 对于具有如下概率密度函数的有限混合模型

$$g(x;G)=\int_{\Theta} f(x;\theta)d\Psi(\theta), \tag{3.3}$$

其中参数 $G=(K,\pi_1,\pi_2,\cdots,\pi_K,\theta_1,\theta_2,\cdots,\theta_K)$, 混合率 $\Psi(\theta)=\sum_{k=1}^{K}\pi_k I(\theta_k\le\theta)$, 且 $\sum_{k=1}^{K}\pi_k=1$, $\pi_k\ge 0$ $(k=1,2,\ldots,K)$, 给定随机独立样本 $\{x_1,x_2,\cdots,x_n\}$, 估计成分个数 K(定阶)、混合比例 π_k 以及成分参数 θ_k. 其中, 定阶是本章研究的重点.

注意到最大化对数似然函数 (**??**) 时, 无法对 K 求导. 尽管模型成分个数未知, 但一般地, 我们可以基于某些信息对其取值范围进行限定. 不妨假定存在一个 K, 使得 $K\ge K_0$(K_0 为真实的成分个数), 这样便可利用第**??**章所研究的极大似然估计方法来对混合比例 π_k 以及成分参数 θ_k 进行估计, 即

$$\hat{G}=\arg\max_{G}\sum_{i=1}^{n}\log\{\sum_{k=1}^{K}\pi_k f(x_i;\theta_k)\}. \tag{3.4}$$

但上述的极大似然法往往会导致两种类型的过拟合:

- 过拟合类型 1: 所拟合的模型中, 某些成分的混合比例 π_k 非常小, 这是由于在成分个数 $K(K\ge K_0)$ 下做参数估计会引入多余的成分;
- 过拟合类型 2: 所拟合的模型包含 “相似” 的成分, 即 $\exists m,n\in\{1,2,\ldots,K\}$, 使得 $\|\theta_m-\theta_n\|$ 非常小, 说明在某真实成分的附近, 出现多个成分参数 θ_k 差别很小的成分.

Chen 和 Khalili (2008) 提出了改进的 SCAD 方法 (modified SCAD, MSCAD): 通过对成分混合比例和成分参数之间的距离施加惩罚, 以减少模型的复杂性. 具体地, 对于预先设定的 $K(K\ge K_0)$, 将各成分参数按大小进行排序 $\theta_1\le\theta_2\le\ldots\le$

θ_K. 记 $\eta_k = \theta_{k+1} - \theta_k$, $k = 1, 2, \ldots, K-1$. 定义如下惩罚对数似然函数

$$\tilde{l}_n(G) = l_n(G) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_{SCAD}(\eta_k), \tag{3.5}$$

其中

$$\begin{aligned} l_n(G) &= \sum_{i=1}^{n} \log\{\sum_{k=1}^{K} \pi_k f(x_i; \theta_k)\}, \\ p'_{SCAD}(\eta) &= \gamma\sqrt{n} I(\sqrt{n}\eta \le \gamma) + \frac{\sqrt{n}(a\gamma - \sqrt{n}\eta)_+}{a-1} I(\sqrt{n}\eta > \gamma), \end{aligned}$$

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

$$\mathcal{M}_K = \{G : \theta_1 \le \theta_2 \le \ldots \le \theta_K, \sum_{k=1}^{K} \pi_k = 1, \pi_k \ge 0\}$$

上, 极大化惩罚对数似然函数 (**??**), 可同时对有限混合模型进行定阶及成分参数的估计. 特别地, 惩罚项 $\log \pi_k$ 能够防止第一类过拟合, 惩罚项 $p_{SCAD}(\cdot)$ 则能够起到防止第二类过拟合的作用.

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的变量选择问题. 这里, 我们用 MCP 罚 (Zhang, 2010) 来代替该项, 即

$$p'_{MCP}(\eta) = \sqrt{n}(\gamma - \frac{\sqrt{n}\eta}{a})_+,$$

其中 $\gamma \ge 0$, $a > 1$.

图**??** 展示了两种罚函数的图像. 可以看出, 两种罚函数对较大的变量均不做压缩. MCP 罚与 SCAD 罚的不同之处在于, 对于那些小的变量, MCP 对其惩罚是一个严格的递减导数, 这进一步防止了模型的第二类过拟合. 而通过数值模拟我们也发现, 在包含相近成份 (即子总体的参数相近) 的混合模型中, SCAD 会倾向于把这些相近的成份合并, 导致过渡正则化; 相反, MCP 则能更好地对模型进行定阶.

