中国科学: 数学 2019 年 第 49 卷 第 8 期: 1159～1182
SCIENTIA SINICA Mathematica
论 文

一种惩罚最大似然方法估计混合回归模型

徐建军[1*], 谭鲜明[2], 张润楚[3]

1. 浙江财经大学金融学院, 杭州 310018;
2. The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA;
3. 南开大学数学学院, 天津 300071
E-mail: jin@zufe.edu.cn, xzt1@psu.edu, zhrch@nankai.edu.cn

收稿日期: 2018-02-16; 接受日期: 2018-06-07; 网络出版日期: 2019-04-16; * 通信作者
国家自然科学基金 (批准号: 71631005)、教育部人文社科 (批准号: 12YJC910011)、浙江省自然科学基金 (批准号: LY18G010015)
国家统计科学研究 (批准号: 2017LY74) 资助项目

摘要 本文考虑具有正态误差假设下混合回归模型的参数估计问题. 由于似然函数的无界性, 混合回归模型普通的最大似然估计不存在. 本文提出一种惩罚最大似然方法来估计混合回归模型的参数, 证明惩罚最大似然估计量 (penalized maximum likelihood estimation, PMLE) 具有强相合和渐近正态性. 通过深入模拟研究, 从估计精确性角度看, 惩罚最大似然估计量有很好的表现. 本文还给出一个音调感知的例子来说明理论结果的应用.

关键词 混合回归 惩罚最大似然估计量 强相合 渐近正态

MSC (2010) 主题分类 91B82, 91B84

1 引言

由于混合回归方法非常适合于数据异质性问题的建模, 使得它在生物学、经济学和金融数据分析等领域有重要的应用. 在过去的二十多年, 无论是业界还是理论界, 对混合回归模型的理论和应用研究产生了众多的研究成果, 一本非常好的混合回归模型综述性分析专著可参见文献 [1]. 在本文研究中, 我们专注于 K 个成分的混合回归模型, 其定义如下:

$$\begin{cases} y_i = \alpha_1 + x_i^{\mathrm{T}}\beta_1 + \varepsilon_{1i}, & \text{以概率 } \pi_1, \\ y_i = \alpha_2 + x_i^{\mathrm{T}}\beta_2 + \varepsilon_{2i}, & \text{以概率 } \pi_2, \\ \quad\vdots \\ y_i = \alpha_K + x_i^{\mathrm{T}}\beta_K + \varepsilon_{Ki}, & \text{以概率 } \pi_K, \end{cases} \tag{1.1}$$

英文引用格式: Xu J J, Tan X M, Zhang R C. A penalized maximum likelihood approach for estimating mixture of regression models (in Chinese). Sci Sin Math, 2019, 49: 1159–1182, doi: 10.1360/SCM-2018-0132

© 2019《中国科学》杂志社

其中 $x_i \in \mathbb{R}^p$ $(i=1,2,\ldots,n)$ 是 p 维的独立随机向量, 具有共同的分布 F_x, 其中记号 T 代表矩阵或向量的转置; $\alpha_k \in \mathbb{R}$ $(k=1,2,\ldots,K)$ 表示 K 个成分的回归截距项; $\beta_k \in \mathbb{R}^p$ $(k=1,2,\ldots,K)$ 表示 K 个成分的回归参数; π_k $(k=1,2,\ldots,K)$ 为 K 个成分的比例参数, 其满足 $\pi_k > 0$, 且有 $\sum_{k=1}^K \pi_k = 1$. 另外, 这里假设 K 个成分的误差项满足 $\varepsilon_{ki} \sim N(0,\sigma_k^2)$ $(k=1,2,\ldots,K)$. 在本文中, 我们还假设 K 是先验已知的常数, 对于模型 (1.1), 感兴趣的未知参数为

$$\boldsymbol{\theta} = (\pi_1, \alpha_1, \beta_1^{\mathrm{T}}, \sigma_1^2, \ldots, \pi_K, \alpha_K, \beta_K^{\mathrm{T}}, \sigma_K^2).$$

相应的模型真实参数可记为 $\boldsymbol{\theta}_0 = (\pi_{10}, \alpha_{10}, \beta_{10}^{\mathrm{T}}, \sigma_{10}^2, \ldots, \pi_{K0}, \alpha_{K0}, \beta_{K0}^{\mathrm{T}}, \sigma_{K0}^2)$. 协变量 x_i 的共同分布记为 F_x, 通常它也是未知的, 但在很多情形下, 我们并不关注协变量分布推断的问题, 故这里把它当作多余参数来看待.

混合回归模型被认为具有结合回归模型和有限混合模型特征的优点, 具体可参见 McLachaln 和 Peel [2] 对该模型的详细说明. 一方面, 在混合回归模型下, 观测数据被认为来自于 K 个潜伏族, 因而, 它也被看成是一种特殊的潜伏族模型. 另一方面, 对于在给定相同潜伏族下的数据, 相依变量 y 和协变量 x 的关系可在共同的回归模型框架下来研究. Wedel 和 Desarbo [3] 推广了混合线性回归到混合广义线性回归模型. 为了陈述的简单性, 本文只考虑简单的混合线性回归模型.

一个基本的统计推断问题是, 寻找好的方法来估计混合回归模型的参数, 在过去几十年里, 人们提出了很多估计方法. 例如, Quandt 和 Ramsey [4] 引入了矩母函数 (moment generating function, MGF) 估计量, 并指出 MGF 优于传统的矩法 (moment method, MM) 估计. Schmidt [5] 给出了一改进版的 MGF 方法. Xu [6] 提出用一连续经验特征函数 (continuous empirical characteristic function, CECF) 方法拟合混合回归模型. 本质上说, 这些方法都可视为矩法的推广或变形.

在正则模型下, 最大似然估计 (maximum likelihood estimation, MLE) 是渐近有效的且可通过 EM 算法 (expectation-maximization algorithm) 容易完成数值的计算 (参见文献 [7]). 特别地, Lindsay [8] 认为 MLE 是当前混合模型参数估计问题中最常用的一种方法. 然而, 在混合回归模型 (1.1) 下的最大似然估计在自然参数空间中是无界函数, 因而, 它不是具有良定义的估计量. 理论上, 最大似然估计是不存在的. 这是因为当某一成分的方差参数 σ_j^2 趋于 0 时, 似然函数趋于无穷, 似然函数无界的说明具体可参见文献 [9]. 在实践中, 似然函数这种退化或奇异问题, 也即当 $\sigma_k = 0$ 时, 在成分数较大的混合回归模型中经常会发生.

为了克服似然函数无界的问题, 人们通常对似然函数的参数空间做一些约束. 例如, Hartely 和 Mallela [10]、Kiefer [11]、Desarbo 和 Cron [12] 都给出了约束最大似然的方法, 通过使约束参数空间为紧集, 使得似然函数在约束参数空间中有界. 具体地, 他们假设 $\sigma_k^2 \geqslant c > 0$, $k=1,2,\ldots,K$, 其中 c 为很小的正常数, 证明了在真实参数属于约束参数空间下, 约束最大似然估计量是相合、渐近正态和有效的. 受到 Hathaway [13] 的启发, Phillips [14] 通过设定两个成分方差比下界的方法, 提出一非紧参数空间的约束, 具体方差比约束可表示为 $\min_{k,j} \frac{\sigma_k}{\sigma_j} \geqslant c > 0$, 其中 c 为任一很小的正常数. 这种约束保证了全局最大似然估计量的存在. 值得注意的是, 当常数 $c=1$ 时, 这等价于各成分的方差满足 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$.

不同常数 c 大小的选择可能会影响约束最大似然估计相合性的结果. Yao [15] 为此提出了一种剖面似然的方法, 在不同方差正态混合分布下, 考虑似然函数无界问题的剖面似然方法, 通过寻找多维参数似然函数最大值, 利用剖面似然的原理, 转化为求一维参数的最大值问题, 然后结合图形搜索似然函数最大值的方法, 完成剖面似然估计的问题. Chen 等 [16] 通过精确估计极端观测点影响似然值的大小, 对方差参数进行适当的惩罚, 提出了一种惩罚似然估计的方法解决最大似然估计量在自然参数空

间中不存在的问题.

在本文中, 我们推广 Chen 等[16] 在正态混合分布下 PMLE 大样本的结果到一般的混合回归模型, 提出一种惩罚最大似然方法估计具有随机协变量和正态误差的混合回归模型的参数. 本文的主要贡献是, 建立了在模型 (1.1) 下 PMLE 的强相合、渐近正态和有效性的结果. 进一步, 我们还通过深入的模拟研究, 探讨了样本量、成分数、混合比例、方差大小、协变量的维度和惩罚项等因子对 PMLE 估计效果影响的大小.

余下内容结构如下. 第 2 节给出主要理论的结果. 一种修正的关于 PMLE 的 EM 算法和一些模拟及一个实际的例子放在第 3 节. 第 4 节总结全文的研究, 讨论了将来可能研究的问题. 所有的证明都放在附录 A 中.

2 主要结果

由模型 (1.1) 生成的随机样本数据是独立的, 且满足如下的联合分布 $F(y \mid x, \boldsymbol{\theta}) \times F_x(x)$, 其中

$$\frac{dF(y \mid x, \boldsymbol{\theta})}{dy} = f(y \mid x, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k} \phi\left(\frac{y - \alpha_k - x^{\mathrm{T}}\beta_k}{\sigma_k}\right), \tag{2.1}$$

其中 $\phi(x)$ 表示标准正态的密度函数. 这里考虑的是具有随机的回归设计阵, 且 x_i 的抽样分布为 F_x.

我们表示参数 $\boldsymbol{\theta}$ 的空间为

$$\Theta = \Big\{\boldsymbol{\theta} : \boldsymbol{\theta} = (\pi_1, \alpha_1, \beta_1^{\mathrm{T}}, \sigma_1^2, \ldots, \pi_K, \alpha_K, \beta_K^{\mathrm{T}}, \sigma_K^2,) \in \mathbb{R}^{K(p+3)},$$
$$\pi_k > 0, \sum_k \pi_k = 1, \sigma_k > 0, k = 1, 2, \ldots, K\Big\}.$$

容易看出, 如果我们重新对混合回归模型的各成分参数排序, 虽参数 $\boldsymbol{\theta}$ 形式改变, 但密度函数公式 (2.1) 仍然保持不变. 因此, 如果两个参数集 $\boldsymbol{\theta}$ 与 $\boldsymbol{\theta}^*$ 之间的差异可通过置换成分参数的位置来消除, 则认为两个参数集是相同的. 由于我们可以对 K 个成分以一定字典序的方式排序, 这样可以保证混合回归模型参数的唯一表示, 因而可避免模型参数不可识别的问题. 例如, 我们可以按照成分比例参数、方差参数和回归参数的次序排序以达到模型参数可识别的目的.

模型参数可识别是做任何有意义统计推断的前提. 为了保证模型 (1.1) 参数可识别, 我们还需要对随机协变量 x_i 做一些假设. 先定义一个 $p-1$ 维的超平面, 设 $\gamma \in \mathbb{R}^p \setminus \{0\}$, 记

$$H(\gamma) = \{x \in \mathbb{R}^p : x^{\mathrm{T}}\gamma = 0\},$$

令 $\mathcal{H}_{p-1}$ 表示 $p-1$ 维超平面张成的空间:

$$\mathcal{H}_{p-1} = \{H(\gamma) : \gamma \in \mathbb{R}^p \setminus \{0\}\}.$$

如下对 F_x 的假设保证了模型 (1.1) 参数的可识别性:

(C1) 协变量的分布 F_x 满足 $F_x(H) = 0, \forall\, H \in \mathcal{H}_{p-1}$.

注 2.1 按照 Hennig[17] 的研究, 只有协变量的样本点 $\{x_i, i = 1, 2, \ldots, n\}$ 分布在空间 $\mathcal{H}_{p-1}$ 中至少 $K+1$ 个超平面上, 具有固定设计阵的混合回归模型才是可识别的. 然而, 本文考虑的是随机设计阵的情形. 注意到条件 (C1) 是对随机协变量 x_i 分布 F_x 的约束要求, 而非观测数据的设计阵 $X_{n\times p}$.

在 (C1) 条件下, 设计阵中样本点分布在小于 $K+1$ 个超平面上的概率为 0. 另一方面, 如果存在某一个 $p-1$ 维超平面 H, 使得 $F_x(H)>0$, 则以正的概率我们可以确定设计阵中样本点落在个数小于 $K+1$ 的超平面上, 因而, 在非零概率下, 混合回归模型的参数是不可识别的. 条件 (C1) 是一个较弱的假设. 显然, 关于任一 σ- 有限测度下的绝对连续分布族满足条件 (C1). 在设计阵不同情形下, 文献 [17] 对有限混合回归模型的参数可识别性问题做了深入的研究.

除了条件 (C1), 为了得到模型 (1.1) 参数估计相合性的结果, 我们还需要如下的假设:

(C2) 随机协变量 x_i 独立于误差项 ε_{kt}, $k=1,2,\ldots,K$, 同时满足 $0<\mathrm{E}_0^X\|x_1\|^2<\infty$, 记号 E_0^X 表示在真实分布 F_x 下的期望.

(C3) 真实的参数 $\boldsymbol{\theta}_0$ 是参数空间 Θ 中的一个内点.

假设 (C2) 和 (C3) 是标准回归模型分析的通常条件 (参见文献 [18]). 在固定协变量情形, 如果记 X 为设计阵, 条件 $0<\mathrm{E}_0^X\|x_1\|^2<\infty$ 意味着当 $n\to\infty$ 时, $n^{-1}X^{\mathrm{T}}X$ 收敛于一个正定的矩阵.

