中国科学 : 数学　2019 年　第 49 卷　第 9 期 : 1225 ∼ 1250
SCIENTIA SINICA Mathematica
论　文

偏正态混合模型的惩罚极大似然估计

金立斌[1], 许王莉[2], 朱利平[3], 朱力行[4,5*]

1. 上海立信会计金融学院统计与数学学院, 上海 201209;
2. 中国人民大学统计学院, 北京 100872;
3. 中国人民大学统计与大数据研究院, 北京 100872;
4. 香港浸会大学数学系, 香港;
5. 北京师范大学统计学院, 北京 100875
E-mail: jinlb1987@hotmail.com, xwlbnu@163.com, zhuliping.stat@yahoo.com, lzhu@hkbu.edu.hk

收稿日期: 2018-04-08; 接受日期: 2018-07-21; 网络出版日期: 2019-05-30; * 通信作者
国家自然科学基金 (批准号: 11801370, 11471335, 11731011 和 11671042)、上海市自然科学基金 (批准号: 18ZR1427200) 和上海立信会计金融学院统计学一级学科 (批准号: 17TS001) 资助项目

摘要　在分析具有异质性和非对称性数据时, 偏正态混合模型提供一种比经典的 Gauss 混合模型更为灵活的建模方式. 然而, 由于无界的似然函数和发散的形状参数, 该模型的极大似然估计并未被正确定义, 进一步导致不理想的推断过程. 为同时解决这两个问题, 本文基于惩罚似然提出一种新的估计方案, 并证明在混合分布的类别个数大于或等于真实的类别个数时, 相应的惩罚极大似然估计是强相合的. 同时, 本文也提出相应的惩罚 EM (expectation maximization) 算法来计算惩罚估计. 最后, 通过模拟分析与现有方法比较研究估计方法在有限样本下的表现, 并采用两个实例说明方法的有效性.

关键词　似然退化　边界估计　偏正态混合模型　惩罚极大似然估计　强相合性

MSC (2010) 主题分类　62F12

1　引言

有限混合模型 (finite mixture models) 广泛地用于科学研究, 其中一个重要的特例就是 Gauss 混合模型 (Gaussian mixture models, GMIX). 近年来, 在处理具有非对称特征的异质性数据时, 学界逐渐开始关注更为广义的偏正态混合模型 (skew normal mixture models, SNMIX). 该模型广泛应用于经济学、生物学和流行病学等诸多领域, 相关实例可参见文献 [1–3]. 同时, 偏正态混合回归模型也已在心理测量学 [4] 和音乐学 [5] 等方面取得了较好的应用.

在混合模型领域, 混合密度的确定一直是受持续关注的热点问题之一, 相关方法可参见文献 [6,7].

英文引用格式: Jin L B, Xu W L, Zhu L P, et al. Penalized maximum likelihood estimation for skew normal mixtures (in Chinese). Sci Sin Math, 2019, 49: 1225–1250, doi: 10.1360/SCM-2018-0256

其中, 经典的极大似然估计 (maximum likelihood estimation, MLE) 依然是最常用的方法之一, 因其在正则参数模型下具有渐近有效性. 在偏正态混合模型中, 目前也出现了多种方案来估计感兴趣的参数, 但大部分依然是关于 MLE 及其计算过程 (参见文献 [1,2]).

然而, 对于 Gauss 混合模型, 经典的极大似然估计无法正确定义 (参见文献 [8,9]). 偏正态混合模型也存在相同的问题, 极大似然估计不相合 (inconsistence) 经常出现. 这种现象在我们分析经典的 Iris 数据时得到了体现. Iris 数据最早由 Fisher[10] 引入统计分析, 现已成为在许多统计推断问题中广泛使用的标杆数据, 如分布理论. 由于该数据观测了 Iris 花 3 个亚类的 150 个样本的 4 个特征, Basford 和 McLachlan[11]、Ray 和 Lindsay[12] 及 Boldea 和 Magnus[13] 利用类别个数为 3 的多元正态混合模型对该数据进行拟合并采用极大似然方法来估计感兴趣的参数.

长期以来, 统计学界一直忽略了对 Iris 数据的边际特征分析, 而我们的分析揭示了极大似然估计的发散问题. 我们考虑数据中两个特征: 萼片长度和花瓣宽度, 并采用类别个数为 3 的 GMIX 和 SNMIX 模型进行拟合. 从分析结果我们发现, 由于存在一个类别方差的估计量接近于 0, GMIX 的 MLE 受到无界的似然函数困扰. 同时, SNMIX 在拟合该数据时也存在与 GMIX 相同的问题, 并且其中一个成分还存在 MLE 发散的问题. 其形状参数的绝对值远大于 100, 这被 Azzalini 和 Arellano-Valle[14] 推荐作为形状参数发散的标志. 事实上, 形状参数的发散在一定程度上是为了处理 Azzalini 和 Capitanio[15] 提出的所谓的数据的偏斜 (frontier) 特性. 上述结果表明, 极大似然方法在著名的 Iris 数据中是失效的. 读者可从第 6.1 小节获得本数据详细的分析结果.

为了避免混合模型的似然退化 (likelihood degeneracy), 目前学界主要存在两种基于似然的方法来重获相合性和有效性. 第一种是限制极大似然估计 (constraint MLE). Redner[16] 证明了在每一个包含参数真值的紧致子空间内, 极大似然估计存在且是强相合的. Hathaway[17] 推荐采用一种非紧致约束下的限制极大似然估计. 然而, Chen 等[18] 指出参数空间的压缩可能将真实参数排除在外. 其他学者关注惩罚似然方法, 其优势是可以在不改变参数空间的前提下处理似然发散问题. Ciuperca 等[19] 和 Chen 等[18] 加入了关于成分方差的惩罚项, 并分别证明了所提出的惩罚极大似然估计 (penalized maximum likelihood estimators, PMLE) 的强相合性. 在多元正态混合模型中, Chen 和 Tan[20] 证明了相应的惩罚估计的相合性; Alexandrovich[21] 对其理论条件进行了进一步修正. Chen[22] 梳理了一般意义下非参数极大似然估计的相合性结果, 尤其是正态混合模型下的惩罚似然估计. Chen 等[23] 提出了针对 Gamma 混合模型的 PMLE, 并证明了其强相合性.

对于偏正态分布, Azzalini 和 Capitanio[15] 发现即使数据能被形状参数 λ 有限的偏正态模型很好近似, λ 的 MLE 依然可能出现在边界上. Azzalini 和 Arellano-Valle[14] 发现 λ 的发散对于模型的推断过程具有不良影响, 且其在有限样本下发生的概率是不可忽略的, 并会造成非常大的计算负担. 为了获得 λ 的一个可信的估计, DiCiccio 和 Monti[24] 采用轮廓似然比检验提出了一个修正估计. Azzalini 和 Arellano-Valle[14] 提出了一种惩罚极大似然估计. 然而, 由于无界的似然函数, DiCiccio 和 Monti[24] 的方法并不能直接应用于偏正态混合模型. 事实上, 偏正态混合模型中 λ_i 的发散已被文献 [2] 中较差的模拟表现反映, 尤其是其中表 3 的结果.