总结所提出的 MMCP 方法: 定义惩罚对数似然函数

$$\tilde{l}_n(G) = l_n(G) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_{MCP}(\eta_k), \tag{3.6}$$

其中 $\log \pi_k$ 对混合比例进行约束, $p_{MCP}(\eta_k)$ 对成分的参数之间的距离进行约束.

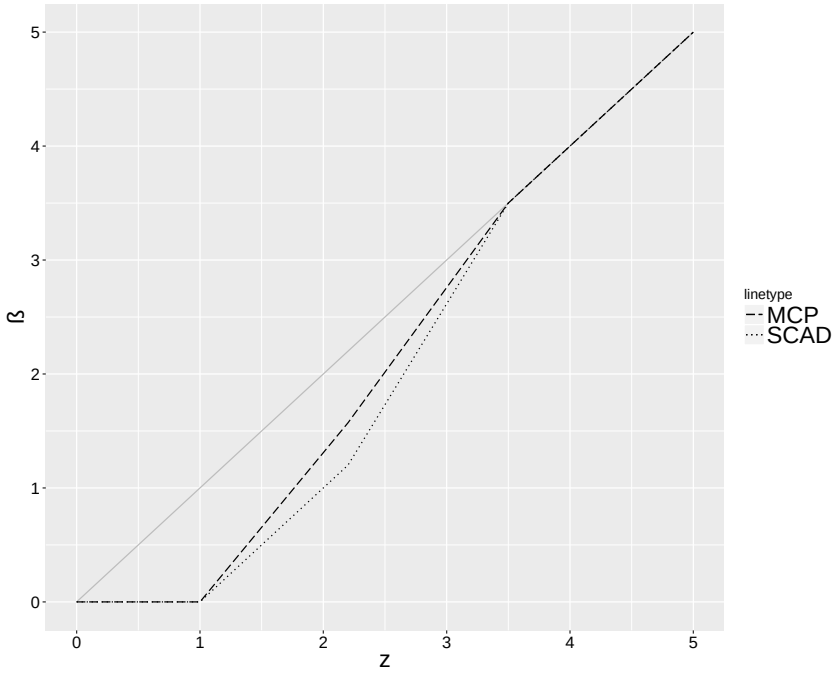


图 3-1: 两种罚函数的图像.

因此, 通过极大化上式, 可同时实现有限混合模型的定阶及参数估计.

3.2.2 MMCP 方法的数值求解

下面, 我们对 EM 算法进行适当修改, 以极大化惩罚对数似然函数 (??).

记 $G = (\pi_1, \pi_2, \ldots, \pi_K, \theta_1, \theta_2, \ldots, \theta_K)$, G_0 为真实参数值. 给定随机独立样本 $\{x_1, x_2, \cdots, x_n\}$, 相应地, 可引入隐变量 $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$, 指示每个样本所属的成分, 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 $\sum_{k=1}^{K} z_{ik} = 1$. 根据所给样本及其属性 (组成的完整数据), 可得惩罚完整对数似然函数

$$\tilde{l}_n^c(G) = l_n^c(G) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_{MCP}(\eta_k), \tag{3.7}$$

其中, $l_n^c(G)$ 如式 (??) 定义.

E 步中, 在给定观测样本和当前参数估计值 $G^{(m)}$ 后, 计算 $\tilde{l}_n^c(G)$ 关于 z_{ik} 的条件期望, 即

$$\begin{aligned} Q(G; G^{(m)}) = & \sum_{i=1}^{n} \sum_{k=1}^{K} E[z_{ik}|x_i; G^{(m)}] \log\{f(x_i; \theta_k)\} - \sum_{k=1}^{K-1} p_{MCP}(\eta_k) \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} \{E[z_{ik}|x_i; G^{(m)}] + \frac{C_K}{n}\} \log \pi_k, \end{aligned} \tag{3.8}$$

其中,

$$E[z_{ik}|x_i;G^{(m)}]=\frac{\pi_k^{(m)}f(x_i;\theta_k^{(m)})}{\sum_{j=1}^{K}\pi_j^{(m)}f(x_i;\theta_j^{(m)})}. \tag{3.9}$$