基于上面的假设, 在给定样本 $\{(y_i,x_i^{\mathrm{T}})\}_{i=1}^n$ 下, 模型 (1.1) 下的对数似然函数为

$$\ell_n(\boldsymbol{\theta},F_x)=\sum_{i=1}^n\log f(y_i\mid x_i,\boldsymbol{\theta})+\sum_{i=1}^n\log dF_x(x_i).$$

对数似然函数被分成两部分: 第一部分仅包含参数 $\boldsymbol{\theta}$; 第二部分仅包含协变量分布 F_x. 为了估计 $\boldsymbol{\theta}$, 我们只需要考虑第一部分对数似然函数, 其定义为

$$\begin{aligned}\ell_n(\boldsymbol{\theta})&=\sum_{i=1}^n\log f_i(y_i\mid x_i,\boldsymbol{\theta})\\&=\sum_{i=1}^n\log\left\{\sum_{k=1}^K\frac{\pi_k}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_i-\alpha_k-x_i^{\mathrm{T}}\beta_k)^2}{2\sigma_k^2}\right)\right\}.\end{aligned}$$

通过适当给定 (α_1,β_1) 的值, 容易发现当 $y_1-\alpha_1-x_1^{\mathrm{T}}\beta_1=0$ 成立时, 再进一步, 如果设 $\sigma_1\to0$, 则有 $\ell_n(\boldsymbol{\theta})\to\infty$. 这说明在参数空间 Θ 下, 参数 $\boldsymbol{\theta}$ 的全局最大似然估计量不存在. 为了解决这个问题, 我们提出加一惩罚函数 $p_n(\boldsymbol{\theta})$ 到 $\ell_n(\boldsymbol{\theta})$, 使得惩罚对数似然函数 $pl_n(\boldsymbol{\theta})=\ell_n(\boldsymbol{\theta})+p_n(\boldsymbol{\theta})$ 在 Θ 上有界. 惩罚最大似然估计量 (PMLE) 定义为

$$\hat{\boldsymbol{\theta}}_n=\arg\max_{\boldsymbol{\theta}\in\Theta}pl_n(\boldsymbol{\theta}).$$

为了保证惩罚对数似然函数的有界性和较好的统计性质, 我们需要惩罚函数 $p_n(\boldsymbol{\theta})$ 满足如下条件:

(A1) $p_n(\boldsymbol{\theta})=\sum_{k=1}^K\tilde{p}_n(\sigma_k)$;

(A2) $\sup_{\sigma>0}\max\{0,\tilde{p}_n(\sigma)\}=o(n)$;

(A3) 对任意的 $\sigma\in(0,1/\sqrt{n}]$, 有

$$\tilde{p}_n(\sigma)\leqslant(2M_0+1)\sqrt{n}\log n\log\sigma,\quad \text{对足够大的 } n \text{ 成立},$$

其中 $M_0=\sum_{k=1}^K\frac{\pi_{k0}}{\sqrt{2\pi}\sigma_{k0}}$;

(A4) $\tilde{p}_n'(\sigma)=\frac{d\tilde{p}_n(\sigma)}{d\sigma}=o(n^{1/2})$ 对任一固定的 $\sigma>0$ 成立.

上述惩罚函数条件的思想来自于 Chen 等 [16]. 条件 (A1) 要求惩罚函数仅依赖于方差参数并具有可加性. 可加性的要求并不一定保证 PMLE 有较好的大样本性质, 但它能极大方便 PMLE 的计算. 条件 (A2) 隐含惩罚项对似然函数不能起主导的影响, 这个要求对于 PMLE 保留似然估计好的统计性质

是非常重要的. 关键 (A3) 惩罚项的条件保证了惩罚对数似然函数的有界性. 最后的条件 (A4) 虽然对 PMLE 相合性成立并非必要, 但它可以避免 PMLE 渐近效率的损失.

对惩罚项 $p_n(\boldsymbol{\theta})$ 的上述 4 个条件要求并不严厉, 相反地, 我们可以很方便地构造满足这些条件的惩罚函数. 例如, 下面的惩罚函数:

$$p_n(\boldsymbol{\theta}) = -c_n \sum_{k=1}^{K} (S_k \sigma_k^{-2} + \log \sigma_k^2), \tag{2.2}$$

如果 $c_n = O(1)$, 则满足条件 (A1)–(A3), 且假设 $c_n = o(\sqrt{n})$, 则满足条件 (A1)–(A4), 我们应用该惩罚函数在后面的模拟研究中, 其中项 $S_k = O(1)$ a.s. $(k = 1, 2, \ldots, K)$ 的引入是为了保证 PMLE 在观测数据线性变换下保持不变, 通常只要满足 $\sqrt{S_k}$ 是观测数据下的线性估计. 在应用上, 数据线性变换下估计量的不变性是理想的统计性质. 在我们模拟研究中, 一般假设 S_k 为某成分方差参数的估计. 例如, 我们采用 K 均值聚类分析的方法 [19], 先把样本分成 K 个子样本, 然后在第 k 个子样本中, 用 y_i 对 x_i 做回归, 记 S_k 为残差方差估计. 因为 $S_k = O(1)$, a.s., 故惩罚项公式 (2.2) 仍然满足 (A1)–(A3).

注 2.2 从 Bayes 角度看, 惩罚函数公式 (2.2) 可视为方差参数 σ_k^2 $(k = 1, 2, \ldots, K)$ 来自于逆 Gamma 先验分布, S_k $(k = 1, 2, \ldots, K)$ 为相应先验分布的众数. 虽然其他的惩罚函数也可用来保证 PMLE 的相合、渐近正态和有效性, 但是惩罚函数公式 (2.2) 还有一计算上的优点, 在 EM 算法中的最大化 M 步时, 可得到统一的解析解, 这是因为逆 Gamma 是正态方差参数的共轭先验分布.

我们知道参数可识别性是相合估计量存在的必要条件. 对于模型 (1.1), 参数可识别性问题的分析由 Hennig [17] 的如下引理给出.

引理 2.1 在假设条件 (C1) 和 (C2) 下, 如果 $F(x, y \mid \boldsymbol{\theta}, F_x) = F(x, y \mid \tilde{\boldsymbol{\theta}}, F_x)$, 则 $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. 其中记号 “$=$” 表示 Θ 上的等价关系, 即通过成分参数的置换后, $\boldsymbol{\theta}$ 等于 $\tilde{\boldsymbol{\theta}}$.

注 2.3 对一般混合分布参数由于成分置换引起识别性问题的讨论, 可参见文献 [20–22]. 在下文, 假设以某种字典序的方式应用到参数空间 Θ, 这样真实的参数表示为一个点, 而非一个等价类.

我们关于 PMLE $\hat{\boldsymbol{\theta}}_n$ 的主要理论结果总结为如下的两个定理.

定理 2.1 在条件 (C1)–(C3) 和 (A1)–(A3) 下, 假设估计量 $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ 满足

$$p\ell_n(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) - p\ell_n(\boldsymbol{\theta}_0) \geqslant c, \quad \forall\, n,$$

其中 c 为任意常数, 则 $\mathrm{P}(\lim_{n\to\infty} \widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = \boldsymbol{\theta}_0) = 1$.

定理 2.1 隐含了 PMLE $\hat{\boldsymbol{\theta}}_n$ 是强相合于 $\boldsymbol{\theta}_{\mathbf{0}}$, 因为我们有 $p\ell_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) - p\ell_n(\boldsymbol{\theta}_0) \geqslant 0$, $\forall\, n$.

为了文章结构精简的考虑, 这里只阐述证明的基本思想, 定理证明的细节部分放在附录 A. 我们证明的关键想法如下:

(1) 当某一成分的方差参数趋于 0 时, 表明数据存在聚集于某一超平面的现象, 从而使似然函数趋于无穷;

(2) 如果聚集于某一超平面上数据的个数能被估计, 则这些数据对似然函数总的贡献也能被估计;

(3) 我们可以进一步估计这些数据对似然函数总贡献的上界;

(4) 基于估计的上界, 当在某成分方差参数很小时, 我们设计惩罚函数超过似然函数总贡献的上界, 从而使 PMLE 保持有界;

(5) 此时, 我们只需关注在参数空间中, 所有成分方差参数都不会趋于 0 的部分, 最后, 应用 Wald 方法 [23], 我们可以证明 PMLE 的强相合性.

更具体他, 证明中的关键步骤是要说明 $\widetilde{\boldsymbol{\theta}}_n$ 的成分方差参数以概率 1 远离 0 点. 为了达到这个目的, 我们只要说明, 当 $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ 中某成分方差趋于 0 时,

$$p\ell_n(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) - p\ell_n(\boldsymbol{\theta}_0) \to -\infty$$

以概率 1 成立. 为了说明这一点, 我们要得到当 $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ 中某成分方差参数接近于 0 时, $\ell_n(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}})$ 的一上界估计值. 可以证明这个上界依赖于观测数据 (y_i, x_i^{T}) 聚集于任一给定超平面附近的个数, 等价地, 我们要估计

$$\sup_{(\alpha,\beta^{\mathrm{T}})\in\mathbb{R}^{p+1}} \#\{i : |y_i - \alpha - x_i^{\mathrm{T}}\beta| \leqslant |\sigma\log\sigma|\},$$

其中, 记号 # 表示对应集合元素的个数. 这是因为当某成分方差参数趋于 0 时, 只有上述这些观测会显著增加对数似然函数的值, 并且任何单个数据的对数似然贡献值都不超过 $\log\frac{1}{\sigma}$. 我们可以得到所有这些数据对对数似然总贡献值的一个上界估计, 然后通过惩罚项来抵消这些数据增加对数似然值的效应, 通过计算对数似然和惩罚项的差知 $p\ell_n(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) - p\ell_n(\boldsymbol{\theta}_0) \to -\infty$, a.s. 一旦我们知道 $\widetilde{\boldsymbol{\theta}}_n$ 的方差参数以概率 1 远离 0, 按照 Wald [23] 的方法, 就容易完成 $\widetilde{\boldsymbol{\theta}}_n$ 相合性的证明.

定理 2.2 在定理 2.1 和 (A4) 条件下, 进一步假设在成分置换约束下参数有唯一的表示, 且在 Θ 中是可识别的, 我们有

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)),$$

其中

$$\boldsymbol{I}(\boldsymbol{\theta}_0) = \mathrm{E}_0^X\left\{\mathrm{E}_0^{Y|X}\left[\frac{\partial\log f(y_1 \mid x_1, \boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right]\left[\frac{\partial\log f(y_1 \mid x_1, \boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\right]\right\}$$

表示 Fisher 信息阵, 符号 $\mathcal{D}$ 表示依分布收敛, 算子 $\mathrm{E}_0^{Y|X}$ 代表在条件密度函数 $f(y \mid x, \boldsymbol{\theta}_0)$ 下的条件期望符号, 并且 E_0^X 是真实分布 F_x 下的期望.

定理 2.2 的证明是直接的, 在 $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ 相合性的基础上, 可应用标准 Kiefer [11] 的方法, 我们把定理 2.2 证明的细节放在附录 A 的最后.

3 模拟和应用

3.1 EM 算法

首先讨论 PMLE 的计算问题. 由于编码简单和在非常一般条件下收敛到局部极大值的理论保证 (参见文献 [24]), 因此, 在有限混合模型数值计算问题中, EM 已成为人们经常使用的一种算法选择. 另外, EM 算法也很容易被修改应用到惩罚似然函数的计算问题中, Green [25] 保证了每次 EM 迭代中, 惩罚似然函数值的增加.

在本文模拟研究中, 我们选择惩罚函数满足 (2.2) 的形式, 即

$$p_n(\boldsymbol{\theta}) = -c_n\sum_{j=1}^{K}(S_j\sigma_j^{-2} + \log\sigma_j^2),$$

这里样本先通过 K 均值聚类后, 对于第 k $(k = 1, 2, \ldots, K)$ 个子样本, 用 y_i 对 x_i 做回归后, 记 S_k 为相应的残差方差估计. 为了评估 c_n 对 PMLE 表现的影响效果, 在模拟中, 我们使用三种不同的 c_n 值做试验, 分别为 $c_n = 1, 1/\sqrt{n}, 1/n$.