在本文中, 为同时克服偏正态混合模型中无界的似然函数和发散的形状参数, 我们采用在对数似然函数中加入惩罚项的方法, 并提出一种惩罚极大似然估计. 本文剩余部分从以下几方面展开. 第 2 节概述一些理论准备工作, 包括偏正态混合模型、技术引理和惩罚项的选择. 第 3 节提供两种情形下所提出的惩罚极大似然估计强相合性的严格证明过程. 第 4 节列出惩罚 EM 算法. 模拟结果和两个实例分析分别在第 5 和 6 节. 第 7 节给出本文的结论. 本文定理 3.1、3.3 和 3.5 的证明过程均在附录 A 给出.

2 理论准备

2.1 偏正态混合模型

Azzalini [25] 提出了偏正态分布 (skew normal distribution), 其密度函数形式如下:

$$f_{SN}(x;\theta)=\frac{2}{\sigma}\phi\bigg(\frac{x-\mu}{\sigma}\bigg)\Phi\bigg(\lambda\frac{x-\mu}{\sigma}\bigg),\quad x\in\mathbb{R},\tag{2.1}$$

其中参数 $\theta=(\mu,\sigma^2,\lambda)\in\Theta\subseteq\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}$. 同时, $\phi(\cdot)$ 和 $\Phi(\cdot)$ 分别是标准正态分布的密度函数和累积分布函数. 显然, 密度函数 (2.1) 依赖于 μ、σ^2 和 λ, 这三个参数分别控制了该分布的位置、尺度和形状 (偏斜度).

基于偏正态分布 (2.1), Lin 等 [1] 提出了相应的偏正态混合模型. 假定混合模型的类别个数为 p, 则其密度函数为

$$f(x;\Psi)=\sum_{k=1}^{p}\pi_k f_{SN}(x;\theta_k)=\int f_{SN}(x;\theta)d\Psi(\theta),\tag{2.2}$$

其中 π_k 和 $\theta_k=(\mu_k,\sigma_k^2,\lambda_k)$ 分别表示混合比率和类别参数, $k=1,\dots,p$. 此外, 我们采用符号 Ψ 同时表示偏正态混合模型内的所有参数和具有如下形式的累积分布函数:

$$\Psi(\theta)=\sum_{k=1}^{p}\pi_k I(\theta_k\leqslant\theta)=\sum_{k=1}^{p}\pi_k I(\mu_k\leqslant\mu,\sigma_k^2\leqslant\sigma^2,\lambda_k\leqslant\lambda),$$

其中 $I(\cdot)$ 是示性函数. 进一步地, Ψ 的参数空间可表示为

$$\begin{aligned}\Gamma=\bigg\{&\Psi=(\pi_1,\dots,\pi_p,\mu_1,\dots,\mu_p,\sigma_1,\dots,\sigma_p,\lambda_1,\dots,\lambda_p):\\&0\leqslant\pi_k\leqslant1,\sum_{k=1}^{p}\pi_k=1,-\infty<\mu_k,\lambda_k<+\infty,\sigma_k\geqslant0\bigg\}.\end{aligned}$$

同时假定真实参数为 Ψ_0, 且 $\Psi_0\in\Gamma$.

假设存在一个样本量为 n 的随机样本 $\{X_1,\dots,X_n\}$ 来自上述偏正态混合模型 (2.2), 那么, 其对数似然函数即可写作

$$\ell_n(\Psi)=\sum_{i=1}^{n}\log f(X_i;\Psi)=\sum_{i=1}^{n}\log\bigg\{\sum_{k=1}^{p}\frac{2\pi_k}{\sigma_k}\phi\bigg(\frac{X_i-\mu_k}{\sigma_k}\bigg)\Phi\bigg(\lambda_k\frac{X_i-\mu_k}{\sigma_k}\bigg)\bigg\}.\tag{2.3}$$

显然, 当 $\mu_k\to X_i$ 和 $\sigma_k\to0$ 且其他参数固定时, 对数似然函数将发散到无穷 [19]. 所以, 对于任意给定的样本量 n, 对数似然函数 $\ell_n(\Psi)$ 在参数空间 Γ 上是无界的. 因此, Ψ 的全局极大似然估计是不相合的 [22]. 同时, 似然比检验统计量将失去优良的渐近性质.

偏正态混合模型中另一个不良性质是 λ_k 的极大似然估计会发散到无穷. 理论上, 这是因为 $\Phi(\cdot)$ 是一个单调递增的函数, 所以, (2.3) 中的 $\ell_n(\Psi)$ 在参数空间 Γ 上求最大化时将出现如下结果:

$$\hat{\lambda}_k=\begin{cases}\infty, & \displaystyle\sum_{i=1}^{n}I(X_i>\mu_k)=n,\\ -\infty, & \displaystyle\sum_{i=1}^{n}I(X_i<\mu_k)=n.\end{cases}\tag{2.4}$$

此时, 限制极大似然估计对减轻形状参数估计的发散性并不是非常有效. 为了说明该问题, 我们不妨假定在参数空间 Γ 上添加一个额外的限制条件 $\max_k\{|\lambda_k|\}\leqslant C$, 其中 C 是一个充分大的正常数. 此时, $\ell_n(\Psi)$ 取到最大值当且仅当形状参数收敛到限制后参数空间的边界上, 表示为

$$\max_k|\hat{\lambda}_k|=C,\quad \text{若}\ \max_k\left|\sum_{i=1}^n\mathrm{sgn}(X_i-\mu_k)\right|=n, \tag{2.5}$$

其中 $\mathrm{sgn}(\cdot)$ 是符号函数. 因此, 限制极大似然估计的效果较差.

为解决上述极大似然方法存在的两个问题, 我们考虑采用惩罚似然方法, 其定义如下:

$$\begin{aligned} &p\ell_n(\Psi)=\ell_n(\Psi)+p_n(\Psi),\\ &p_n(\Psi)=\sum_{k=1}^p p_{1n}(\sigma_k)+\sum_{k=1}^p p_{2n}(\lambda_k). \end{aligned} \tag{2.6}$$

因此, Ψ 的惩罚极大似然估计可定义为

$$\tilde{\Psi}=\arg\max_{\Psi}\ p\ell_n(\Psi).$$

在给定的合理惩罚项下, (2.6) 中的惩罚似然函数 $p\ell_n(\Psi)$ 在参数空间 Γ 上是有界的, 这样就保证了惩罚极大似然估计的存在性. 同时, 为了获得 $\tilde{\Psi}$ 的相合性, 必须仔细选择惩罚函数 $p_{1n}(\sigma)$ 和 $p_{2n}(\lambda)$. 我们选择这样的 $p_{1n}(\sigma)$: 当 σ 趋向于 0 或者无穷时, 其函数值趋向于负无穷; 并挑选这样的 $p_{2n}(\lambda)$: 当 $|\lambda|$ 趋向于无穷时, 其函数值趋向于负无穷.

2.2 技术引理

在 Gauss 混合模型中, Chen 等 [18] 提出了一种建立惩罚估计的强相合性的新技术. 基于 Bernstein 不等式, 他们证明了一个深刻的结论: 落在位置参数的一个小邻域内的样本量存在一致的上界. 同时, Chen 等 [18] 也指出, 正态假设在证明过程中并未起到至关重要的作用, 这也为该结论的拓展奠定了理论基础. 最近, Chen [22] 将上述结论推广至无分布 (distribution-free) 的情形. 忽略证明过程, 我们将 Chen [22] 的主要结论总结为如下引理.