M 步为求解 $G^{(m+1)}=\arg\max_G Q(G;G^{(m)})$. 注意到关于混合比例 π_k, 有限制条件 $\sum_{j=1}^{K}\pi_j=1$, 应用拉格朗日乘子法, 有

$$\frac{\partial[Q(G;G^{(m)})+\lambda(\sum_{j=1}^{K}\pi_j-1)]}{\partial\pi_k}=0.$$

可得

$$\pi_k^{(m)}=\frac{\sum_{i=1}^{n}E[z_{ik}|x_i;G^{(m)}]+C_K}{n+KC_K},k=1,2,\cdots,K.$$

由于 $p_{MCP}(\eta)$ 不光滑, 无法使用 Newton-Raphson 方法来关于 θ_k 极大化 $Q(G;G^{(m)})$. 这里, 参考 Fan 和 Li (2001) 的做法, 将 $p_{MCP}(\eta)$ 近似替换为

$$\tilde{p}_{MCP}(\eta;\eta_k^{(m)})=p_{MCP}(\eta_k^{(m)})+\frac{p'_{MCP}(\eta_k^{(m)})}{2\eta_k^{(m)}}(\eta^2-\eta_k^{(m)2}). \tag{3.10}$$

于是, 问题转化为求解如下方程组

$$\begin{cases}\sum_{i=1}^{n}E[z_{i1}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_1}\log f(x_i;\theta_1)-\frac{\partial\tilde{p}_{MCP}(\eta_1;\eta_1^{(m)})}{\partial\theta_1}=0,\\ \sum_{i=1}^{n}E[z_{ik}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_k}\log f(x_i;\theta_k)-\frac{\partial\tilde{p}_{MCP}(\eta_{k-1};\eta_{k-1}^{(m)})}{\partial\theta_k}-\frac{\partial\tilde{p}_{MCP}(\eta_k;\eta_k^{(m)})}{\partial\theta_k}=0,\\ \qquad\qquad k=2,3,\cdots,K-1,\\ \sum_{i=1}^{n}E[z_{iK}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_K}\log f(x_i;\theta_K)-\frac{\partial\tilde{p}_{MCP}(\eta_{K-1};\eta_{K-1}^{(m)})}{\partial\theta_K}=0.\end{cases}$$

将式 (**??**) 带入上述方程组中, 可得

$$\begin{cases}\sum_{i=1}^{n}E[z_{i1}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_1}\log f(x_i;\theta_1)+\frac{\partial}{\partial\theta_1}\{\frac{p'_{MCP}(\eta_1^{(m)})}{2\eta_k^{(m)}}(\theta_2-\theta_1)^2\}=0,\\ \sum_{i=1}^{n}E[z_{ik}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_k}\log f(x_i;\theta_k)-\frac{\partial}{\partial\theta_k}\{\frac{p'_{MCP}(\eta_{k-1}^{(m)})}{2\eta_{k-1}^{(m)}}(\theta_k-\theta_{k-1})^2\}\\ \qquad+\frac{\partial}{\partial\theta_k}\{\frac{p'_{MCP}(\eta_k^{(m)})}{2\eta_k^{(m)}}(\theta_{k+1}-\theta_k)^2\}=0,k=2,3,\cdots,K-1,\\ \sum_{i=1}^{n}E[z_{iK}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_K}\log f(x_i;\theta_K)-\frac{\partial}{\partial\theta_K}\{\frac{p'_{MCP}(\eta_{K-1}^{(m)})}{2\eta_{K-1}^{(m)}}(\theta_K-\theta_{K-1})^2\}=0.\end{cases}$$

整理, 可得

$$\begin{cases} \sum_{i=1}^{n} E[z_{i1}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_1}\log f(x_i;\theta_1) - \frac{p'_{MCP}(\eta_1^{(m)})}{\eta_k^{(m)}}\theta_2 + \frac{p'_{MCP}(\eta_1^{(m)})}{\eta_k^{(m)}}\theta_1 = 0, \\ \sum_{i=1}^{n} E[z_{ik}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_k}\log f(x_i;\theta_k) - \frac{p'_{MCP}(\eta_{k-1}^{(m)})}{\eta_{k-1}^{(m)}}\theta_k + \frac{p'_{MCP}(\eta_{k-1}^{(m)})}{\eta_{k-1}^{(m)}}\theta_{k-1} \\ \qquad -\frac{p'_{MCP}(\eta_k^{(m)})}{\eta_k^{(m)}}\theta_{k+1} + \frac{p'_{MCP}(\eta_k^{(m)})}{\eta_k^{(m)}}\theta_k = 0, k = 2,3,\cdots,K-1, \\ \sum_{i=1}^{n} E[z_{iK}|x_i;G^{(m)}]\frac{\partial}{\partial\theta_K}\log f(x_i;\theta_K) - \frac{p'_{MCP}(\eta_{K-1}^{(m)})}{\eta_{K-1}^{(m)}}\theta_K + \frac{p'_{MCP}(\eta_{K-1}^{(m)})}{\eta_{K-1}^{(m)}}\theta_{K-1} = 0. \end{cases}$$