为了构造 EM 算法, 对于每个观测样本 $\{(y_1, x_1^{\mathrm{T}}), (y_2, x_2^{\mathrm{T}}), \ldots, (y_n, x_n^{\mathrm{T}})\}$, 我们假设存在对应的潜变量 z_i, $i \in 1, \ldots, n$, 其中 $z_i = k$ $(k = 1, 2, \ldots, K)$ 代表第 i 个数据来自于第 k 个成分的回归模型. 因此, 完全数据的对数似然函数加上惩罚项具有如下表示:

$$p\ell_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} I\{z_i = k\} \times \left\{ \log \pi_k - \log \sigma_k - \frac{(y_i - \alpha_k - x_i^{\mathrm{T}}\beta_k)^2}{2\sigma_k^2} \right\} + p_n(\boldsymbol{\theta}).$$

设定初始值为 $\boldsymbol{\theta}^{(0)}$, EM 算法在 E 步和 M 步之间来回迭代. 在 E 步, 给定当前的参数值, 设为

$$\boldsymbol{\theta}^{(m)} = (\pi_1^{(m)}, \alpha_1^{(m)}, \beta_1^{T(m)}, \sigma_1^{2(m)}, \ldots, \pi_K^{(m)}, \alpha_K^{(m)}, \beta_K^{T(m)}, \sigma_K^{2(m)}),$$

在条件 $\{(y_i, x_i^{\mathrm{T}})\}_{i=1}^n$ 和 $\boldsymbol{\theta}^{(m)}$ 下, 首先计算完全数据下惩罚对数似然函数的期望值, 记为

$$\pi_{ik}^{(m+1)} = \mathrm{E}\{I\{z_i = k\} \mid (y_i, x_i^{\mathrm{T}}), \boldsymbol{\theta}^{(m)}\} = \frac{\pi_k^{(m)} \frac{1}{\sigma_k^{(m)}} \phi\big(\frac{y_i - \alpha_k^{(m)} - x_i^{\mathrm{T}}\beta_k^{(m)}}{\sigma_k^{(m)}}\big)}{\sum_{j=1}^{K} \pi_j^{(m)} \frac{1}{\sigma_j^{(m)}} \phi\big(\frac{y_i - \alpha_j^{(m)} - x_i^{\mathrm{T}}\beta_j^{(m)}}{\sigma_j^{(m)}}\big)},$$

对于 $i = 1, \ldots, n$, $k = 1, 2, \ldots, K$. 我们有

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}) &= \mathrm{E}\{\ell_n^c(\boldsymbol{\theta}) + p_n(\boldsymbol{\theta}) \mid \{x_i, y_i\}_{i=1}^n, \boldsymbol{\theta}^{(m)}\} \\ &= \sum_{k=1}^{K} (\log \pi_k) \sum_{i=1}^{n} \pi_{ik}^{(m+1)} - \frac{1}{2}\sum_{k=1}^{K} (\log \sigma_k^2) \sum_{i=1}^{n} \pi_{ik}^{(m+1)} \\ &\quad - \frac{1}{2}\sum_{k=1}^{K} \sigma_k^{-2} \sum_{i=1}^{n} \pi_{ik}^{(m+1)} (y_i - \alpha_k - x_i^{\mathrm{T}}\beta_k)^2 + p_n(\boldsymbol{\theta}). \end{aligned}$$

在 M 步, 计算函数 $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)})$ 关于 $\boldsymbol{\theta}$ 的极大化点. 容易证明, 按照 (2.2) 定义的惩罚函数, $\boldsymbol{\theta}^{(m+1)}$ 极大化点应该具有下列形式:

$$\begin{aligned} &\pi_k^{(m+1)} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \pi_{ik}^{(m+1)}, \\ &(\alpha_k^{(m+1)}, \beta_k^{\mathrm{T}(m+1)}) = \left(\sum_{i=1}^{n} \pi_{ik}^{(m+1)} \widetilde{x}_i \widetilde{x}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} \pi_{ik}^{(m+1)} \widetilde{x}_i^{\mathrm{T}} y_i, \\ &\sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \pi_{ik}^{(m+1)} (y_i - \alpha_k^{(m+1)} - x_i^{\mathrm{T}}\beta_k^{(m+1)})^2 + 2S_k c_n}{\sum_{i=1}^{n} \pi_{ik}^{(m+1)} + 2c_n}, \end{aligned}$$

其中 $\widetilde{x}_i = (1, x_i^{\mathrm{T}})^{\mathrm{T}}$. EM 计算达到被估参数满足一定收敛标准时停止. 在我们模拟研究中, 我们停止 EM 计算的标准是 $\|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}\|_\infty \leqslant 10^{-6}$, 其中 $\|\boldsymbol{\theta}\|_\infty$ 表示向量 $\boldsymbol{\theta}$ 最大绝对值范数.

3.2 模拟结果

为了评估应用惩罚最大似然方法估计混合回归模型参数的有限样本表现, 以及评估一些因子, 如样本量、成分数、混合比例、方差大小、协变量的维度和惩罚项等对估计效果影响的大小, 我们设计的模拟研究见表 1.

从表 1 中知, 我们考虑具有成分数 $K = 2, 3$ 的混合线性回归模型; 混合比例设置为相同和不相同两种情形; 有大、中和小三种成分方差; 协变量的维数为 $p = 1, 2, 3$; 惩罚函数公式 (2.2) 中的 $c_n = 0$,

表 1　模拟研究的设计

K	2, 3
π_k	$1/K+(k-(K+1)/2)\times b$, $b=0,\ 0.1$
σ_k^2	$a\times(0.5,1.0,2.0)$, $a=1,0.5,0.25$
p	$1,2,3$
c_n	$0,1/n,1/\sqrt{n},1$
n	$100,200,400$

$1/n$, $1/\sqrt{n}$, 1. 这样总数上有 $2\times2\times3\times3\times4=144$ 种模型构造. 对于每一种构造, 我们考虑在三种样本大小 $n=100$、$n=200$ 和 $n=400$ 下, 分别评估 PMLE 的表现. 在下面分析中, 我们记不同惩罚项分别为 P0 ($c_n=0$)、P1 ($c_n=1/n$)、P2 ($c_n=1/\sqrt{n}$) 和 P3 ($c_n=1$). 注意到 P0 对应于修正的 MLE (删除退化根后的最大似然估计值). 对每种模型构造下回归系数 $\theta_k=(\alpha_k,\beta_k^{\mathrm{T}})^{\mathrm{T}}$ 的设置可参见表 2.

给定一模型构造下的情形, 如 (K, (π_k, σ_k^2, θ_k, $k=1,2,\ldots,K$)), 我们通过如下的方式产生 n 个随机样本. 首先产生 p 维均值为 0, 协方差阵为 $\Sigma=(\sigma_{sl})$ 的多元正态样本 $(\eta_{i1},\ldots,\eta_{ip})_{i=1}^n$, 其中 $\sigma_{sl}=\rho^{|s-l|}$, 这里 $s,l=1,2,\ldots,p$, $\rho=0.1$. 其次, 令 $x_{ij}=2(\Psi(\eta_{ij})-0.5)$, 其中 Ψ 为标准正态概率分布函数, 因而, x_{ij} 为 $[-1,1]$ 上的均匀分布. 注意到, 如果 $p>1$, 各变量 $x_{i1},\ldots,x_{ip}$ 是相关的, 协变量的相关性符合很多实际数据的情况. 最后, 我们通过随机抽样分布 $\Pr(C_i=k)=\pi_{k0}$ 产生每一观测的潜变量, 且在潜变量给定状态下, 产生 $\epsilon_i\mid(C_i=k)\sim N(0,\sigma_{k0}^2)$ $(k=1,2,\ldots,K)$ 分布下的随机数. 被解释变量按照模型 (1.1) 容易得到.

通常惩罚对数似然函数具有多峰形态, 尝试运行多个不同的初值寻找惩罚对数似然函数最大值点是必要的. 在我们的模拟研究和实际数据分析中, 应用 EM 算法时, 我们使用 10 个随机产生的初值. 为了产生这 10 个初值, 首先应用 K 均值聚类法对 n 个观测样本 $\{(y_i,x_i^{\mathrm{T}})\}_{i=1}^n$ 分成 K 类. 然后, 对每个类中的数据, 用 y_i 对 x_i 做回归, 这样我们得到了每类中的回归系数 $\tilde{\theta}_{k0}$ 和残差方差 $\tilde{\sigma}_{k0}^2$. 注意到 $\tilde{\theta}_{k0}$ 是 $p+1$ 维的向量, 有一个截距项和 p 个斜率. 最后, 固定第 k 个成分的方差参数为 $\tilde{\sigma}_{k0}^2$ $(k=1,2,\ldots,K)$, 设定第 k 个成分的混合比例为相应第 k 类数据的个数与样本容量 n 的比值, 10 个初值通过随机产生 $N(\tilde{\theta}_{k0},(\tilde{\sigma}_{k0}^2/4)\boldsymbol{I}_{p+1})$ 分布下第 k 个成分的回归系数得到, 其中 $\boldsymbol{I}_{p+1}$ 是 $p+1$ 维的单位阵.

从这 10 个初值的任意一个开始, 经过多次迭代后, 按照某种停止规则, EM 算法会停止于参数空间中的一个点. 在模拟中, 我们设定停止的标准是满足 $\|\boldsymbol{\theta}^{(m+1)}-\boldsymbol{\theta}^{(m)}\|_\infty\leqslant10^{-6}$, 或者迭代的次数超过 1,000 次. 另外, 如果对任意的 $1\leqslant j\leqslant K$, 满足 (i) $\sigma_j^2<1\times10^{-16}$, 或者 (ii) $\sum_{i=1}^n I(\pi_{ij}>1\times10^{-16})\leqslant p+1$, 其中 $I(\cdot)$ 表示示性函数, 则称 EM 算法停止于一奇异点. 每次我们都记录 EM 算法收敛到奇异

表 2　回归系数的设置

	K	
	2	3
$p=1$	$\theta_1=(-3.0,1.0)^{\mathrm{T}}$	$\theta_3=(1.0,-1.0)^{\mathrm{T}}$
	$\theta_2=(-1.0,0.5)^{\mathrm{T}}$	
$p=2$	$\theta_1=(-3.0,1.0,1.0)^{\mathrm{T}}$	$\theta_3=(1.0,-1.0,-0.5)^{\mathrm{T}}$
	$\theta_2=(-1.0,0.5,0.5)^{\mathrm{T}}$	
$p=3$	$\theta_1=(-3.0,1.0,1.0,1.0)^{\mathrm{T}}$	$\theta_3=(1.0,-1.0,-0.5,0.5)^{\mathrm{T}}$
	$\theta_2=(-1.0,0.5,0.5,0.5)^{\mathrm{T}}$	

点的次数. 最后, 对任一模拟的数据集, 去除奇异估计值后, 通过比较 10 次的最大似然值或最大惩罚似然值, 我们得到修正的 MLE 和 PMLE. 每种情况, 我们都重复 500 次, 并保存估计和相关的信息.

为了评估这些估计量 P0、P1、P2 和 P3 的精确性, 我们计算在不同模型构造下估计量的偏差、标准差和均方误差根. 为了更好地看出不同情形下估计量偏差、标准差和均方误差根的变化, 以及避免展示过多的表格, 我们只给出了两成分时不同模型构造下偏差、标准差和均方误差根, 见图 1–4, 三成分不同模型构造下也有相应类似的结果. 这些模型构造主要依赖于方差 (大, 中, 小)、混合比例 (相同, 不相同) 和协变量 x 的维度 ($p = 1, 2, 3$), 一共有 18 个两成分和 18 个三成分模型, 模型构造的方式见表 3. 不管对两成分还是三成分模型, 我们在图中只给出共同参数偏差、标准差和均方误差根的变化, 这些共同参数分别为每个成分下的截距、x_1 对应的斜率、成分方差和混合比例.

从图 1 上观察, 我们可以看到随着样本量的增加, 对所有惩罚项和模型构造下, 偏差、标准差和均方误差根都是下降的. 进一步, 我们发现方差尺度系数 a 对估计精度效应的影响是显著的, 对任一惩罚项, a 值越大, 所有估计的偏差、标准差和均方误差根值都越大. 最后, 我们还发现惩罚项对估计的偏差、标准差和均方误差根有显著影响, 明显地, P1、P2 和 P3 估计下的截距项相对于 P0 有更小的偏差、标准差和均方误差根, 而从图 1 上看 P1、P2 和 P3 之间的这些差异则不明显.

对于其他参数估计, 我们发现方差尺度系数 a 和样本量 n 对估计的偏差、标准差和均方误差根有相同的影响效果, (见图 2–4). 然而, 惩罚项对这些不同参数的估计精度会产生不同的影响. 例如, 在图 2 中, 关于 P0、P1、P2 和 P3 下的斜率估计, 我们没有发现存在显著的差异, 但是对成分方差和混合比例而言, 我们观察到 P0、P1、P2 和 P3 之间存在显著的差异, (见图 3 和 4). 整体上看, 我们尝试所有模拟情形, 当估计存在差异时, 估计量 P1、P2 和 P3 的表现要优于 P0.

为了进一步评估这些因素对估计精度的影响程度, 我们做了方差分析, 通过设定某参数估计的偏差或均方误差根作为响应变量、样本量 (n)、方差尺度系数 (a)、混合比例差异度 (b)、协变量的维数 (p) 和惩罚项 (c_n) 作为因子. 表 4 呈列了一些显著性的因子 (p 值 < 0.05), 这些因子以 p 值升序的方式排列, 它们对给定参数估计的偏差或均方误差根能产生显著的影响. 例如, 当成分数 $K = 2$ 时, 对于估计参数 β_{11}, 方差尺度系数、样本量和惩罚项按重要程度依次对该参数估计的偏差或均方误差根产生显著影响, 而其他因子则没有显著性表现. 为了比较三个非零的 c_n, 分别对应于估计量 P1、P2 和 P3, 分析 c_n 是否存在不同的显著表现, 我们移除估计量 P0 后做相同的方差分析. 方差分析显著性因子结果罗列在对应于 P1、P2 和 P3 列的表格中. 我们注意到估计量 P1、P2 和 P3 对 β_{11} 偏差的影响是显著不同的, 越大的 c_n, 存在越大的偏差; 然而, P1、P2 和 P3 对均方误差根的影响不存在显著的差异.