引理 2.1 令 $X_1,\ldots,X_n$ 为独立同分布的样本观测, 且来自一个密度函数为 $f(x)$ 的绝对连续分布 F. 假设 $f(x)$ 是连续函数且 $M=\sup_x f(x)<\infty$, 并将经验分布函数表示为

$$F_n(x)=n^{-1}\sum_{i=1}^n I(X_i\leqslant x).$$

因此, 当 $n\to\infty$ 时, 几乎必然

$$\sup_{x\in\mathbb{R}}\{F_n(x+\epsilon)-F_n(x)\}\leqslant 2M\epsilon+10n^{-1}\log n$$

对于所有 $\epsilon>0$ 一致成立.

值得注意的是, 引理 2.1 已经排除了对于每个 ϵ, 该上界可能被违反的 0 概率事件. 进一步地, 偏正态混合模型的密度函数和分布函数显然满足引理 2.1 中宽松的分布假定. 因此, 令 $\epsilon=|\sigma\log\sigma|$, 其中 $\sigma>0$ 且 σ 很小. 对引理 2.1 修改后, 我们可以获得偏正态混合模型下的结论:

引理 2.2 假设 X_i $(i=1,\ldots,n)$ 是独立同分布的随机样本, 且服从如 (2.2) 定义的密度函数为 $f(x;\Psi_0)$ 的偏正态混合分布. 除了一个不依赖于 σ 的 0 概率事件, 当 $n\to\infty$ 时, 下列结论几乎必然成立:

$$\sup_{\mu\in\mathbb{R}}\sum_{i=1}^{n} I(|X_i-\mu|\leqslant|\sigma\log\sigma|)\leqslant 4Mn|\sigma\log\sigma|+10\log n,$$

其中 $M=\sup_x f(x;\Psi_0)$.

2.3 惩罚项的选择

引理 2.1 和 2.2 实际上提供了决定惩罚项的尺度的理论基础. 为了保证所提出的惩罚极大似然估计的相合性, 假设惩罚项 $p_{1n}(\sigma)$ 和 $p_{2n}(\lambda)$ 满足以下条件:

(C1) $\forall\sigma>0$, $p_{1n}(\sigma)=o(n)$ 和 $\sup_{\sigma>0}\max\{0,p_{1n}(\sigma)\}=o(n)$ 成立.

(C2) 当 $\sigma<n^{-1}\log n$ 且 n 很大时, $p_{1n}(\sigma)\leqslant(\log n)^2\log\sigma$.

(C3) $p_{2n}(\lambda)$ 是一个连续函数且在 $\lambda=0$ 处取到最大值. 当 $|\lambda|\to\infty$ 时, $p_{2n}(\lambda)$ 趋向于负无穷. 此外, $p_{2n}(0)=0$.

(C4) $p_{1n}(\sigma)$ 和 $p_{2n}(\lambda)$ 相对于 σ 和 λ 分别可微, 且当 $n\to\infty$ 时, $p'_{1n}(\sigma)=o(n^{1/2})$ 和 $p'_{2n}(\lambda)=o(n^{1/2})$ 成立.

条件 (C1) 实际上对惩罚函数 $p_{1n}(\sigma)$ 的上界和下界作了限制, 同时条件 (C2) 要求惩罚函数 $p_{1n}(\sigma)$ 必须充分陡峭以阻止 $\sigma^2\to 0$. 条件 (C3) 是为了限制惩罚项 $p_{2n}(\lambda)$ 对似然函数的影响. 条件 (C4) 确保了惩罚极大似然估计的极限分布的存在性. 此外, 上述两个惩罚函数显然存在且不唯一. 因此, 使用者能够自由地选择合理的惩罚项, 这表明上述数理条件的约束性并不强.

本文中, 令 s_n^2 表示样本方差, 我们推荐采用下面的两个惩罚函数:

$$\begin{aligned} p_{1n}(\sigma)&=-a_n\left\{\frac{s_n^2}{\sigma^2}+\log\left(\frac{\sigma^2}{s_n^2}\right)-1\right\},\\ p_{2n}(\lambda)&=-b_n\{\lambda^2-\log(1+\lambda^2)\}, \end{aligned}\tag{2.7}$$

其中 a_n 和 b_n 分别是惩罚函数 $p_{1n}(\sigma)$ 和 $p_{2n}(\lambda)$ 的调节参数, 两者均为正数.

注意到, 我们推荐的惩罚函数很容易通过条件 (C1)–(C4) 的验证. 惩罚函数 $p_{1n}(\sigma)$ 的形式从 Bayes 理论角度代表了在参数 σ^2 上加了一个服从逆 Gamma 分布的先验分布. 同时, Chen 等 [18] 指出这个惩罚函数 $p_{1n}(\sigma)$ 具有保持估计的尺度不变性 (scale invariance) 的优势.

为处理偏度参数发散, Azzalini 和 Arellano-Valle [14] 提出了一种惩罚似然估计, 其惩罚函数为

$$p(\lambda)=-0.876\log(1+0.856\lambda^2).$$

与该惩罚项 $p(\lambda)$ 相比, 本文在 (2.7) 中所提出的 $p_{2n}(\lambda)$ 具有三方面优势: 首先, 作为一个凸函数, 我们采用的 $p_{2n}(\lambda)$ 在 0 附近非常平坦, 但在远离 0 时非常陡峭. 因此, 在 λ 的取值正常时, 该惩罚函数对于似然函数的影响非常小; 但是对于发散的 λ, $p_{2n}(\lambda)$ 非常敏感. 以 $b_n=0.01$ 为例, 图 1(a) 给出 $p_{2n}(\lambda)$ 的函数形状. 惩罚函数 $p_{2n}(\lambda)$ 代表在 $1+\lambda^2$ 上添加服从 Gamma 分布的先验密度, 而 $p(\lambda)$ 在 λ 上引入 Cauchy 分布先验, 两者代表的先验密度函数如图 1(b) 所示. 尽管两种先验估计都等于 0, Gamma 型先验密度函数在 λ 属于一个正常取值空间时具有明显更大的概率, 意味着 Gamma 型惩罚函数比 Cauchy 型惩罚函数具有更强的稳健性. 此外, Gamma 型惩罚项 $p_{2n}(\lambda)$ 不会增加 EM 型算法的计算复杂度, 而 Cauchy 型惩罚项不具有此优点, 第 4 节给出详细的解释.

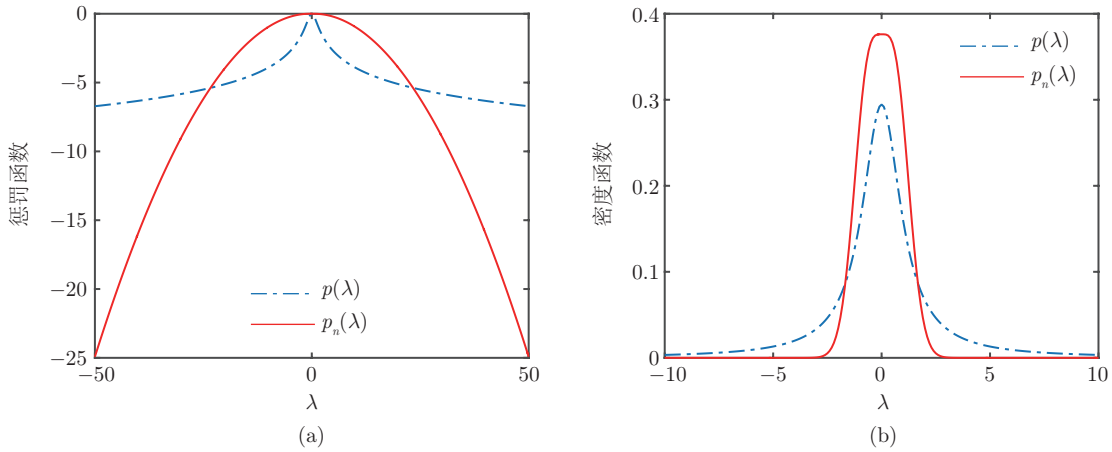

图 1　(a) $p(\lambda)$ 和 $p_{2n}(\lambda)$ 的函数形状; (b) $p(\lambda)$ 和 $p_{2n}(\lambda)$ 代表的先验密度函数