以上方程组包含 K 个方程, 求解可得模型的 K 个成分参数 θ_k, $k = 1, 2, \cdots, K$.

于是, 任意给定一个参数初始值, $G^{(0)}$. 重复上述 E 步和 M 步, 直到参数估计序列收敛, 便可得到参数 $G = (\pi_1, \pi_2, \ldots, \pi_K, \theta_1, \theta_2, \ldots, \theta_K)$ 的估计值, 我们只需将那些 θ_k 相同的成分对应的混合比例 π_k 进行合并, 即可同时完成有限混合模型的定阶与成分参数估计.

下面阐述可调参数的选择问题. 关于 C_K, Chen 等 (2001) 指出如果所有成分参数 θ_k 处于 $[-M, M]$ 或 $[M^{-1}, M](M$ 为某个较大的正数) 内, 一个合适的取值便是 $C_K = \log M$. 这里, 我们参考 Xu 和 Chen (2015) 的做法, 令 $M = \max\{x_1, x_2, \cdots, x_n\}$. 关于 MMCP 惩函数中的 γ 和 a, 前者采用 10-折交叉验证选出最优参数, 后者则根据 Breheny 和 Huang (2011) 的建议, 我们令 $a = 3$.

3.3 数值实验

为了说明本文所提 MMCP 方法的性能, 我们将其与 MSCAD 在数值实验上进行研究. 下面分别从高斯混合模型和泊松混合模型中产生样本数据并检验两种方法的定阶效果.

例 3.1 首先, 我们从如下高斯混合模型的密度函数

$$g(x;G) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\}.$$

中生成样本数据. 10 个模型的参数设定如表**??**所示, 且所有成分的标准差 $\sigma_k = 1$.

可以看出, 在所有混合模型中, 最复杂模型的混合成分个数为 7, 参考 Ishwaran 等 (2001), 我们令所有模型的成分个数初始值为 15, 对于那些成分均值之差小于 10^{-3} 的成分, 我们将相应的子群体合并, 同时对混合比例 π_k 进行加总 (Xu 和 Chen, 2015). 关于其他参数初始值的设定、迭代终止的准则, 与章节**??** 一样.

表?? 至表?? 展示了所有高斯混合模型基于 500 次蒙特卡洛模拟的结果. 可以看出在大部分模型中, MMCP 对成分个数 K 的估计准确率都高于 MSCAD (除了模型 4 和模型 7, MMCP 略低于 MSCAD). 特别地, MSCAD 方法倾向于将相似子群体合并, 这使得 MSCAD 对成分数的估计总是偏低, 尤其是当成分参数的距离很近时 (如模型 3、模型 5、模型 8 和模型 10), 过渡正则化使得 MSCAD 方法的估计效果很差, 而本文提出的 MMCP 方法则能较好的对 K 进行估计.

表 3-1: 两成分高斯混合模型定阶模拟结果.

K_0	$\hat{K}$	模型 1 (众数为 2)		模型 2 (众数为 2)		模型 3 (众数为 1)	
		MSCAD	MMCP	MSCAD	MMCP	MSCAD	MMCP
2	1	355	83	302	38	420	299
	2	77	264	112	275	45	99
	3	66	130	85	164	35	100
	4	2	20	1	21	-	2
	5	-	2	-	2	-	-
	6	-	1	-	-	-	-

注: K_0、$\hat{K}$ 分别表示参数 K 的真实值、估计值, 众数为模型对应的密度函数图像中“峰”的个数. 表中数值代表在 500 次蒙特卡洛模拟中, 各个方法得到相应 $\hat{K}$ 值的次数, 如 355 表示在模型 1 的 500 次模拟中, MSCAD 方法有 355 次将 K 估计为 1. 因此, K_0 所在行的数值越大, 相应方法的估计效果越好.