由于混合比例和方差等这些因子来源于实际问题本身, 它们不受研究者的控制, 而惩罚项是我们完全能控制的一个因子. 基于这个原因, 虽然整个模拟研究给出了更多的结果, 但我们主要分析 c_n 如何影响估计的精确性. 总体而言, 不管两成分还是三成分模型, 对大多数的参数, c_n 的确显著影响估计的偏差, 越小的 c_n, 如 P0, 并不必然会产生更小的偏差, 而且 P0 相对于 P1、P2 和 P3, 对大多数参数估计有更大的均方误差根. 关于均方误差根在 P1、P2 和 P3 间的显著性差异仅存在于混合比例和成分方差, 看上去 P3 ($c_n = 1$) 有相对更小的均方误差根.

P0 比 P1、P2 和 P3 有更大偏差的现象看上去好像是违反人们的直觉. 一种合理的解释应该是增加惩罚项导致惩罚对数似然函数凸性的增加, 从而使 PMLE 的计算变的更稳定. 为了验证这个解释, 在模拟研究中, 我们还比较了 P0、P1、P2 和 P3 平均的 EM 迭代次数. 对于每一种模拟构造下的任一方法, 在 5,000 次 EM 计算中, 其中包含 500 次重复样本和每个样本下的 10 个初值, 我们统计了平

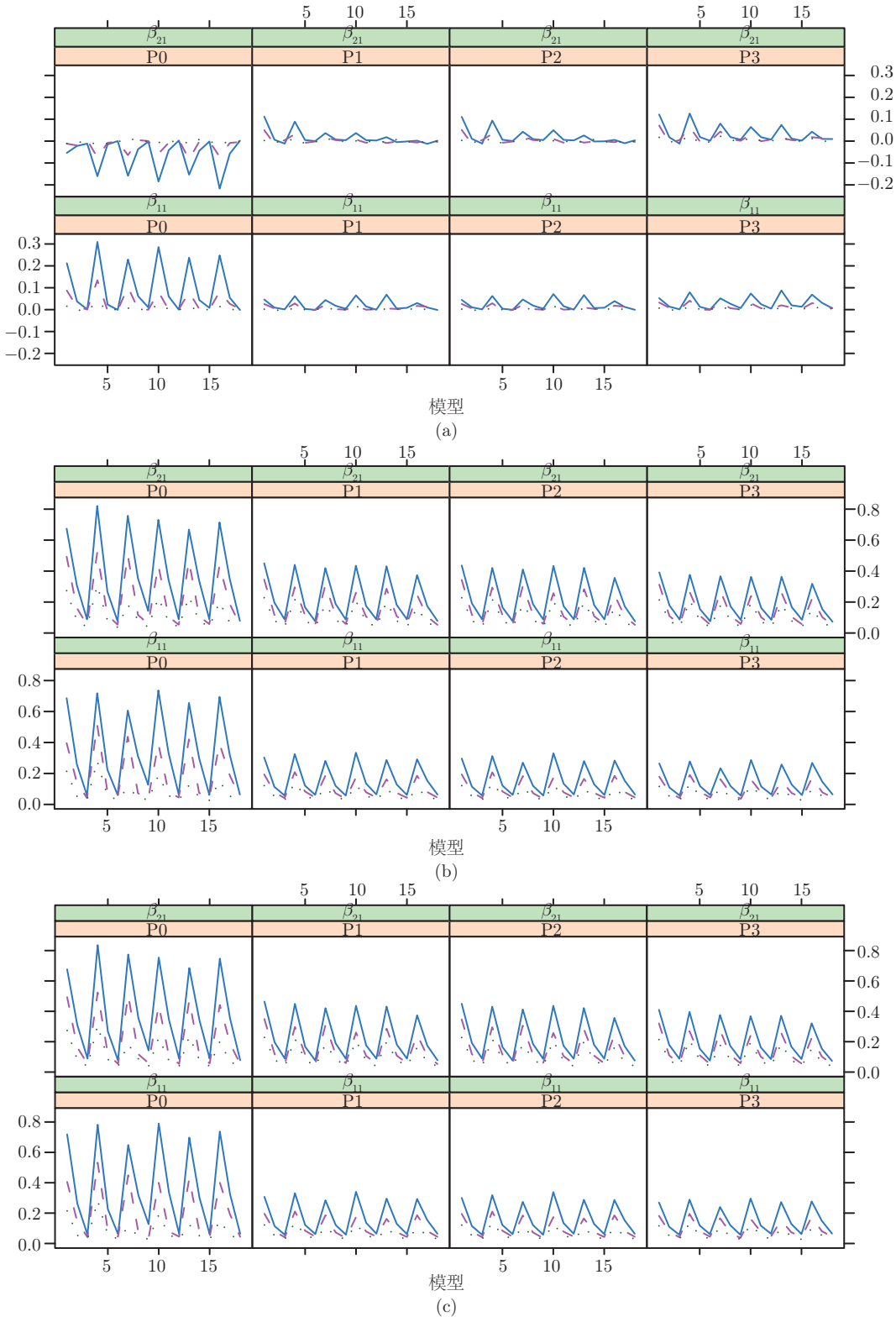


图 1　两成分混合回归模型截距项估计的偏差 (a)、标准误 (b) 和均方误差根 (c). 样本量分别为 100 (实线)、200 (虚线) 和 400 (点虚线); 横轴表示表 3 中依次构造的 18 个模型

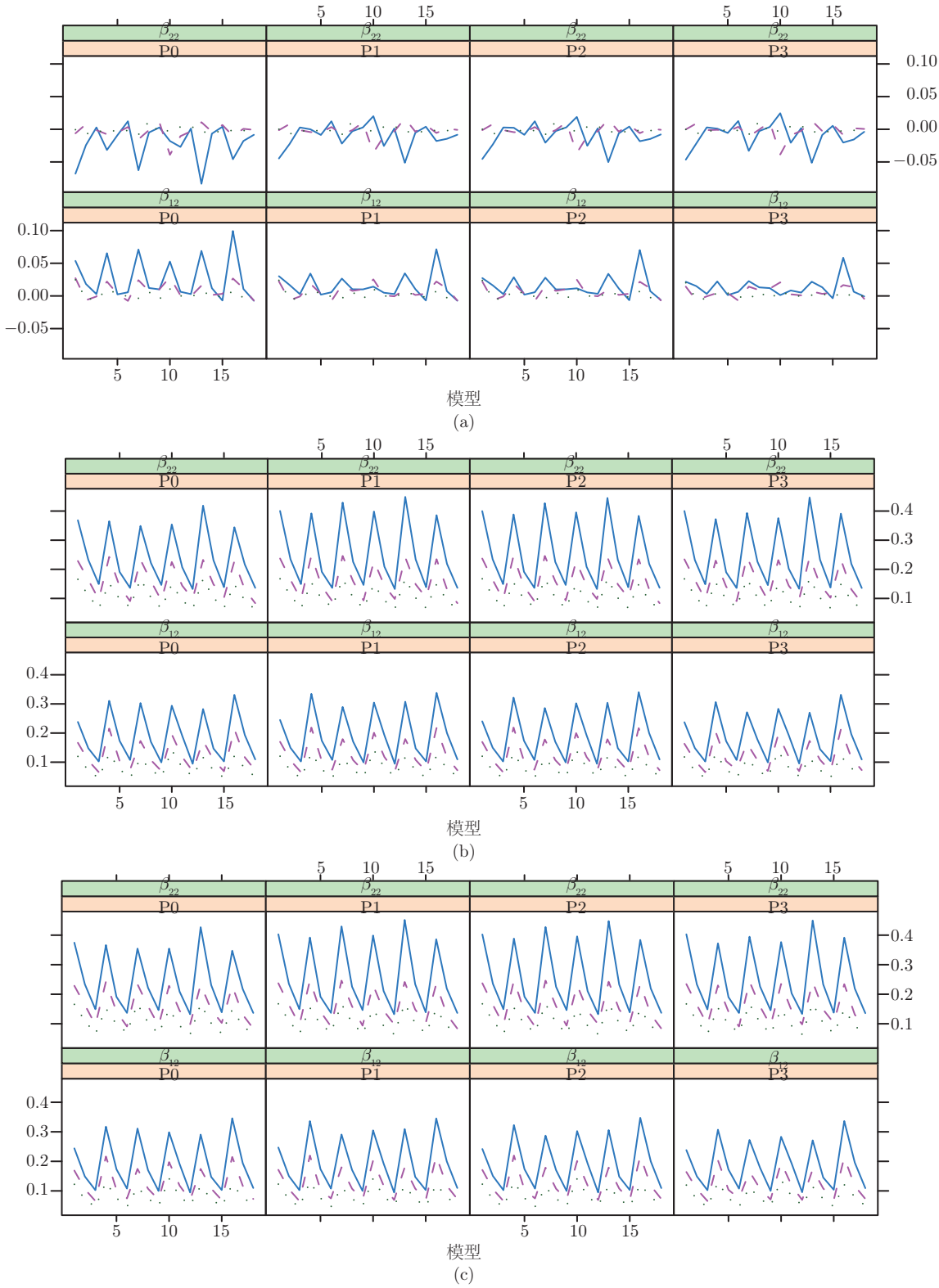
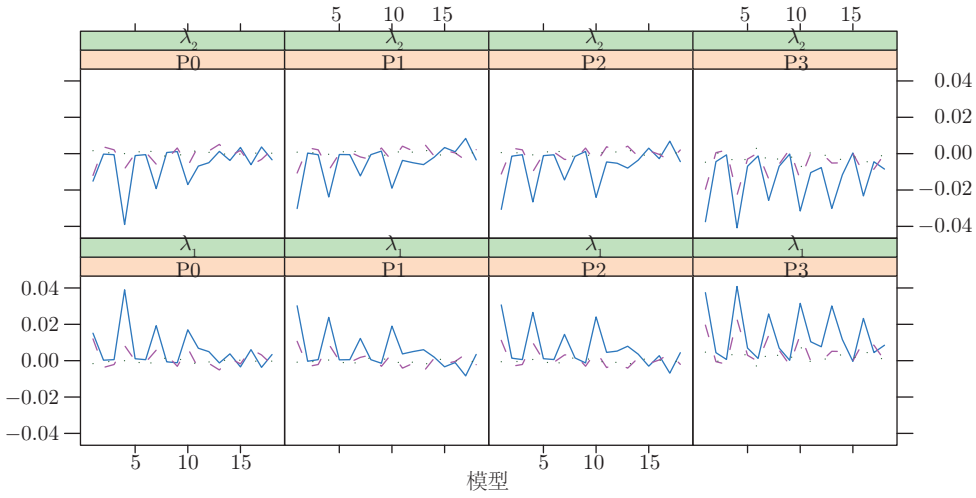
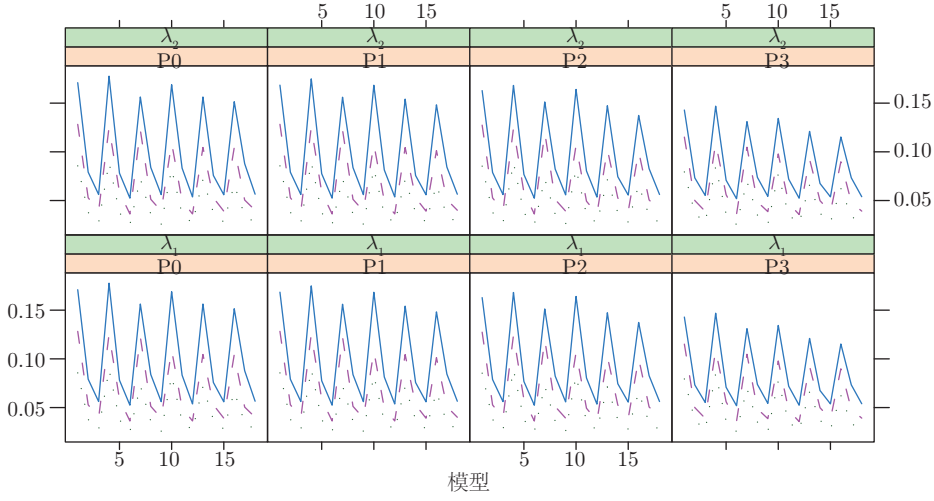


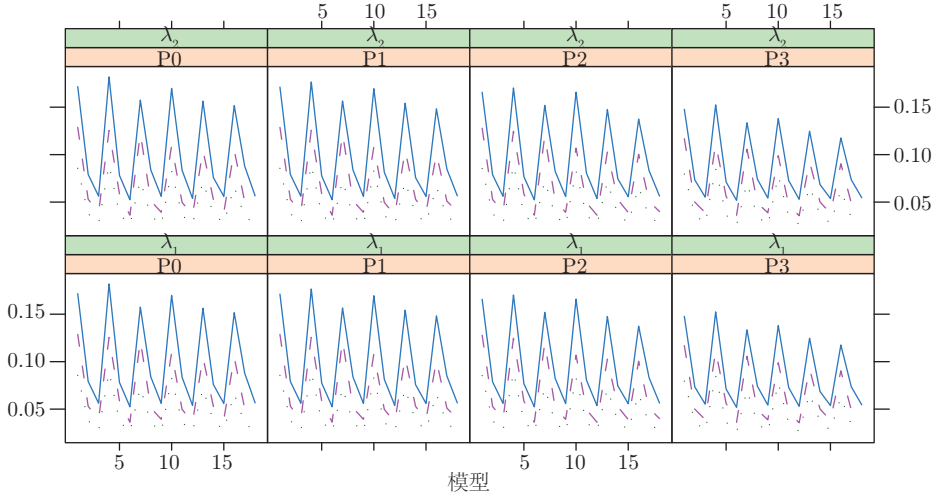
图 2　两成分混合回归模型斜率估计的偏差 (a)、标准误 (b) 和均方误差根 (c). 样本量分别为 100 (实线)、200 (虚线) 和 400 (点虚线); 横轴表示表 3 中依次构造的 18 个模型