此外, a_n 和 b_n 控制了惩罚项的强度, 其合理选择应依赖于样本量 n. 通过设定 $a_n = n^{-1}$, Chen 等[18] 在正态混合模型中构建了一个具有优良表现的惩罚估计. 其动机可能是为了使惩罚项充分小, 以保持似然方法的最优性质, 文献 [18, 表 6] 中的结果充分反映了这一点. 因此, 本文也采用 $a_n = n^{-1}$. 基于相似的原因且为举例说明, 本文设定 $b_n = 0.05(\log n)^{-1}$, 第 5.2 小节解释了采用 $b_n = O\{(\log n)^{-1}\}$ 的原因.

3　惩罚极大似然估计的强相合性

3.1　$p = p_0$ 时估计的相合性

令 $M = \sup_x f(x;\Psi_0)$ 和 $K_0 = \mathrm{E}_{\Psi_0}\{\log f(X;\Psi_0)\}$, 其中 $\mathrm{E}_{\Psi_0}(\cdot)$ 表示相对于密度函数 $f(x;\Psi_0)$ 取期望. 首先选择一个充分小的正数 ϵ_0 满足以下条件:

$$4pM\epsilon_0\log^2\epsilon_0 \leqslant 1,$$
$$\log\epsilon_0 + \frac{\log^2\epsilon_0}{2} \geqslant p(2-K_0).$$

此外, 我们再选择一个足够大的正数 η_0 满足

$$\eta_0 > \max_k\{|\lambda_{0k}|\}, \quad k = 1,\ldots,p,$$

其中 λ_{0k} 是 Ψ_0 的元素. ϵ_0 和 η_0 的选择显然依赖于真实参数 Ψ_0, 而与样本量 n 无关. 由于 ϵ_0 和 η_0 显然存在, 借助这两个常数, 就能方便地定义以下参数空间:

$$\Gamma_\sigma = \{\Psi\in\Gamma : \min\{\sigma_k\} \leqslant \epsilon_0, k = 1,\ldots,p\},$$
$$\Gamma_\lambda = \{\Psi\in\Gamma : \max\{|\lambda_k|\} \geqslant \eta_0, k = 1,\ldots,p\},$$
$$\Gamma^* = \Gamma - \Gamma_\sigma\cup\Gamma_\lambda.$$

在上述参数空间中, 我们将看到惩罚函数的影响. 当参数落在空间 Γ_σ 中时, 混合分布的参数值就至少有一个类别标准差接近 0. 惩罚项 $p_{1n}(\sigma)$ 将反击这种情形, 使得满足 $\sigma\in\Gamma_\sigma$ 的惩罚极大似然估

计具有一个衰减的概率. 相似地, Γ_λ 代表至少存在一个 $|\lambda_k|$ 发散到无穷, 所提出的惩罚估计将排除在 Γ_λ 中的取值.

首先证明惩罚估计的相合性. 为使结论更清楚, 首先按升序重新排列类别标准差如下:

$$\sigma_{(1)} \leqslant \cdots \leqslant \sigma_{(p)}.$$

同时, 相应的混合比率和类别参数也分别表示为 $\pi_{(k)}$ 和 $\theta_{(k)} = (\mu_{(k)}, \sigma^2_{(k)}, \lambda_{(k)})$, 其中 $k \in \{1, \ldots, p\}$. 因此, 对于某个较小的 τ_0 和 $\tau \in \{1, \ldots, p\}$, 我们可将参数空间 Γ_σ 划分成如下子集:

$$\Gamma^\tau_\sigma = \big\{\Psi \in \Gamma_\sigma : \sigma_{(1)} \leqslant \cdots \leqslant \sigma_{(\tau)} \leqslant \tau_0 < \epsilon_0 \leqslant \sigma_{(\tau+1)} \leqslant \cdots \leqslant \sigma_{(p)}\big\}.$$

特别地, 当 $\tau = p$ 时,

$$\Gamma^p_\sigma = \big\{\Psi \in \Gamma_\sigma : \sigma_{(1)} \leqslant \cdots \leqslant \sigma_{(p)} \leqslant \tau_0 < \epsilon_0\big\}.$$

定理 3.1 假设真实的密度函数为 $f(x; \Psi_0)$, 且令惩罚似然函数 $p\ell_n(\Psi)$ 定义为 (2.6) 中的形式, 其中的惩罚项 $p_n(\Psi)$ 满足假设条件 (C1)–(C3), 那么, 当 $\Psi \in \Gamma^p_\sigma$ 且 $n \to \infty$ 时,

$$\sup_{\Gamma^p_\sigma} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty$$

几乎必然成立.

定理 3.1 的证明过程见附录 A, 此外, 当 Γ^τ_σ 且 $1 \leqslant \tau \leqslant p-1$ 时, 我们可以得到与定理 3.1 相似的结论, 相应的结论总结为如下定理.

定理 3.2 在与定理 3.1 相同的假设下, 仅改变 $\Psi \in \Gamma^\tau_\sigma$, 其中 τ 满足 $1 \leqslant \tau \leqslant p-1$, 则

$$\sup_{\Psi \in \Gamma^\tau_\sigma} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty$$

在 $n \to \infty$ 时几乎必然成立.

注意到 $\Gamma_\sigma = \bigcup_{\tau=1}^p \Gamma^\tau_\sigma$. 所以, 根据定理 3.1 和 3.2, 我们可总结出参数 Ψ 的惩罚极大似然估计以概率 1 属于 Γ^c_σ. 下面进一步证明偏度参数 λ 的惩罚极大似然估计是有界的. 考虑参数空间 $\Psi \in \Gamma^c_\sigma \cap \Gamma_\lambda$.

定理 3.3 在与定理 3.1 相同的条件下, 当 $n \to \infty$ 时, 我们可以证明下列结论几乎必然成立:

$$\sup_{\Psi \in \Gamma^c_\sigma \cap \Gamma_\lambda} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty.$$

在上述三个定理中, 我们事实上排除了所提出的惩罚极大似然估计 $\tilde{\Psi}$ 落在参数空间 $\Gamma_\sigma \cup \Gamma_\lambda = \Gamma_\sigma \cup \{\Gamma^c_\sigma \cap \Gamma_\lambda\}$ 内的可能性. 因此, 这充分表明, $\tilde{\Psi} \in \Gamma^*$ 是概率 1 成立的. 所以, 所提出的惩罚估计 $\tilde{\Psi}$ 的强相合性陈述如下.