表 3-2: 四成分高斯混合模型定阶模拟结果.

K_0	$\hat{K}$	模型 4 (众数为 4)		模型 5 (众数为 1)		模型 6 (众数为 2)	
		MSCAD	MMCP	MSCAD	MMCP	MSCAD	MMCP
4	1	1	-	29	17	14	8
	2	1	-	58	56	80	27
	3	43	32	388	345	298	261
	4	139	123	21	63	79	130
	5	226	211	4	19	29	64
	6	77	95	-	-	-	9
	7	12	38	-	-	-	1
	8	1	1	-	-	-	-

例 3.2 从如下泊松混合模型

$$g(x;G)=\sum_{k=1}^{K}\pi_k\frac{\theta_k^x}{x!}\exp(-\theta_k).$$

中生成样本数据. 7 个模型的参数设定如表?? 所示, 其中每个模型我们均考虑样本量 $n=100,500$ 两种情形, 并分别进行 500 次蒙特卡洛模拟.

同样地, 在应用 MMCP 方法时, 令所有模型的成分个数初始值为 15, 其他参

表 3-3: 七成分高斯混合模型定阶模拟结果.

K_0	$\hat{K}$	模型 7 (众数为 7)		模型 8 (众数为 1)		模型 9 (众数为 3)		模型 10 (众数为 2)	
		MSCAD	MMCP	MSCAD	MMCP	MSCAD	MMCP	MSCAD	MMCP
	2	-	-	2	2	-	-	3	-
	3	-	-	27	18	13	6	27	6
	4	7	4	73	63	40	26	50	35
	5	23	18	213	161	167	112	251	158
	6	44	20	125	116	152	155	77	118
7	7	128	98	59	119	92	110	53	110
	8	88	78	1	12	23	53	17	27
	9	66	85	-	8	12	30	2	24
	10	28	42	-	1	1	7	1	2
	11	10	36	-	-	-	1	-	1
	12	8	16	-	-	-	-	-	-
	13	2	7	-	-	-	-	-	-

数初始值的设定及迭代终止的准则, 详见章节**??**.

表**??** 至表**??** 展示了所有泊松模型在两种样本量情况下的估计结果. 可以看出, 随着样本量的增大, 两种方法对于模型定阶的准确率都有所提升, 再者, 所有模型中, MMCP 对成分数 K 的估计准确率都高于 MSCAD.

表 3-4: 两成分泊松混合模型定阶模拟结果.

K_0	$\hat{K}$	模型 1		模型 2		模型 3	
		MSCAD	MMCP	MSCAD	MMCP	MSCAD	MMCP
n=100							
	1	446	279	308	59	312	239
2	2	54	221	192	441	188	244
	3	-	-	-	-	-	17
n=500							
	1	442	270	252	38	231	191
2	2	58	230	248	462	268	308
	3	-	-	-	-	1	1

3.4 小结

本章研究了有限混合模型的定阶问题. 首先, 我们阐述了信息准则法和惩罚最小距离法计算效率低及过拟合的问题; 接着, 基于以上工作, 对惩罚似然方法 (Chen 和 Khalili, 2008) 进行改进, 提出 MMCP 方法. 这种方法引入了两个惩罚函数: 一个对混合比例进行约束, 另一个则应用回归模型中变量选择的思想, 对

表 3-5: 三成分泊松混合模型定阶模拟结果.

K_0	$\hat{K}$	模型 4		模型 5	
		MSCAD	MMCP	MSCAD	MMCP
n=100					
	1	223	91	166	60
	2	122	192	220	240
3	3	155	217	114	199
	4	-	-	-	1
n=500					
	1	235	84	148	49
	2	111	153	295	260
3	3	154	262	57	186
	4	-	1	-	5

表 3-6: 四成分泊松混合模型定阶模拟结果.

K_0	$\hat{K}$	模型 6		模型 7	
		MSCAD	MMCP	MSCAD	MMCP
n=100					
	1	25	6	18	6
	2	152	124	64	54
	3	313	304	390	342
4	4	10	62	28	80
	5	-	4	-	18
n=500					
	1	38	8	11	4
	2	134	78	57	38
	3	300	249	381	282
4	4	27	157	45	138
	5	1	8	6	37
	6	-	-	-	1

成分参数距离进行约束, 从而解决了两种类型的过拟合. 特别地, MMCP 能够同时实现定阶及成分参数的估计. 最后, 数值试验显示, 相比 MSCAD 方法, MMCP 方法对有限混合模型进行定阶的准确率更高.