(a)

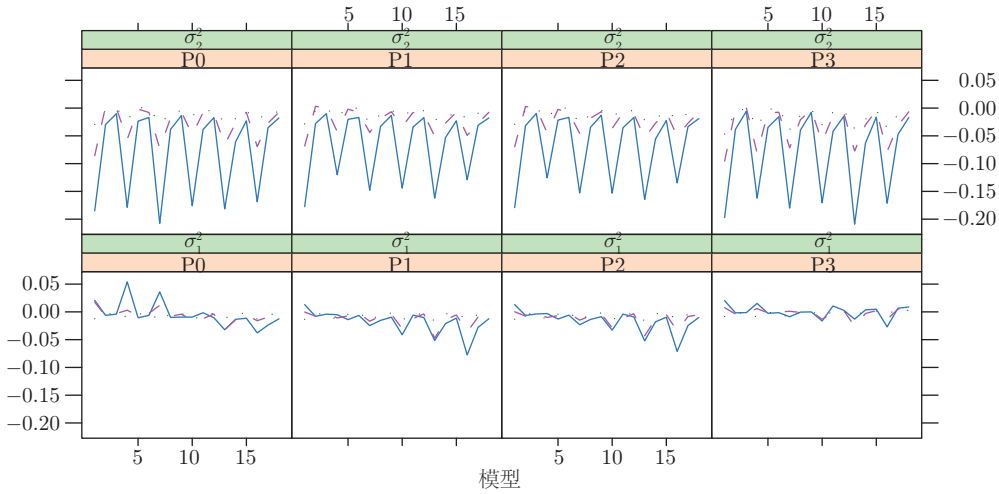


(b)

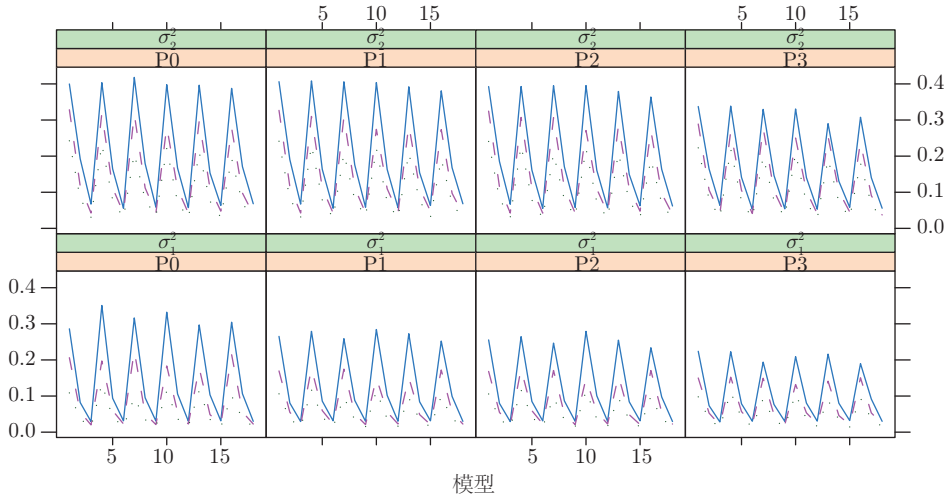


(c)

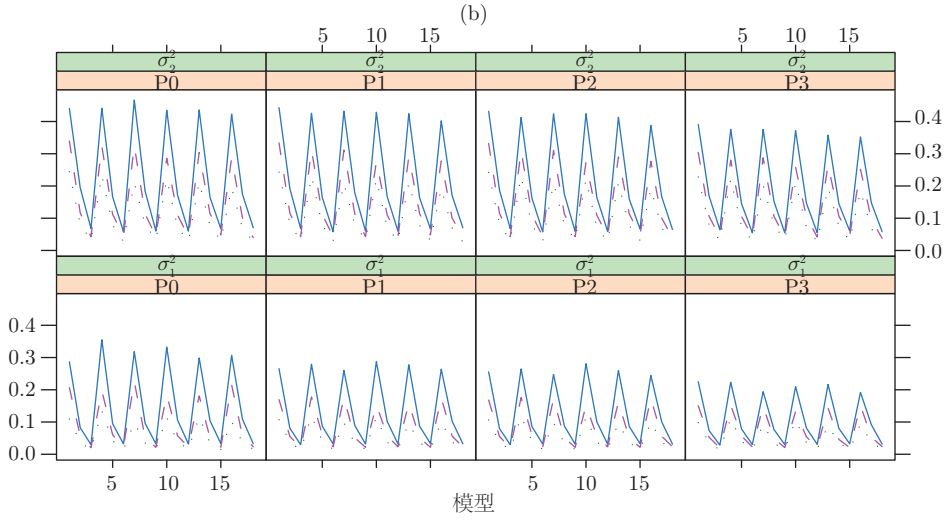
图 3　两成分混合回归模型混合比例估计的偏差 (a)、标准误 (b) 和均方误差根 (c). 样本量分别为 100 (实线)、200 (虚线) 和 400 (点虚线); 横轴表示表 3 中依次构造的 18 个模型



(a)



(b)



(c)

图 4　两成分混合回归模型方差估计的偏差 (a)、标准误 (b) 和均方误差根 (c). 样本量分别为 100 (实线)、200 (虚线) 和 400 (点虚线); 横轴表示表 3 中依次构造的 18 个模型

表 3　模型的构造 ($K = 2$ 或 $K = 3$)

#	a	b	p	#	a	b	p	#	a	b	p
1	1.00	0.0	1	7	1.00	0.0	2	13	1.00	0.0	3
2	0.50	0.0	1	8	0.50	0.0	2	14	0.50	0.0	3
3	0.25	0.0	1	9	0.25	0.0	2	15	0.25	0.0	3
4	1.00	0.1	1	10	1.00	0.1	2	16	1.00	0.1	3
5	0.50	0.1	1	11	0.50	0.1	2	17	0.50	0.1	3
6	0.25	0.1	1	12	0.25	0.1	2	18	0.25	0.1	3

表 4　方差分析表: 显著性因子按 p 值由低到高排列 (p 值 < 0.05)

参数	因子	因子	因子	因子
	$K = 2$ 偏差		$K = 2$ 均方误差根	
	P0、P1、P2 和 P3	P1、P2 和 P3	P0、P1、P2 和 P3	P1、P2 和 P3
β_{11}	a, n, c_n	a, n, c_n	a, n, c_n	a, n
β_{21}	a, n, c_n, p	a, n, p, c_n	a, n, c_n	a, n, p, b
β_{12}	a, n, c_n	a, n	a, n, b	a, n, b
β_{22}	a, n	a, n	a, n, b	a, n, b
σ_1^2	a, p, c_n, n	a, c_n, p, n	a, n, c_n	a, n, c_n
σ_2^2	a, n	a, n	a, n, p, c_n	a, n, p, c_n
π_1	a, n, c_n, p	a, n, p, c_n	a, n, p, c_n	a, n, p, c_n
	$K = 3$ 偏差		$K = 3$ 均方误差根	
	P0、P1、P2 和 P3	P1、P2 和 P3	P0、P1、P2 和 P3	P1、P2 和 P3
β_{11}	a, n, c_n, b	a, n, b	a, n, c_n, b	a, n, b, p
β_{21}	a, n, c_n, b, p	a, n, b	a, n, c_n, b	a, n, b
β_{31}	a, n, c_n	a, n	a, n, p, c_n	a, n, p
β_{12}	a, c_n, n	a	a, n, b	a, n, b
β_{22}	a, n, b, p	a, n, b	a, n, b, c_n	a, n, b
β_{32}	a, p, n, c_n	a, p, n	a, n, b	a, n, b
σ_1^2	a, b, n, c_n, p	a, b, n, c_n	a, n, b, c_n	a, n, b, c_n
σ_2^2	a, n, p, c_n	a, n, c_n, p	a, n, c_n	a, n, c_n
σ_3^2	a, n, p	a, n	a, n, b	a, n
π_1	a, n, b, c_n	a, n, b, c_n	a, n, c_n, b	a, n, b, c_n
π_2	a, c_n, p	a, c_n, n	a, n, c_n, p, b	a, n, c_n, p, b
π_3	a, n, b	a, n, b	a, n, b, c_n, p	a, n, b, c_n, p

均的 EM 迭代次数. 图 5(a) 展示了在 36 个模型构造和 3 个样本大小下估计量 P0 平均的 EM 迭代次数. 总的说来, 计算 P0 时平均的 EM 迭代次数随方差尺度系数 (a) 和成分数 (K) 增加而增加, 而随样本量 (n) 增加而减少. 在计算 P1、P2 和 P3 时, 我们发现也存在相同的现象. 为了比较 P1、P2 和 P3 对 P0 相对的运算量, 在 36 个模型构造和 3 种样本大小下, 共 5,000 次 EM 计算中, 我们统计了 P1、P2 和 P3 对 P0 时 EM 迭代次数的中位比值. 图 5(b) 和 5(c) 分别反映了 P2 和 P3 对 P0 的

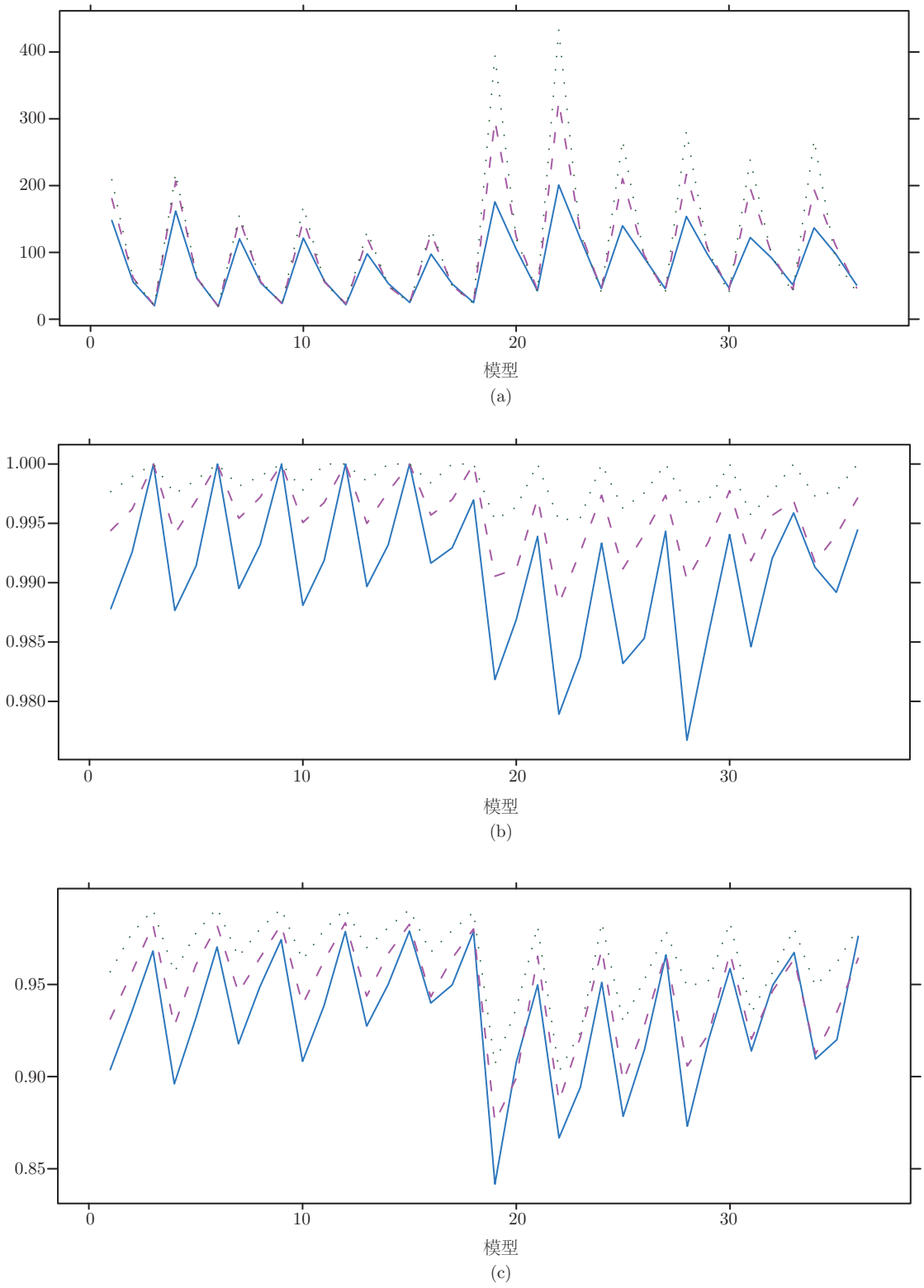


图 5　P0 估计量的平均迭代次数 (a), P2 对 P0 估计量的平均迭代次数的中位比 (b), P3 对 P0 估计量的平均迭代次数的中位比 (c). 样本量 100 (实线)、样本量 200 (虚线) 和样本量 400 (点虚线). 横轴代表模型的构造数目, 前 18 个代表两成分混合回归模型, 后 18 个代表三成分混合回归模型

中位比值变化特征. 我们发现整体上 P2 和 P3 比 P0 有更少的迭代次数, 并且该比值随 n、a 和 K 增加而减少. 这些结论说明估计量 P2 和 P3 计算上比 P0 更有效率. 在这里, 我们忽略了 P1 对 P0 的比较, 这是因为 P1 和 P0 在各种模型构造和样本量情形下, 它们的 EM 迭代次数几乎没有什么区别. 总之, 我们注意到更大的惩罚项, 即更大的 c_n, 导致更多计算效率的获得, 特别对应于更大的成分方差和更多的成分数情形时, 因为这对应着更复杂的似然函数. 这些结论与我们的假设是一致的, 即惩罚项会增加对数似然函数的凸性以使 PMLE 的计算变的更稳定.