定理 3.4 同样假设在与定理 3.1 相同的条件下, Ψ 是一个类别个数为 p_0 且满足下列不等式的混合分布:

$$p\ell_n(\Psi) - p\ell_n(\Psi_0) \geqslant c > -\infty,$$

那么, 在 $n \to \infty$ 时, 几乎必然有 $\Psi \to \Psi_0$.

我们不妨将参数空间 Γ^* 的形式重写为

$$\Gamma^* = \Big\{\Psi \in \Gamma : \min_k\{\sigma_k\} \geqslant \epsilon_0, \max_k\{|\lambda_k|\} \leqslant \eta_0\Big\}.$$

显然, $\Psi \in \Gamma^*$ 实际上等价于在类别标准差上加入一个正的下界, 同时在偏度参数的绝对值上加入一个正的上界. 进一步地, 当 $\Psi \in \Gamma^*$ 时, 因为 $p\ell_n(\tilde{\Psi}) - p\ell_n(\Psi_0) \geqslant 0$, 且有 $p_n(\Psi) = o(1)$, 因此易得

$$\ell_n(\tilde{\Psi}) - \ell_n(\Psi_0) \geqslant p_n(\tilde{\Psi}) - p_n(\Psi_0) = o(1).$$

因为参数空间 Γ^* 是正则的, 所以采用 Wald [26] 的证明技巧可以很容易得到惩罚估计 $\tilde{\Psi}$ 的强相合性, 即使似然函数添加了一个 $o(n)$ 的惩罚项. 此外, 由于 $p = p_0$, $\tilde{\Psi}$ 的所有元素因此都将几乎必然收敛到 Ψ_0 中相对应的部分.

此外, 令

$$S_n(\Psi) = \frac{\partial \ell_n(\Psi)}{\partial \Psi} \quad \text{和} \quad S_n'(\Psi) = \frac{\partial^2 \ell_n(\Psi)}{\partial \Psi \partial \Psi^T}$$

分别表示 $\ell_n(\Psi)$ 的得分向量和二阶偏导矩阵. 由于偏正态混合模型在 Ψ_0 处是正则的, 那我们就能得到正定的 Fisher 信息阵

$$nI(\Psi_0) = -\mathrm{E}\{S_n'(\Psi_0)\} = \mathrm{E}\{S_n^{\mathrm{T}}(\Psi_0) S_n(\Psi_0)\}.$$

根据经典的渐近技术和条件 (C4): $p_{1n}'(\sigma) = o(n^{1/2})$, $p_{2n}'(\lambda) = o(n^{1/2})$, 我们可以得到

$$\tilde{\Psi} - \Psi_0 = -\{S_n'(\Psi_0)\}^{-1} S_n(\Psi_0) + o_p(n^{1/2}).$$

因此, 我们所提出的惩罚极大似然估计具有渐近正态性和有效性.

定理 3.5 在与定理 3.1 相同的假设和条件 (C4) 下, 当 $n \to \infty$ 时, 依分布地,

$$\sqrt{n}(\tilde{\Psi} - \Psi_0) \to N(\mathbf{0}, I^{-1}(\Psi_0)),$$

其中 $I^{-1}(\Psi_0)$ 表示 Fisher 信息阵 $I(\Psi_0)$ 的逆矩阵.

3.2 $p > p_0$ 时估计的相合性

在实际应用中, p_0 经常是未知的, 人们可能仅了解混合模型秩的一个上界, 即 $p_0 < p < \infty$. 在这种情形下, 通过将 $\tilde{\Psi}$ 和 Ψ_0 处理成同一空间内的混合分布, Chen 等 [18] 及 Chen 和 Tan [20] 分别在一元和多元正态混合模型中证明了所提出的惩罚估计的相合性. 本文中, 为了测度混合分布 Ψ 和 Ψ_0 之间的差异, 首先定义一个如下距离:

$$D(\Psi, \Psi_0) = \int_{\Theta} |\Psi(\theta) - \Psi_0(\theta)| \exp(-|\theta|) d\theta, \tag{3.1}$$

其中 $\theta = (\mu, \sigma^2, \lambda) \in \Theta \subseteq \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$, $|\theta| = |\mu| + \sigma^2 + |\lambda|$ 且 $d\theta = d\mu d\sigma^2 d\lambda$. (3.1) 中定义的距离具有两个优良的性质: (i) $D(\Psi, \Psi_0)$ 是有界的, 因为

$$0 \leqslant D(\Psi, \Psi_0) \leqslant \int_{\Theta} \exp(-|\theta|) d\theta < \infty;$$

(ii) $D(\Psi, \Psi_0) \to 0$ 意味着 Ψ 依分布收敛于 Ψ_0, 为进一步证明相合性提供了理论基础. 因此, 我们可以得到如下定理.

定理 3.6 除了 $p_0 < p < \infty$ 以外, 在定理 3.1 的相同假设下, Ψ 是任意的具有 p 个类别且满足以下条件的混合分布:

$$p\ell_n(\Psi) - p\ell_n(\Psi_0) \geqslant c > -\infty,$$

那么当 $n \to \infty$ 时, Ψ 几乎必然趋向于 Ψ_0.

4 惩罚 EM 算法

在 SNMIX 模型的计算方面, Lin 等[1] 提出了两个拓展的 EM 算法来计算模型的极大似然估计, 分别是 ECM (expectation/conditional maximization) 算法和 ECME (expectation/conditional maximization either) 算法. 考虑到惩罚 EM 算法的渐近性质和快速的收敛速率[27], 我们提出两种惩罚的 EM 型算法来计算所提出的惩罚估计 $\tilde{\Psi}$.

考虑全数据 $(X,Z)=\{X_j,Z_j\}_{j=1}^n$, 其中隐类别指示向量 $Z_j=(Z_{1j},\ldots,Z_{pj})$ 服从一个一次实验的多项分布, 且单元概率为 $\pi_1,\ldots,\pi_p$. 我们不妨将其写作 $Z_j\sim\mathcal{M}(1;\pi_1,\ldots,\pi_p)$. 此外, $Z_1,\ldots,Z_n$ 相互独立. 基于类别指示向量, 对于每一个样本 X_j $(j=1,\ldots,n)$, 我们可以给出偏正态混合模型的一个分层随机表达式为

$$\begin{aligned}&X_j\mid\tau_j,Z_{ij}=1\sim N(\mu_i+\delta_i\tau_j,(1-\delta_i^2)\sigma_i^2),\\&\tau_j\mid Z_{ij}=1\sim TN_{[0,+\infty)}(0,\sigma_i^2),\\&Z_j\sim\mathcal{M}(1;\pi_1,\ldots,\pi_p),\end{aligned}\tag{4.1}$$

其中 $\delta_i=\lambda_i/\sqrt{1+\lambda_i^2}$, $TN_{[0,+\infty)}(0,\sigma^2)$ 表示截断正态分布 (truncated normal distribution). 此外, $\tau_1,\ldots,\tau_n$ 也是相互独立的.