第 4 章　总结与展望

4.1　总结

有限混合模型是一种解决复杂分布的工具. 由于几乎所有概率密度都能用有限混合分布逼近, 因此它在理论和应用上均被广泛研究. 作为实际操作中基础又重要的工作, 有限混合混合模型中的参数估计及模型选择比单一模型的复杂得多. 本文便对这两个问题进行研究.

一方面, 针对有限混合模型的参数估计问题, 本文提出了三种改进的 EM 算法, 加快了经典 EM 算法的收敛速度. 通过模拟数据和实际数据分析得出, 新方法对参数估计的准确性和稳定性更高. 另一方面, 针对有限混合模型的定阶问题, 本文提出了 MMCP 方法, 同时实现混合模型的定阶及其它参数的估计. 相比 MSCAD 方法, MMCP 方法对混合模型的定阶准确率更高.

注意到第二章是在有限混合模型成分个数给定的条件下进行的. 在实际应用中, 有些研究对象可能有较多的理论依据以给出混合模型的成分个数, 或者数据本身可以清晰给出成分个数; 然而有些研究对象则需要事先进行定阶, 我们在例 2.3 中通过将其与 AIC 和 BIC 两种信息准则法相结合进行定阶, 但该方法的计算效率依旧有待提升. 第三章提出了 MMCP 方法, 该方法的数值求解过程较为复杂, 且其渐进性质有待证明, 这些都是值得进一步研究的问题.

4.2　未来的工作

下面展开几个拓展性问题, 即不同成分的混合模型的参数估计、多元混合模型的定阶及算法优化问题.

4.2.1　不同成分的混合模型的参数估计

本文所研究的混合模型可以是高斯混合模型、泊松混合模型、二项混合模型等, 但混合模型中各个子群体的分布是相同. 更为一般化的问题, 是研究不同成分的混合模型. 第二章所提出的三种改进算法的思想依旧是可以应用的. 但是第三章中对子群体分布参数之间的距离进行惩罚的思想, 如何应用到不同成分的混合模型中, 是我们未来将进行的工作.

4.2.2 多元混合模型的定阶

我们所展示的混合模型参数估计及定阶问题都是一元的, 如前文所述, 三种改进的 EM 算法可以被推广到多元情形, 并且无需额外工作量. 对于 MMCP 方法, 下面简单阐述改进的方法:

假定随机独立样本 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 均来自多元有限混合模型

$$g(\boldsymbol{x}; \boldsymbol{G}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}; \boldsymbol{\theta}_k),$$

其中 $\boldsymbol{G} = (K, \pi_1, \pi_2, \ldots, \pi_K, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K)$, $\boldsymbol{\theta}_k$ 为各成分的参数向量. 于是可得对数似然函数

$$l_n(\boldsymbol{G}) = \sum_{i=1}^{n} \log\{g(\boldsymbol{x}_i; \boldsymbol{G})\}.$$

本文第三章中, 我们首先对成分参数按大小进行排序, 再将成分参数距离定义为相邻参数的差. 这里, 我们对成分参数重新下定义

$$\eta_k^* = \min\{\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_j\| : j = k+1, k+2, ..., K\}.$$

可以看出, 本文讨论的情形只是当上式 $\boldsymbol{\theta}_k$ 取一维的情形. 接着, MMCP 方法相应地可改为

$$\tilde{l}_n(\boldsymbol{G}) = l_n(\boldsymbol{G}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_{MCP}(\eta_k^*).$$

我们的想法是: 给定某个较大的 K 值, 最大化上式以完成混合模型的定阶及参数估计, 同时防止两类过拟合情况. 当然此方法的求解也是一个较复杂的难题.

4.2.3 算法优化问题

最值问题一直是被广泛研究的经典问答. 文中第二章提出了三种改进 EM 算法, 新算法加快了收敛速度, 同时比经典 EM 算法更加准确及稳定, 但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. 第三章通过修改的 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, 另一方在每次迭代的 M 步都需要求解方程组, 这两者都消耗了大量的计算时间, 同样地, 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] 陈希孺. 高等数理统计学 [M]. 合肥: 中国科学技术大学出版社, 2009.

[3] 郑忠国, 童行伟, 赵慧. 高等统计学 [M]. 北京: 北京大学出版社, 2012.