在模拟研究中, 当成分数 $K=2$ 时, 我们计算上没有发生任何奇异性, 而在 $K=3$ 时, 当样本量为 100, 在模型 1、7、13 和 16 下, 计算 P0 时分别出现 11、5、1 和 1 次的奇异点. 在模拟中退化问题并不严重的原因可能在于较小的成分数和各成分间有较好的分离性. 尽管这样, 我们相信在更复杂模型下, 似然估计的退化现象应该是一个严重的问题, 例如, 出现更大的 K, 更大的方差比, 这时候似然计算上的退化问题在很多文献中都已提到过, 如文献 [13,26]. 虽然退化问题在计算上可以通过尝试更多的初值来缓解, 但是本文中的 PMLE 从根本上解决了这个退化问题并且减轻了计算上的负担.

总之, PMLE 理论上提出是为了解决混合回归模型似然函数无界性问题, 我们从模拟中发现 PMLE 比 MLE 更好的证据, 这是因为 PMLE 比 MLE 估计上有更好的精度, 更少的运算量以及完全消除计算上的奇异问题, 而且所有取得的这些好处不需要额外的计算成本, 只需要简单修改相应的 EM 迭代公式. 在我们的模拟研究中, 三个 PMLE, 即 P1、P2 和 P3, 在大多数参数估计中, 虽然表现基本相似, 基于定理 2.1 和 2.2 的结论, 我们建议在实际应用上还是推荐使用 P1 和 P2.

3.3 实际例子

本小节应用 PMLE 的方法来分析一个音调感知的数据集. 音调感知的数据集来自于 Cohen[27] 的实验. 在这个实验中, 受过训练的音乐家演奏一个纯基音, 然后通过一个拉伸比为 2 (stretch ratio = 2.0) 的方式加上一个电泛音, 这对应于谐音模式, 谐音通常被固定音高的传统乐器所产生. 实验中, 这个音乐家被要求在基音上调音到八度音. 调音相当于给出了可调音对基音的比值 (tuned = 2.0), 当比值为 2 时表示对应所有拉伸比下的正确调音. 这里分析的数据集产生于对同一音乐家做了 150 次调音试验的结果.

这个实验目的是研究实际的音调 (x) (即拉伸比) 和音乐家感知的音调 (y) (也即调音) 之间的关系. 数据集从 R 软件包 fpc 获得, 可参见 https://cran.r-project.org/web/packages/fpc/index.html. 实验的设计是想确定下面的两种音调感知理论, 哪一种是合适的 (参见文献 [28]). 第一种理论称为部分匹配假说 (partial matching hypothesis, PMH), 该理论认为受试者应用泛音来调音调, 相当于调拉伸比. 第二种理论称为间隔记忆假说 (interval memory hypothesis, IMH), 该理论预计受试者会以 2 : 1 基准频率调试名义上的八度音, 而不管拉伸比的值.

对应于上面两种理论的如下两成分混合回归模型被 Cohen[28] 设定:

$$\begin{cases} y_i = x_i + \varepsilon_{1i}, & \text{以概率}\,\pi, \\ y_i = 2 + \varepsilon_{2i}, & \text{以概率}\,1-\pi. \end{cases} \tag{3.1}$$

这个数据集曾经被多名学者做过分析, 参见文献 [29–31]. 根据模型 (3.1), 模型真实参数可被视为

$$\boldsymbol{\theta}_0 = (\alpha_{10}, \beta_{10}, \sigma_{10}^2, \alpha_{20}, \beta_{20}, \sigma_{20}^2, \pi_{10}) = (0, 1, \sigma_{10}^2, 2, 0, \sigma_{20}^2, \pi_{10}).$$

人们对模型 (3.1) 参数估计的主要分析方法有 Bayes 执行下的最大似然和稳健回归的 EM 算法.

对应于两种音调感知的理论, 我们发现在散点图上存在明显的两条回归线. 我们应用 PMLE 方法来估计混合回归模型的参数. 众所周知, EM 算法中初值的构造对参数估计的结果有重要的影响. 基于这个原因, 我们通过两种方法来产生初值. 一种是已经在模拟研究中使用过的基于 K 均值聚类的 EM 算法 (clustered EM, CEM). 另外一种被称为随机 EM 算法 (stochastic EM, SEM), 由 Viele 和 Tong[30] 提出和使用. 具体地, 他们通过三个步骤产生初值: 第一步, 混合比例通过 Dirichlet 分布随机产生; 第二步, 截距和斜率参数由数据集随机抽样产生; 第三步, 在正态误差下做回归分析, 得到模型的均方误差 (mean square error, MSE), 然后各成分方差参数由 $(0, \mathrm{MSE})$ 上的均匀分布随机产生. 在两种初值构造的策略中, 我们都按照相应的方法产生 10 个初值, 然后用 PMLE 获得参数估计的结果.

在这个例子中, 我们用两成分的混合回归模型拟合数据. 估计量 MLE 和 P2 下的两条回归线分别画在图 6 的左右两边, 唯一差别在于上面的两条回归线用 CEM 算法得到, 而下面的两条回归线对应于 SEM 算法. 从图 6 上看, 我们认为本文所提的 PMLE 方法能够检测哪一种音调感知理论更有说服力.

基于 CEM 和 SEM 算法, 表 5 和 6 分别列出了修正 MLE 和 PMLE 参数估计和相应标准差的结果, 通过计算参数估计量 Fisher 信息阵的逆矩阵, 按照提取其主对角线上元素平方根得到相应参数估计的标准误.

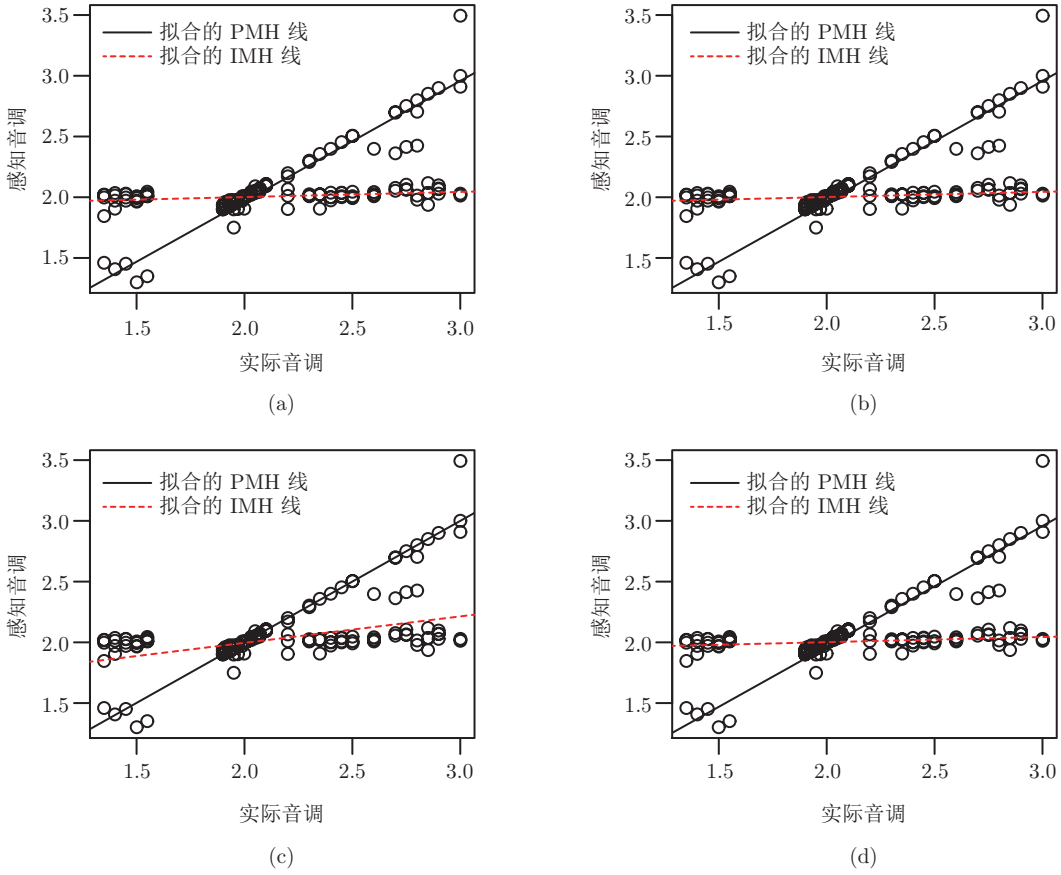


图 6 音调感知数据集来自于 Cohen[27], 估计的两条回归线: (a) 基于 CEM 算法的 MLE 估计; (b) 基于 CEM 算法的 P2 估计; (c) 基于 SEM 算法的 MLE 估计; (d) 基于 SEM 算法的 P2 估计. 实的回归线代表符合部分匹配假说 (PMH), 虚的回归线代表符合间隔记忆假说 (IMH)

表 5 CEM 算法下音调感知数据集估计量 MLE, P1 和 P2 结果的比较, 括号中数值为相应的标准误

估计量/参数	α_1	α_2	β_1	β_2	σ_1^2	σ_2^2	π_1	对数似然值
MLE	−0.0193	1.9164	0.9923	0.0425	0.0176	0.0021	0.3023	141.1984
	(0.1022)	(0.0227)	(0.0441)	(0.0102)	(0.0042)	(0.0003)	(0.0484)	
P1	−0.0193	1.9164	0.9923	0.0426	0.0177	0.0021	0.3021	141.1983
	(0.1023)	(0.0227)	(0.0442)	(0.0102)	(0.0042)	(0.0003)	(0.0484)	
P2	−0.0200	1.9161	0.9925	0.0427	0.0181	0.0022	0.3002	141.1877
	(0.1037)	(0.0229)	(0.0447)	(0.0103)	(0.0043)	(0.0004)	(0.0483)	

表 6 SEM 算法下音调感知数据集估计量 MLE, P1 和 P2 结果的比较, 括号中数值为相应的标准误

估计量/参数	α_1	α_2	β_1	β_2	σ_1^2	σ_2^2	π_1	对数似然值
MLE	0.0032	1.5608	0.9989	0.2176	0.0000	0.0471	0.3719	145.4168
	(0.0042)	(0.0971)	(0.0019)	(0.0432)	(0.0000)	(0.0069)	(0.0421)	
P1	0.0030	1.5602	0.9989	0.2176	0.0000	0.0475	0.3778	145.1228
	(0.0047)	(0.0976)	(0.0021)	(0.0434)	(0.0000)	(0.0071)	(0.0429)	
P2	−0.0196	1.9162	0.9924	0.0426	0.0179	0.0021	0.3013	141.1959
	(0.1030)	(0.0228)	(0.0445)	(0.0103)	(0.0042)	(0.0003)	(0.0484)	

根据表 5, 我们注意到在 CEM 算法下, 修正 MLE 和 P1, P2 的参数估计结果几乎是相同的. 然而, 在表 6 中, 我们发现修正 MLE, P1 和 P2 的估计结果有显著的不同, 这反映了估计量 MLE 可能收敛到参数空间的边界点, 因为存在某一个成分的方差参数非常接近于 0. 实际上, 修正 MLE 下的 $\sigma_1^2 = 2.0472 \times 10^{-5}$, P1 下为 $\sigma_1^2 = 2.5327 \times 10^{-5}$, 然而在 P2 下有 $\sigma_1^2 = 0.0179$. 值得注意的是, Viele 和 Tong [30] 也获得某成分方差参数很小的类似修正 MLE 的结果.

总之, P2 估计下的结果与两种音调感知理论较为相符. 从该例子看, 在 SEM 算法下的 P2 估计量能更好地用两成分混合回归模型拟合观测数据.

4 讨论

本文提出一种惩罚最大似然方法来估计具有正态误差混合回归模型的参数. 这个方法很好地解决了混合回归模型下似然函数无界的问题, 我们的理论证明了 PMLE 是强相合、渐近正态和有效的. 数值研究结果表明估计效率的获得不会增加计算的成本.

许多其他混合模型也有似然函数无界的现象. 例如, 似然函数无界也是切换回归模型中的一个问题, 在切换回归模型中, 有 K 种机制被 $K-1$ 个切换点 $(T_1 < T_2 < \cdots < T_{K-1})$ 暂时分离, 在不同的机制下, 被解释变量与协变量之间的关系有不同的回归模型. 更一般地, 混合回归模型的误差如果服从位置 - 尺度分布族, 则似然函数仍有无界的问题, 可参见文献 [32]. 在测量误差给定分布假设下, Wedel 和 Bijmolt [33] 使用混合模型的似然方法分析了相异点数据问题. 我们期望惩罚似然方法应用这些模型也可以获得相合和渐近有效的估计.

在混合回归模型中, 本文假设成分数 K 在研究中是先验已知的. 然而在很多实际问题中, K 是未知且是需要估计的, 这有点类似于聚类分析中类个数的确定问题. 虽然本文中没有涉及 K 的讨论, 但我们相信成分数 K 的选择是非常有挑战性的一个研究问题, 这可以留给未来做进一步研究. 对于

$K > K_0$, 我们有理由相信 PMLE 方法会对成分数 K 的选择和相合性的证明充当重要的角色.