根据 (4.1), 忽略可加的常数项, 全数据对数似然函数可表示为

$$\ell_c(\Psi)=\sum_{i=1}^p\sum_{j=1}^n Z_{ij}\left\{\log(\pi_i)-\log(\sigma_i^2)-\frac{1}{2}\log(1-\delta_i^2)-\frac{\tau_j^2-2\delta_i\tau_j(x_j-\mu_i)+(x_j-\mu_i)^2}{2\sigma_i^2(1-\delta_i^2)}\right\}.\tag{4.2}$$

根据 Bayes 理论, 有

$$\tau_j\mid(X_j=x_j,Z_{ij}=1)\sim TN_{[0,+\infty)}(\mu_{\tau_{ij}},\sigma_{\tau_i}^2),$$

其中 $\mu_{\tau_{ij}}=\delta_i(x_j-\mu_i)$, $\sigma_{\tau_i}=\sigma_i\sqrt{1-\delta_i^2}$. 因此, 对于目前的参数 $\Psi^{(t)}=(\pi_1^{(t)},\ldots,\pi_p^{(t)},\theta_1^{(t)},\ldots,\theta_p^{(t)})$, 其中 $\theta_k^{(t)}=(\mu_k^{(t)},\sigma_k^{2(t)},\lambda_k^{(t)})$, 令

$$\mu_{\tau_{ij}}^{(t)}=\delta_i^{(t)}(x_j-\mu_i^{(t)}),\quad \sigma_{\tau_i}^{(t)}=\sigma_i^{(t)}\sqrt{1-\delta_i^{2(t)}}\quad\text{且}\quad \delta_i^{(t)}=\frac{\lambda_i^{(t)}}{\sqrt{1+\lambda_i^{(t)2}}}.$$

ECM 算法可按如下步骤进行迭代.

E- 步: 计算条件期望

$$\begin{aligned}\alpha_{ij}^{(t)}&=\mathrm{E}(Z_{ij}\mid X_j=x_j,\Psi^{(t)})=\frac{\pi_i^{(t)}f_{SN}(x_j;\theta_i^{(t)})}{\sum_{k=1}^p\pi_k^{(t)}f_{SN}(x_j;\theta_k^{(t)})},\\\beta_{ij}^{(t)}&=\mathrm{E}(\tau_j\mid X_j=x_j,Z_{ij}=1,\Psi^{(t)})=\mu_{\tau_{ij}}^{(t)}+\sigma_{\tau_i}^{(t)}\Delta_{ij}^{(t)},\\\gamma_{ij}^{(t)}&=\mathrm{E}(\tau_j^2\mid X_j=x_j,Z_{ij}=1,\Psi^{(t)})=\mu_{\tau_{ij}}^{2(t)}+\sigma_{\tau_i}^{2(t)}+\mu_{\tau_{ij}}^{(t)}\sigma_{\tau_i}^{(t)}\Delta_{ij}^{(t)},\end{aligned}$$

其中 $\Delta_{ij}^{(t)}=\phi(\varrho_{ij}^{(t)})/\Phi(\varrho_{ij}^{(t)})$ 且 $\varrho_{ij}^{(t)}=\lambda_i^{(t)}\frac{x_j-\mu_i^{(t)}}{\sigma_i^{(t)}}$. 此外, 我们可以得到 $\mathrm{E}(Z_{ij}\tau_j\mid X_j,\Psi^{(t)})=\alpha_{ij}^{(t)}\beta_{ij}^{(t)}$ 和 $\mathrm{E}(Z_{ij}\tau_j^2\mid X_j,\Psi^{(t)})=\alpha_{ij}^{(t)}\gamma_{ij}^{(t)}$. 因此, 目标函数可表示为

$$Q(\Psi\mid\Psi^{(t)})=\mathrm{E}(\ell_c(\Psi)+p_{1n}(\Psi)\mid X,\Psi^{(t)})$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{n}\alpha_{ij}^{(t)}\Big\{\log(\pi_i) - \log(\sigma_i^2) - \frac{1}{2}\log(1-\delta_i^2) - \frac{\gamma_{ij}^{(t)} - 2\delta_i\beta_{ij}^{(t)}(x_j-\mu_i) + (x_j-\mu_i)^2}{2\sigma_i^2(1-\delta_i^2)}\Big\} + \sum_{k=1}^{p}p_{1n}(\sigma_k) + \sum_{k=1}^{p}p_{2n}(\lambda_k). \tag{4.3}$$

CM- 步: 在限制条件 $\sum_{k=1}^{p}\pi_k = 1$ 下, 相对于 Ψ 最大化目标函数 $Q(\Psi \mid \Psi^{(t)})$.

(1) 更新 $\pi_i^{(t)}$ 为 $\pi_i^{(t+1)} = n^{-1}\sum_{j=1}^{n}\alpha_{ij}^{(t)}$;

(2) 更新 $\mu_i^{(t)}$ 为

$$\mu_i^{(t+1)} = \frac{\sum_{j=1}^{n}\alpha_{ij}^{(t)}x_j - \delta_i^{(t)}\sum_{j=1}^{n}\alpha_{ij}^{(t)}\beta_{ij}^{(t)}}{\sum_{j=1}^{n}\alpha_{ij}^{(t)}}.$$

(3) 固定 $\mu_i = \mu_i^{(t+1)}$, 令

$$S_{0i}^{(t)} = \sum_{j=1}^{n}\alpha_{ij}^{(t)}\gamma_{ij}^{(t)}, \quad S_{1i}^{(t)} = \sum_{j=1}^{n}\alpha_{ij}^{(t)}\beta_{ij}^{(t)}(x_j - \mu_i^{(t+1)})$$

和

$$S_{2i}^{(t)} = \sum_{j=1}^{n}\alpha_{ij}^{(t)}(x_j - \mu_i^{(t+1)})^2,$$

根据 (2.7) 中 $p_{1n}(\sigma)$ 的定义, 获得更新的 $\sigma_i^{2(t+1)}$ 如下:

$$\sigma_i^{2(t+1)} = \frac{S_{0i}^{(t)} - 2\delta_i^{(t)}S_{1i}^{(t)} + S_{2i}^{(t)} + 2a_n(1-\delta_i^{2(t)})s_n^2}{2(1-\delta_i^{2(t)})(a_n + \sum_{j=1}^{n}\alpha_{ij}^{(t)})}.$$

(4) 固定 $\mu_i = \mu_i^{(t+1)}$ 和 $\sigma_i = \sigma_i^{(t+1)}$, 令 $\alpha_i^{(t)} = \sum_{j=1}^{n}\alpha_{ij}^{(t)}$, 考虑等价变换

$$p_{2n}(\lambda_i) = -b_n\Big\{\frac{1}{1-\delta_i^2} + \log(1-\delta_i^2) - 1\Big\},$$

所以, $\lambda_i^{(t+1)}$ 是以下方程的根:

$$\delta_i^3\sigma_i^{2(t+1)}(2b_n + \alpha_i^{(t)}) - (1+\delta_i^2)S_{1i}^{(t)} + \delta_i(S_{0i}^{(t)} + S_{2i}^{(t)} - \sigma_i^{2(t+1)}\alpha_i^{(t)}) = 0.$$

(4*) 对于文献 [14] 的惩罚函数 $p(\lambda) = -c_1\log(1+c_2\lambda^2)$, 其中 $c_1 = 0.876$, $c_2 = 0.856$. 此时, $\lambda_i^{(t+1)}$ 的值可通过求解以下方程获得:

$$\sigma_i^{2(t+1)}(\delta_i - \delta_i^3)\Big(\alpha_i^{(t)} - \frac{2c_1c_2}{1-(1-c_2)\delta_i^2}\Big) + (1+\delta_i^2)S_{1i}^{(t)} - \delta_i(S_{0i}^{(t)} + S_{2i}^{(t)}) = 0.$$

对以上的 ECM 算法稍加修正, 就可以得到拟合偏正态混合模型的 ECME 算法. 只要将其中的第 4 个 CM- 步替换为如下的 CML- 步.