[4] AHMAD K E, JAHEEN Z F, MODHESH A A. Estimation of a discriminant function based on small sample size from a mixture of two gumbel distributions[J]. Communications in Statistics - Simulation and Computation, 2010, 39(4):713-725.

[5] AKAIKE H. Information theory and an extention of the maximum likelihood principle[C]//Second International Symposium on Information Theory, 1973: 267-281.

[6] BEHBOODIAN J. On a mixture of normal distributions[J]. Biometrika, 1970, 57(1):215-217.

[7] BIERNACKI C, CELEUX G, GOVAERT G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models[J]. Computational Statistics & Data Analysis, 2003, 41(3-4):561-575.

[8] BILMES J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models[R]. California: TR-97-021, 1998.

[9] BISHOP C M. Pattern recognition and machine learning (information science and statistics)[M]. New York: Springer-Verlag, Inc, 2006, 430-432.

[10] BÖHNING D. A review of reliable maximum likelihood algorithms for semiparametric mixture models[J]. Journal of Statistical Planning and Inference, 1995, 47(1-2):5-28.

[11] 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.

[12] CHARNIGO R, SUN J. Testing homogeneity in a mixture distribution via the L2 distance between competing models[J]. Journal of the American Statistical Association, 2004, 99(466):488-498.

[13] CHEN H, CHEN J, KALBFLEISCH J D. A modified 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.

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] CHEN J, LI P. Testing the order of a normal mixture in mean[J]. Communications in Mathematics and Statistics, 2016, 4(1):21-38.

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

[22] COHEN A C. Estimation in mixtures of two normal distributions[J]. Technometrics, 1967, 9(1):15-28.

[23] DAY N E. Estimating the components of a mixture of normal distributions[J]. Biometrika, 1969, 56(3):463-474.

[24] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, 1977, 39(1):1-38.

[25] 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.

[26] 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.

[27] FU Y, CHEN J, LI P. Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions[J]. Journal of Statistical Planning and Inference, 2008, 138(3):667-681.

[28] GAREL B. Asymptotic theory of the likelihood ratio test for the identification of a mixture[J]. Journal of Statistical Planning and Inference, 2005, 131(2):271-296.

[29] GHOSH J K, SEN P K. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results[C]// In 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.

[30] 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.

[31] HASSELBLAD V. Estimation of finite mixtures of distributions from the exponential family[J]. Journal of the American Statistical Association, 1969, 64(328):1459-1471.

[32] HASSELBLAD V. Estimation of parameters for a mixture of normal distributions[J]. Technometrics, 1966, 8(3):431-444.

[33] HOLZMANN H, MUNK A, GNEITING T. Identifiability of finite mixtures of elliptical distributions[J]. Scandinavian Journal of Statistics, 2006, 33(4):753–

763.

[34] HOLZMANN H, MUNK A, STRATMANN B. Identifiability of finite mixtures - with applications to circular distributions[J]. Sankhyā: The Indian Journal of Statistics (2003-2007), 2004, 66(3):440-449.

[35] ISHWARAN H, JAMES L F, SUN J. Bayesian model selection in finite mixtures by marginal density decompositions[J]. Journal of the American Statistical Association, 2001, 96(456):1316-1332.

[36] JAMES L F, PRIEBE C E, MARCHETTE D J. Consistent estimation of mixture complexity[J]. Annals of Statistics, 2001:1281-1296.

[37] 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.

[38] LEHMANN, E L. Theory of point estimation[M]. New York: JohnWiley, 1998.

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

[40] LI P, CHEN J. Testing the order of a finite mixture[J]. Journal of the American Statistical Association, 2010, 105(491):1084-1092.

[41] LIU C, RUBIN D B. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence[J]. Biometrika, 1994, 81(4):633-648.

[42] LOUIS T A. Finding the observed information matrix when using the EM algorithm[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1982: 226-233.

[43] MCLACHLAN G J. The EM algorithm and extensions(second edition)[M]. New York: Wiley & Sons, Inc, 2008, 61-65.

[44] MELNYKOV V, MELNYKOV I. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components[J]. Computational Statistics & Data Analysis, 2012, 56(6):1381-1395.

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

[46] 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.

[47] PEARSON K. Contribution to the mathematical theory of evolution[J]. Philosophical Transactions of the Royal Society of London A, 1894, 186:71-110.

[48] 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.