致谢 感谢编委和两位审稿人提出的宝贵建议.

参考文献

1 Frühwirth-Schnatter S. Finite Mixture and Markov Switching Models. New York: Springer, 2006
2 McLachlan G, Peel D. Finite Mixture Models, Wiley Series in Probability and Statistics. New York: John Wiley & Sons, 2000
3 Wedel M, DeSarbo W S. A mixture likelihood approach for generalized linear models. J Classification, 1995, 12: 21–55
4 Quandt R E, Ramsey J B. Estimating mixtures of normal distributions and switching regressions. J Amer Statist Assoc, 1978, 73: 730–738
5 Schmidt P. An improved version of the Quandt-Ramsey MGF estimator for mixtures of normal distributions and switching regressions. Econometrica, 1982, 50: 501–516
6 Xu D H. Asset returns, volatility and value-at-risk. PhD Thesis. Ontario: University of Western Ontario, 2007
7 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Stat Methodol, 1977, 39: 1–38
8 Lindsay B G. Mixture Models: Theory, Geometry and Applications. Hayward: Institute for Mathematical Statistics, 1995
9 Day N E. Estimating the components of a mixture of normal distributions. Biometrika, 1969, 56: 463–474
10 Hartley M J, Mallela P. The asymptotic properties of a maximum likelihood estimator for a model of markets in disequilibrium. Econometrica, 1977, 45: 1205–1220
11 Kiefer N M. Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica, 1978, 46: 427–434
12 DeSarbo W S, Cron W L. A maximum likelihood methodology for clusterwise linear regression. J Classification, 1988, 5: 249–282
13 Hathaway R J. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Statist, 1985, 13: 795–800
14 Phillips R F. A constrained maximum-likelihood approach to estimating switching regressions. J Econometrics, 1991, 48: 241–262
15 Yao W X. A profile likelihood method for normal mixture with unequal variance. J Statist Plann Inference, 2010, 140: 2089–2098
16 Chen J H, Tan X M, Zhang R C. Consistency of penalized MLE for normal mixtures in mean and variance. Statist Sinica, 2008, 18: 443–465
17 Hennig C. Identifiability of models for clusterwise linear regression. J Classification, 2000, 17: 273–296
18 Amemiya T. Regression analysis when the dependent variable is truncated normal. Econometrica, 1973, 41: 997–1016
19 MacQueen J B. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967, 281–297
20 Stephens M. Dealing with label switching in mixture models. J R Stat Soc Ser B Stat Methodol, 2000, 62: 795–809
21 Yao W X, Lindsay B G. Bayesian mixture labeling by highest posterior density. J Amer Statist Assoc, 2009, 104: 758–767
22 Yao W X. Label switching and its solutions for frequentist mixture models. J Stat Comput Simul, 2015, 85: 1000–1012
23 Wald A. Note on the consistency of the maximum likelihood estimate. Ann Math Statist, 1949, 20: 595–601
24 Wu C F. On the convergence properties of the EM algorithm. Ann Statist, 1983, 11: 91–103
25 Green P J. On use of the EM algorithm for penalized likelihood estimation. J R Stat Soc Ser B Stat Methodol, 1990, 52: 443–452
26 Redner R A, Walker H F. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev, 1984, 26: 195–239
27 Cohen E. Inharmonic Tone Perception. PhD Thesis. Palo Alto: Stanford University, 1980
28 Cohen E A. Some effects of inharmonic partials on interval perception. Music Percept, 1984, 1: 323–349
29 De Veaux R D. Mixtures of linear regressions. Comput Statist Data Anal, 1989, 8: 227–245
30 Viele K, Tong B. Modeling with mixtures of linear regressions. Stat Comput, 2002, 12: 315–330
31 Bai X Q, Yao W X, Boyer J E. Robust fitting of mixture regression models. Comput Statist Data Anal, 2012, 56: 2347–2359
32 Tanaka K, Takemura A. Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small. Bernoulli, 2006, 12: 1003–1017

33 Wedel M, Bijmolt T H A. Mixed tree and spacial representations of dissimilarity judgments. J Classification, 2000, 17: 243–271
34 Vapnik V N, Chervonenkis A Y. On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab Appl, 1971, 16: 264–280
35 Devroye L. Bounds for the uniform deviation of empirical measures. J Multivariate Anal, 1982, 12: 72–79
36 Redner R. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann Statist, 1981, 9: 225–228

附录 A　定理 2.1 和 2.2 的证明

我们本质上按照 Chen 等[16] 的思想完成定理 2.1 的证明. 主要的想法是估计那些极端数据对对数似然的贡献. 为简单且不失一般性, 在证明中, 我们只考虑成分数 $K=2$ 的混合回归模型, 推广到一般 K 的结果是直接的, 只是证明表述上复杂一些. 显然, 当某成分方差参数 σ_k $(k=1,2)$ 趋于 0 时, 只有那些距离 $|y_i-\alpha_k-x_i^{\mathrm{T}}\beta_k|$ 非常小的观测对对数似然函数产生很大的贡献, 每一个这样的观测, 对数似然函数贡献的上界是 $|\log\sigma_k|$. 基于对这些数据个数的估计, 我们能够计算出惩罚项的大小, 以便使 PMLE 有很好的统计大样本性质. 基于这个目的, 我们先建立如下的引理.

引理 A.1　设 $\{(y_i,x_i^{\mathrm{T}})\}_{i=1}^n$ 为来自于模型 (1.1) 下的一随机样本且令 $M_0=\frac{\pi_{10}}{\sqrt{2\pi}\sigma_{10}}+\frac{\pi_{20}}{\sqrt{2\pi}\sigma_{20}}$, 当 $0<\sigma<\mathrm{e}^{-2}$ 时, 则除了一独立于 σ 的零测度集, 我们有

$$\Omega_n(\sigma)\leqslant\sqrt{n}\log n+2M_0n|\sigma\log\sigma|\quad\text{a.s.},$$

其中

$$\Omega_n(\sigma)=\sup_{(\alpha,\beta)\in\mathbb{R}^{p+1}}\#\{(y_i,x_i^{\mathrm{T}}):|y_i-\alpha-x_i^{\mathrm{T}}\beta|\leqslant|\sigma\log\sigma|,t=1,2,\ldots,n\}.$$

为了证明引理 A.1, 我们应用 VC 族理论的一些结果, 可参见文献 [34]. 为便于参考, 我们简要阐述 VC 族的一些理论和相关的记号.

记 $F_n(C)=\frac{1}{n}\sum_{i=1}^n I_{(z_i\in C)}$, 其中 $z_i=(x_i^{\mathrm{T}},y_i)$. 再设

$$C=C(\alpha,\beta,\sigma)=\{(y,x^{\mathrm{T}}):|y-\alpha-x^{\mathrm{T}}\beta|\leqslant|\sigma\log\sigma|\}.\tag{A.1}$$

定义

$$\mathcal{C}=\{C(\alpha,\beta,\sigma):(\alpha,\beta)\in\mathbb{R}^{p+1},\sigma\in\mathbb{R}^+\},\tag{A.2}$$

即 $\mathcal{C}$ 是 $\mathbb{R}^{p+1}$ 中的一个集族. 对于 $z_1,\ldots,z_n\in\mathbb{R}^{p+1}$, 记

$$\Delta^{\mathcal{C}}(z_1,\ldots,z_n)=\#\{C\cap\{z_1,\ldots,z_n\}:C\in\mathcal{C}\}\leqslant 2^n,$$
$$m^{\mathcal{C}}(n)=\sup\{\Delta^{\mathcal{C}}(z_1,\ldots,z_n):z_1,\ldots,z_n\in\mathbb{R}^{p+1}\}.$$

如果对某 $n\geqslant 1$, 有 $m^{\mathcal{C}}(n)<2^n$, 则 $\mathcal{C}$ 称为 Vapnik-Cervonenkis 族 (简称为 VC 族). 最小的 $n\geqslant 1$, 满足 $m^{\mathcal{C}}(n)<2^n$, 称为 VC 族 $\mathcal{C}$ 的指数, 并记为 $V(\mathcal{C})$. 从上面的定义知, 下面关于 VC 族的两条性质是显然的.

引理 A.2　令 $\mathcal{C}$、$\mathcal{C}_1$ 和 $\mathcal{C}_2$ 为 VC 族, 则下面的两个集族也是 VC 族:

(1) $\mathcal{D}\subset\mathcal{C}$;

(2) $\mathcal{C}_1\cap\mathcal{C}_2=\{C_1\cap C_2:C_1\in\mathcal{C}_1,C_2\in\mathcal{C}_2\}$.

在实际应用中, 一些有趣的集类都是 VC 族, 例如, 令 $\mathcal{D}=\{\{z: z^{\mathrm{T}}\theta \leqslant t\}, (\theta^{\mathrm{T}}, t)\in \mathbb{R}^{p+1}\}$, 则 $\mathcal{D}$ 是 VC 族, 因为 $m^{\mathcal{D}}(n)\leqslant (2n)^p$. 其结果是, 通过应用前面 VC 族的两条性质, 我们知道 (A.2) 中的 $\mathcal{C}$ 是 VC 族.

下面的引理来自于 Devroye[35].

引理 A.3 令 $\mathcal{C}$ 为可测的 VC 族. 对所有的 $M>0$, $n\geqslant 1$, 如果 $v\geqslant V(\mathcal{C})$, 则

$$\mathrm{P}\Big(\sup_{\mathcal{C}}|F_n(C)-\mathrm{E}F_n(C)|>M\Big)\leqslant 4\mathrm{e}^8\left(\frac{n^2\mathrm{e}}{v}\right)^v\exp(-2M^2).$$

有了这些准备工作, 我们可以证明引理 A.1.

引理 A.1 的证明 令 C 和 $\mathcal{C}$ 分别按照 (A.1) 和 (A.2) 定义, 显然有

$$\begin{aligned}&\sup_{(\alpha,\beta^{\mathrm{T}})\in\mathbb{R}^{p+1}}\#\{(y_i,x_i^{\mathrm{T}}):|y_i-\alpha-x_i^{\mathrm{T}}\beta|\leqslant|\sigma\log\sigma|, t=1,2,\ldots,n\}\\&=\sup_{C\in\mathcal{C},\sigma\text{ 固定}}nF_n(C)\leqslant\sup_{C\in\mathcal{C}}n|F_n(C)-\mathrm{E}F_n(C)|+\sup_{C\in\mathcal{C},\sigma\text{ 固定}}n|\mathrm{E}F_n(C)|\\&\leqslant\sup_{C\in\mathcal{C}}n|F_n(C)-\mathrm{E}F_n(C)|+\sup_{C\in\mathcal{C},\sigma\text{ 固定}}n|\mathrm{E}F_n(C)|.\end{aligned}$$

容易验证

$$\sup_{C\in\mathcal{C},\sigma\text{ 固定}}n|\mathrm{E}F_n(C)|\leqslant 2M_0n|\sigma\log\sigma|. \tag{A.3}$$

在引理 A.3 中的常数 M 用 $\log n/\sqrt{n}$ 替换, 然后应用 Borel-Cantelli 引理, 可得

$$\sup_{C\in\mathcal{C}}n|F_n(C)-\mathrm{E}F_n(C)|\leqslant\sqrt{n}\log n,\quad \text{a.s.} \tag{A.4}$$

结合 (A.3) 和 (A.4), 并注意到在 (A.4) 中的零测集独立于 σ, 我们就得到引理 A.1 中的结果, 即

$$\Omega_n(\sigma)\leqslant\sqrt{n}\log n+2M_0n|\sigma\log\sigma|\quad\text{a.s.}$$

证毕. □

在给定引理 A.1 的结果下, 定理 2.1 的证明直接可根据文献 [16,23] 中的定理得到.

定理 2.1 的证明 令 $K_0(x_i,\boldsymbol{\theta}_0)=\mathrm{E}_0^{Y|X}[\log f(y_i\mid x_i,\boldsymbol{\theta}_0)]$ 和 $\mathrm{E}_0^X K(X_1,\boldsymbol{\theta}_0)=K_0(\boldsymbol{\theta}_0)$, 其中算子 $\mathrm{E}_0^{Y|X}$ 表示密度 $f(y\mid x,\boldsymbol{\theta}_0)$ 下的条件期望, 且 E_0^X 表示分布 F_x 下的期望, 也即, $X_1\sim F_x(x)$. 不失一般性, 假设 $K_0(\boldsymbol{\theta}_0)<\infty$ 和 $\sigma_1\leqslant\sigma_2$.

分割参数空间 Θ 为如下的三部分:

$$\Theta_1=\{\boldsymbol{\theta}\in\Theta:\sigma_1\leqslant\sigma_2\leqslant\varepsilon_0\},\quad \Theta_2=\{\boldsymbol{\theta}\in\Theta:\sigma_1\leqslant\tau_0,\sigma_2\geqslant\varepsilon_0\},\quad \Theta_3=\Theta-\Theta_1\cup\Theta_2,$$

其中 $\varepsilon_0>\tau_0>0$ 为两个很小的正常数, 满足如下的四个条件:

(1) $0<\tau_0\leqslant\varepsilon_0<\mathrm{e}^{-2}$;

(2) $(4M_0+2)\varepsilon_0(\log\varepsilon_0)^2\leqslant 1$;

(3) $-\log\varepsilon_0-\frac{1}{2}(\log\varepsilon_0)^2\leqslant 2K_0(\boldsymbol{\theta}_0)-4$;

(4) $(2M_0+1)\tau_0(\log\tau_0)^2\leqslant\frac{2}{5}\delta$,

其中 δ 为一个正常数且在后面证明中被定义.