CML- 步: 计算

$$(\lambda_1^{(t+1)}, \ldots, \lambda_p^{(t+1)}) = \underset{\lambda_1,\ldots,\lambda_p}{\arg\max}\Big[\sum_{i=1}^{p}p_{2n}(\lambda_i) + \sum_{j=1}^{n}\log\Big\{\sum_{i=1}^{p}\pi_i^{(t+1)}f_{SN}(x_j;\mu_i^{(t+1)},\sigma_i^{2(t+1)},\lambda_i)\Big\}\Big].$$

正如 Lin 等 [1] 指出的, 当 $p = 1$ 或者 $\lambda_1, \ldots, \lambda_p$ 全相等时, ECME 算法比 ECM 算法具有更快的收敛速率. 但是除此之外, 计算估计时选择 ECM 算法更有效率. 因为第 4 个 CM- 步中一维方程根的搜索速率远远快于 CML- 步中多维参数的优化. 此外, 为了控制算法的收敛, 当目标函数的相对变化率小于临界值 10^{-6} 时, 我们就停止该 EM 型算法.

注 4.1 与文献 [1] 中的第 4 个 CM- 步相比, 本文惩罚 ECM 算法的第 4 个 CM- 步具有相似的结构和相同的计算复杂度. 同时, 我们也注意到在 CM- 步: (4*) 中 Azzalini 和 Arellano-Valle [14] 的惩罚函数显著增加了计算的复杂程度. 此外, 值得注意的是, 代表了逆 Gamma 型先验分布的方差惩罚函数也具有保持计算效率的优点 (详见文献 [18,19]).

注 4.2 Ridolfi 和 Idier [28] 提出了构建方差惩罚函数的三大要素, 即 (i) 具有恰当的先验密度函数; (ii) 修正似然的缺陷; (iii) 保持显式的估计方程. 然而对于惩罚似然方法, 尤其是 SNMIX 模型在偏度参数方向上没有显示估计方程的情形, 第三个条件是不恰当的. 本文认为应放松第三个条件, 并将之修改为 "保持计算的效率", 即在 EM 算法中加入惩罚项不应增加计算负担.

5 模拟分析

5.1 惩罚函数的比较

第一个数值模拟是为了比较下列两个惩罚估计的表现: 我们提出的仅包含 (2.7) 中的 $p_{2n}(\lambda)$ 的惩罚极大似然估计与 Azzalini 和 Arellano-Valle [14] 提出的惩罚估计, 简称 MPLE (maximum penalized likelihood estimator). 为方便比较, 参数设置和样本量的取值与文献 [14, 图 5] 中所对应的模拟例子相同, 即 $\theta = (0, 1, 5)$ 且 $n = \{50, 100, 250, 350, 500, 1000\}$. 此外, 我们所采用的重复次数是 5,000.

图 2 显示 λ 的两个惩罚估计的偏差 (biases) 和均方根误差 (root mean squared errors, RMSEs). 其中第一行是固定偏度参数值 $\lambda = 5$ 但样本量 $n = \{50, 100, 250, 350, 500, 1000\}$ 的模拟结果; 第二行是固定样本量 $n = 100$ 但 $\lambda = 1, \ldots, 10$ 的结果. 因此, 我们可考察当样本量上升时, 偏差和和均方根误差下降的趋势. 图 2 的第一行显示了这一结果, 同时也反映了两个惩罚估计的相合性. 图 2 的第二行揭示了在小样本下 ($n = 100$), 偏度参数 λ 变大时, 这两个估计的表现. 相对于文献 [14] 中的 MPLE, 我们的惩罚估计 (PMLE) 一致表现更优. 同时, λ 取值在 6 附近时, 我们的 PMLE 的偏差可以达到非常小.

此外, 我们采用双对数化尺度来简化对于曲线的理解. 尤其对于图 2(a) 中, 所提出的 MPLE 的偏差以近似 $n^{-3/2}$ 阶的速率下降. 这可能是由于随着样本量 n 增加, (2.7) 中定义的惩罚项 $p_{2n}(\lambda)$ 以 $\log(n)^{-1}$ 阶的速率下降, 因此, PMLE 的偏差减少的比 MPLE 更为缓慢. 所以, 总的来说, 在小样本和中等规模样本下, 本文提出的 PMLE 明显地优于文献 [14] 中的 MPLE. 出现这种现象的原因可能是, 我们的方法在理论上存在一定的优势 (详见第 2.3 小节).

5.2 $p = p_0$ 时的模拟分析

本小节通过数值模拟来验证所提出的惩罚极大似然估计的相合性. 我们依然采用重复次数为 5,000, 并考虑样本量 $n = \{100, 200\}$ 下的两个类别个数为 2 的 SNMIX 模型. 表 1 给出所考虑模型的参数设定.

对于每个模型, 借助 ECM 算法和惩罚 ECM 算法计算似然函数和惩罚似然函数的局部最大值, 我们可以获得其相应的估计量. 为处理算法的初始化问题, Chen 等 [18] 采用了真实的混合分布作为初始值; Basso 等 [2] 提出了采用结合 K- 均值算法和矩估计的方法来定义初始值. 在本节的模拟中, 我们同时采用两种方案并考察其表现. 此外, 为了克服混合模型中存在的标签交换 (label switching) 效应 [7], 我们在偏正态混合模型中采用 Celeux 等 [29] 提出的关于位置参数的方法.