[49] RICHARDSON S, GREEN P J. On Bayesian analysis of mixtures with an unknown number of components[J]. Journal of the Royal Statistical Society, 1997, 59(4):731-792.

[50] ROEDER K. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies[J]. Journal of the American Statistical Association, 1990, 85(411):617-624.

[51] SARAIVA E F, LOUZADA F, MILAN L. Mixture models with an unknown number of components via a new posterior split–merge MCMC algorithm[J]. Applied Mathematics & Computation, 2014, 244(2):959-975.

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

[53] ŠTEPÁNOVÁ K, VAVREČKA M. Estimating number of components in Gaussian mixture model using combination of greedy and merging algorithm[J]. Pattern Analysis & Applications, 2016:1-12.

[54] TEICHER H. Identifiability of finite mixtures[J]. Annals of Mathematical Statistics, 1963, 34(4):1265-1269.

[55] TEICHER H. On the mixture of distributions[J]. Annals of Mathematical Statistics, 1960, 31(1):55-73.

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

[57] UEDA N, NAKANO R. Deterministic annealing EM algorithm[M]. Elsevier Science Ltd. 1998:271–282.

[58] WEI G C G, Tanner M A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms[J]. Journal of the American Statistical Association, 1990, 85(411):699-704.

[59] WOO M J, SRIRAM T. Robust estimation of mixture complexity for count data[J]. Computational Statistics and Data Analysis, 2007, 51(9):4379-4392.

[60] WU C F J. On the convergence properties of the EM algorithm[J]. The Annals of Statistics, 1983: 95-103.

[61] 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.

[62] YIN J, ZHANG Y, GAO L. Accelerating expectation-maximization algorithms with frequent updates[C]// IEEE International Conference on CLUSTER Computing. IEEE, 2012:275-283.

[63] 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.

[64] ZHANG C H. Nearly unbiased variable selection under minimax concave penalty[J]. Annals of Statistics, 2010, 38(2):894-942.

硕士研究生期间的科研成果

论文:

XU G, WANG W, He R, Cao J. Modified EM algorithms for parameter estimation in finite mixture models[J]. Communications in Statistics - Theory and Methods, 2018, Under Review.

XU G, YANG S, WANG W. The ridge iterative regression and the data-augmentation lasso[J]. Acta Mathematica Sinica, 2018, Under Review.

课题:

混合模型及改进 EM 算法的理论与应用. 浙江省自然科学基金 Y19A010012. 排名: 3/5.

竞赛:

第十四届中国研究生数学建模竞赛. 二等奖. 排名: 2/3.

第十三届全国研究生数学建模竞赛. 三等奖. 排名: 1/3.

致　谢

在论文完整之际, 我想感谢硕士研究生阶段所有帮助和支持我的老师、同学和亲友们.

首先, 我要感谢我的导师王伟刚老师. 无论在科研还是工作生活中, 王老师都给了我很大的指导和鼓励. 读研以来, 我前进的每一步背后都倾注了王老师大量的心血. 王老师学识渊博、功底深厚, 对于科学问题总有独到的见解, 本文的不少创新点正是在与老师的讨论当中诞生的. 尽管王老师平日里要负责学院里的工作和繁重的教学任务, 但他对于科研的认真态度从不懈怠, 对于学生的学习和生活也时常放在心上. 非常幸运能在硕士研究生阶段跟着王老师学习, 师恩难忘, 在此向王老师表示最诚挚的感谢!

其次, 我要感谢讨论班的明瑞星老师、董雪梅老师. 明老师的科研态度严谨认真, 能够全面深刻地看待问题, 在课上、讨论班上总能提出创新、专业的观点, 明老师对于本论文的选题和创新点也提供了很大的帮助; 董老师为人和蔼, 经常在讨论班对学生的疑问作出透彻的解释, 给出很多相应的解决方案. 未来的学习和工作中, 我将以老师们为榜样不断前进.

此外, 我要感谢在讨论班一起学习的肖敏学姐、李亚旭学长以及各位同学. 肖学姐和李学长对我论文写作过程中的指导, 让我获益匪浅. 也要感谢班里的同学, 在平时一起学习、打球、游泳, 使我的生活充满了乐趣.

最后, 感谢我家人、朋友、亲戚, 在我读研期间给我的生活提供了各方面的关怀, 使我保持良好的状态投入到学业中.

再次感谢所有帮助我的人们!