定理 2.1 的证明分为三步.

第 1 步 对于子集 Θ_1, 我们要说明

$$\sup_{\boldsymbol{\theta}\in\Theta_1} p\ell_n(\boldsymbol{\theta}) - p\ell_n(\boldsymbol{\theta}_0) \to -\infty, \quad \text{a.s.,} \ \text{当 } n\to\infty \text{ 时}.$$

定义如下的指标集:

$$A = \{t : |y_i - \alpha_1 - x_i^{\mathrm{T}}\beta_1| \leqslant |\sigma_1 \log\sigma_1|, t = 1, 2, \ldots, n\},$$
$$B = \{t : |y_i - \alpha_2 - x_i^{\mathrm{T}}\beta_2| \leqslant |\sigma_2 \log\sigma_2|, t = 1, 2, \ldots, n\}.$$

对任意的指标集 S, $S \subset \{1, 2, \ldots, n\}$, 定义

$$\ell_n(\boldsymbol{\theta} \mid S) = \sum_{t\in S} \log\left[\frac{\pi}{\sigma_1}\phi\left(\frac{y_i - \alpha_1 - x_i^{\mathrm{T}}\beta_1}{\sigma_1}\right) + \frac{1-\pi}{\sigma_2}\phi\left(\frac{y_i - \alpha_2 - x_i^{\mathrm{T}}\beta_2}{\sigma_2}\right)\right],$$

其中 $\phi(.)$ 表示标准正态密度函数. 注意到

$$\ell_n(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta} \mid A) + \ell_n(\boldsymbol{\theta} \mid A^cB) + \ell_n(\boldsymbol{\theta} \mid A^cB^c).$$

对于等式右边第一项, 有

$$\ell_n(\boldsymbol{\theta} \mid A) \leqslant (\#A) \times \log\left(\frac{1}{\sigma_1}\right).$$

当一个观测数据 (x_i^{T}, y_i) 来自于指标集 A^cB 时, 有

$$\begin{aligned}\frac{1}{\sigma_1}\exp\left(-\frac{(y_i - \alpha_1 - x_i^{\mathrm{T}}\beta_1)^2}{2\sigma_1^2}\right) &\leqslant \frac{1}{\sigma_1}\exp\left(-\frac{1}{2}(\log\sigma_1)^2\right)\\ &= \exp\left(-\log\sigma_1 - \frac{1}{2}(\log\sigma_1)^2\right)\\ &\leqslant \exp\left(-\log\sigma_2 - \frac{1}{2}(\log\sigma_2)^2\right)\\ &\leqslant \frac{1}{\sigma_2}.\end{aligned}$$

因而,

$$\ell_n(\boldsymbol{\theta} \mid A^cB) \leqslant (\#B) \times \log\left(\frac{1}{\sigma_2}\right).$$

根据引理 A.1 和假设 (A1)–(A3), 有

$$\ell_n(\boldsymbol{\theta} \mid A) + p_n(\boldsymbol{\theta}) \leqslant (2M_0 + 1)n\sigma_1(\log\sigma_1)^2 \leqslant (2M_0 + 1)n\varepsilon_0(\log\varepsilon_0)^2,$$
$$\ell_n(\boldsymbol{\theta} \mid A^cB) + p_n(\boldsymbol{\theta}) \leqslant (2M_0 + 1)n\sigma_2(\log\sigma_2)^2 \leqslant (2M_0 + 1)n\varepsilon_0(\log\varepsilon_0)^2.$$

对于指标集在 A^cB^c 中的观测数据, 有

$$\ell_n(\boldsymbol{\theta} \mid A^cB^c) \leqslant \sum_{t\in A^cB^c}\left[-\log\varepsilon_0 - \frac{1}{2}(\log\varepsilon_0)^2\right].$$

注意到 $\frac{n}{2} \leqslant \#A^cB^c \leqslant n$, 再分别结合条件 (2)、(3) 和前面三项的三个不等式, 得到

$$p\ell_n(\boldsymbol{\theta}) \leqslant n + \frac{1}{2}n(2K_0(\boldsymbol{\theta}_0) - 4).$$

因为 $p\ell_n(\boldsymbol{\theta}_0) = nK_0(\boldsymbol{\theta}_0) + o(n)$, a.s., 我们有

$$p\ell_n(\boldsymbol{\theta}) - p\ell_n(\boldsymbol{\theta}_0) \leqslant -n + o(n), \quad \text{a.s.},$$

故当 $n \to \infty$ 时, 以概率 1, $\widetilde{\boldsymbol{\theta}}_n \notin \Theta_1$.

第 2 步 在这一步, 我们要说明

$$\sup_{\boldsymbol{\theta}\in\Theta_2} p\ell_n(\boldsymbol{\theta}) - p\ell_n(\boldsymbol{\theta}_0) \to -\infty, \quad \text{a.s., 当 } n \to \infty \text{ 时}.$$

在 Θ_2 上定义距离函数

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{T}}) = \sum_{i=1}^{2p+3} |\mathrm{arctan}\theta_i - \mathrm{arctan}\theta_i^{\mathrm{T}}|,$$

其中 θ_i 表示 $\boldsymbol{\theta}$ 的第 i 个分量. 在这个距离 d 下, Θ_2 是全有界集并且在包含所有极限点时可以使 Θ_2 紧致化. 记 $\bar{\Theta}_2$ 表示紧集. 进一步, 定义

$$h(y \mid x, \boldsymbol{\theta}) = a\frac{\pi}{\sqrt{2}}\phi\left(\frac{y - \alpha_1 - x^{\mathrm{T}}\beta_1}{\sqrt{2}\sigma_1}\right) + b\frac{(1-\pi)}{\sigma_2}\phi\left(\frac{y - \alpha_2 - x^{\mathrm{T}}\beta_2}{\sigma_2}\right),$$

其中

$$a = \begin{cases} 0, & \sigma_1 = 0, \quad \text{或 } \alpha_1 = \infty, \quad \text{或 } \|\beta_1\| = \infty, \\ 1, & \text{其他}, \end{cases} \qquad b = \begin{cases} 0, & \sigma_2 = \infty, \quad \text{或 } \alpha_2 = \infty, \quad \text{或 } \|\beta_2\| = \infty, \\ 1, & \text{其他}. \end{cases}$$

关于函数 $h(y \mid x, \boldsymbol{\theta})$, 容易验证下面的三条性质成立:

(1) $\int_{-\infty}^{+\infty} h(y \mid x, \boldsymbol{\theta})dy \leqslant 1$;

(2) $\sup_{\boldsymbol{\theta}\in\bar{\Theta}_2} h(y \mid x, \boldsymbol{\theta}) \leqslant \frac{1}{\varepsilon_0}$;

(3) $h(y \mid x, \boldsymbol{\theta})$ 作为 $\boldsymbol{\theta}$ 的函数是非负的且关于密度 $f(y \mid x, \boldsymbol{\theta}_0)$ 在紧集 $\bar{\Theta}_2$ 上几乎处处连续.

对任意的 $\boldsymbol{\theta} \in \bar{\Theta}_2$, 定义 $L_0(x_i, \boldsymbol{\theta}) = \mathrm{E}_0^{Y|X} \log h(y_i \mid x_i, \boldsymbol{\theta})$. 注意到

$$\frac{1}{n}\sum_{i=1}^{n} L_0(x_i, \boldsymbol{\theta}) \to \mathrm{E}_0 L_0(x_1, \boldsymbol{\theta}) = L_0(\boldsymbol{\theta}) \quad \text{a.s.}$$

令 $\boldsymbol{\theta}^* = \arg\cdot\max_{\boldsymbol{\theta}\in\bar{\Theta}_2} L_0(\boldsymbol{\theta})$, 再设

$$\delta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}[K_0(x_i, \boldsymbol{\theta}_0) - L_0(x_i, \boldsymbol{\theta}^*)] = K_0(\boldsymbol{\theta}_0) - L_0(\boldsymbol{\theta}^*).$$

我们可以选择 ε_0 足够小, 使得 $\boldsymbol{\theta}_0 \notin \bar{\Theta}_2$, 因而, $h(y_i \mid x_i, \boldsymbol{\theta}^*) \neq f(y_i \mid x_i, \boldsymbol{\theta}_0)$. 结合假设 (C1), 应用条件 Jensen 不等式, 有 $L_0(x_i, \boldsymbol{\theta}^*) < K_0(x_i, \boldsymbol{\theta}_0)$ a.s., 因而,

$$L_0(\boldsymbol{\theta}^*) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} L_0(x_i, \boldsymbol{\theta}^*) < \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} K_0(x_i, \boldsymbol{\theta}_0) = K_0(\boldsymbol{\theta}_0).$$

定义 $\delta = \delta(\tau_0) = K_0(\boldsymbol{\theta}_0) - L_0(\boldsymbol{\theta}^*)$, 则 $\delta(\tau_0) > 0$. 显然, $\delta(\tau_0)$ 是 τ_0 的递减函数. 这隐含我们能找到足够小的 τ_0 使得条件 (4) 成立.

根据 $\sup_{\boldsymbol{\theta}\in\bar{\Theta}_2}\sum_{i=1}^n \log h(y_i \mid x_i,\boldsymbol{\theta}) \leqslant o(n)+n(K_0(\boldsymbol{\theta}_0)-\frac{9}{10}\delta)$ a.s. 和条件 (4), 我们有

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_2} p\ell_n(\boldsymbol{\theta})-p\ell_n(\boldsymbol{\theta}_0) \leqslant & \sup_{\sigma_1\leqslant\varepsilon_0}\left[\sum_{t\in A}\log\frac{1}{\sigma_1}-p_n(\boldsymbol{\theta})\right]+\sup_{\boldsymbol{\theta}\in\bar{\Theta}_2}\sum_{i=1}^n \log h(y_i \mid x_i,\boldsymbol{\theta})\\ & -\sum_{i=1}^n \log f(y_i \mid x_i,\boldsymbol{\theta}_0)+o(n)\\ \leqslant & (2M_0+1)n\tau_0(\log\tau_0)^2-\frac{9}{10}\delta n+o(n)\\ \leqslant & -\frac{1}{2}\delta n+o(n)\to-\infty \quad \text{a.s.}\end{aligned}$$

第 3 步 根据上面的两步知, $\widetilde{\boldsymbol{\theta}}_n\in\Theta_3$ 以概率 1 成立. 对于 $\boldsymbol{\theta}\in\Theta_3$, 有 $p_n(\boldsymbol{\theta})=o(n)$. 因为对每个 $n\in\mathbb{N}$, 有 $p\ell_n(\widetilde{\boldsymbol{\theta}}_n)-p\ell_n(\boldsymbol{\theta}_0)\geqslant c$, 此时 $\ell_n(\widetilde{\boldsymbol{\theta}}_n)-\ell_n(\boldsymbol{\theta}_0)\geqslant o(n)$. 限制 $\boldsymbol{\theta}\in\Theta_3$ 等价于给方差参数设定了一个正的下界. 按照文献 [36], 可得 $\mathrm{P}(\lim_{n\to\infty}\widetilde{\boldsymbol{\theta}}_n=\boldsymbol{\theta}_0)=1$. 这完成了定理 2.1 的证明. □

基于定理 2.1, 除了一些技术细节, 定理 2.2 的证明是简单和直接的.

定理 2.2 的证明 类似于 Kiefer [11] 的方法, 首先, 容易证明 $\boldsymbol{I}(\boldsymbol{\theta}_0)$ 为一个有限的正定阵. 根据惩罚函数 $p\ell_n(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}_0$ 的光滑性以及 PMLE 是惩罚对数似然函数的最大值点, 我们有

$$\frac{\partial\ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}+p_n'(\hat{\boldsymbol{\theta}}_n)=\frac{\partial p\ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}=0.$$

根据中值定理, 有

$$\frac{\partial\ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}-\frac{\partial\ell_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}=\frac{\partial^2\ell_n(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0),$$

其中 $\boldsymbol{\theta}^*$ 位于 $\boldsymbol{\theta}_0$ 和 $\hat{\boldsymbol{\theta}}_n$ 之间. 直接地, 可以验证

$$\frac{1}{n}\frac{\partial^2\ell_n(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}=-\boldsymbol{I}(\boldsymbol{\theta}_0)+o_p(1),\quad \frac{\partial\ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}=-p_n'(\hat{\boldsymbol{\theta}}_n)=o(\sqrt{n}),\quad \frac{1}{\sqrt{n}}\frac{\partial\ell_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\to N(0,\boldsymbol{I}(\boldsymbol{\theta}_0)).$$

因而, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\to N(0,\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0))$. 这完成了定理 2.2 的证明. □

A penalized maximum likelihood approach for estimating mixture of regression models

Jianjun Xu, Xianming Tan & Runchu Zhang

Abstract This paper considers the parameter estimation problem for mixture of regression models with normal errors. Because of unboundedness of the likelihood function, the ordinary maximum likelihood estimator for mixture of regression models does not exist. We propose a penalized maximum likelihood approach for estimating the parameters in a mixture of regression models. We prove that the penalized maximum likelihood estimator (PMLE) is strongly consistent and asymptotically normal. Through extensive simulation studies, the proposed new estimator is shown to perform quite well in terms of estimation accuracy. We also present a tone perception example to illustrate the applications of the theoretical results.

Keywords **mixture regression, penalized maximum likelihood estimator, strong consistency, asymptotic normality**

MSC(2010) **91B82, 91B84**

doi: 10.1360/SCM-2018-0132