模型 I 模型 I 的密度函数体现为双峰并且具有很好的分离特征. 表 2 给出 5,000 次重复中 $\hat{\sigma}^2$ 和

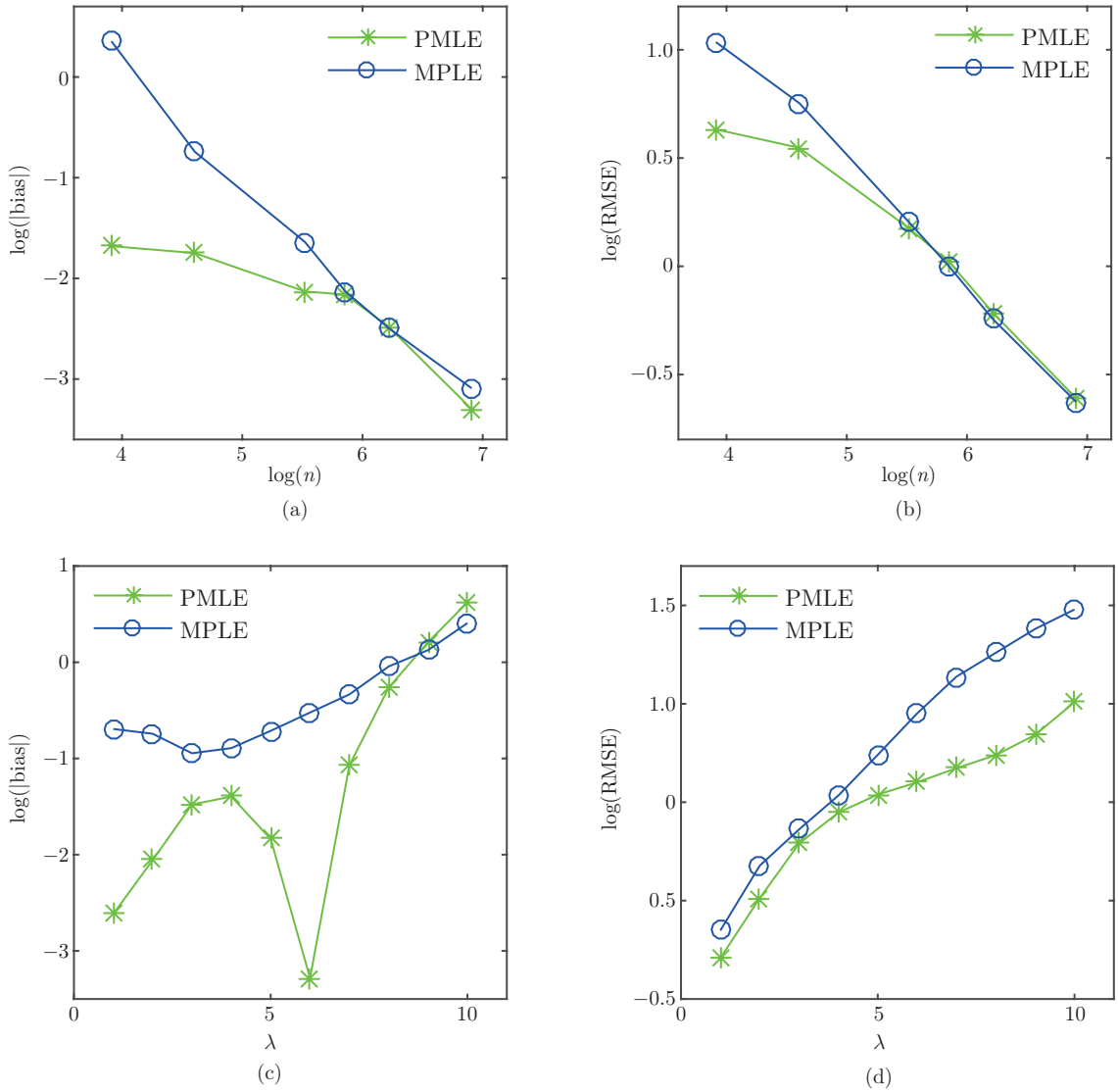

图 2　两个惩罚估计的模拟分析

表 1　模拟模型的参数设定

模型	参数设定 $(SN(\mu,\sigma^2,\lambda))$
I	$0.5SN(-2,1,2)+0.5SN(2,2,1)$
II	$0.5SN(-1,2,1)+0.5SN(1.5,2,-1)$

表 2　模型 I 中极大似然估计的结果

参数	真值初始化		K- 均值初始化	
	$n=100$	$n=200$	$n=100$	$n=200$
n $(\lvert\hat{\lambda}\rvert>100)$	205	4	214	3

$|\hat{\lambda}|$ 有缺陷的频率. 本文将 $\hat{\sigma}^2<10^{-10}$ 作为 σ^2 的估计值存在缺陷的标志, 同时将 $|\hat{\lambda}|>100$ 作为其发散的标志. 注意到惩罚估计中不存在具有缺陷的估计值.

表 2 的结果表明, 在模型 I 中, σ^2 的极大似然估计并没有缩减到 0. 然而, 尽管密度函数分离得很好, 模型的极大似然估计依然受到 λ 的发散性的困扰, 这在两种初始化方案下都得到了体现.

表 3 进一步计算了 MLE 和 PMLE 的偏差和均方根误差. 从表 3 可以看出, PMLE 的偏差和均方根误差随着样本量 n 的上升显著地下降, 这体现了其相合性. 同时值得注意的是, 在估计 λ_1 和 λ_2 的表现上, PMLE 远远优于 MLE, 因为 MLE 小样本情形下非常不可信, 而 PMLE 改进了这一缺点. 总的来说, 除了参数 μ_1, PMLE 的表现都显著地优于 MLE. 同时, 可能是由于分离良好的核密度函数, 所有其他参数的 MLE 和 PMLE 表现都不错.

偏度参数 MLE 的不可靠表现再次验证了 DiCiccio 和 Monti [24] 的理论预期, 即 λ 的估计准确度在样本量较大的情形下才能保证. 一般而言, 当样本量 $n = 1{,}000$ 时, 混合分布的极大似然估计应该是可信的. 在模型 I 中, 我们在 $n = 1{,}000$ 时进行了一个独立的模拟研究, 其结果总结在图 3 中. 此时, 不仅估计出的混合分布 $\hat{\Psi}$, 并且 $\hat{\Psi}$ 的所有元素都收敛到了 Ψ_0.

此外, 为了处理偏正态分布中 λ 的估计发散性, DiCiccio 和 Monti [24] 提出了一种较易实现的方法. 他们将 λ 的估计定义为, 使得 $H_0 : \lambda = \breve{\lambda}$ 在 5% 的显著性水平下不能被拒绝, 且取值最小的 $\breve{\lambda}$. 其所使用的检验统计量是轮廓似然比检验, 并依分布收敛于 χ^2_1.

令 $\hat{\Lambda}$ 表示 $\Lambda = (\lambda_1, \ldots, \lambda_p)$ 的极大似然估计, 且真值为 $\Lambda_0 = (\lambda_{10}, \ldots, \lambda_{p0})$. 由于在模型 I 中不存在 $\sigma^2 \to 0$ 的情形, 因此可以将 DiCiccio 和 Monti [24] 的方法拓展到偏正态混合模型. 记 $\breve{\Lambda}$ 是 $\hat{\Lambda}$ 的一个修正估计 (modified estimator, ME) 且 $\breve{\Lambda} = (\breve{\lambda}_1, \ldots, \breve{\lambda}_p)$. 此时, $\breve{\Lambda}$ 可以通过以下方式获得: 令复合原假设 $H_0 : \Lambda = \breve{\Lambda}$ 在 5% 的显著性水平下不被轮廓似然比检验拒绝, 且使得 $\|\breve{\Lambda} - \hat{\Lambda}\|_1$ 取到最大的 $\breve{\Lambda}$. 其中的轮廓似然比统计量的极限分布为 χ^2_ν, 其自由度定义为

$$\nu = \sum_{k=1}^{p} I(|\lambda_k| \geqslant 30).$$

表 3 模型 I 的偏差和均方根误差 (括号内)

	参数	$n = 100$		$n = 200$	
		MLE	PMLE	MLE	PMLE
真值初始化	$\hat{\mu}_1$	0.027 (0.26)	0.036 (0.24)	0.015 (0.19)	0.020 (0.18)
	$\hat{\mu}_2$	0.071 (0.46)	0.070 (0.43)	0.051 (0.35)	0.055 (0.35)
	$\hat{\sigma}_1^2$	0.162 (0.74)	0.081 (0.59)	0.072 (0.45)	0.048 (0.41)
	$\hat{\sigma}_2^2$	0.055 (0.84)	0.024 (0.77)	0.029 (0.62)	0.017 (0.59)
	$\hat{\lambda}_1$	7.208 (40.9)	0.528 (1.82)	0.571 (4.36)	0.299 (1.25)
	$\hat{\lambda}_2$	5.065 (32.0)	0.491 (1.62)	0.358 (1.94)	0.233 (1.05)
	$\hat{\pi}_1$	0.008 (0.03)	0.005 (0.03)	0.004 (0.02)	0.003 (0.02)
K- 均值初始化	$\hat{\mu}_1$	0.088 (0.37)	0.097 (0.36)	0.043 (0.23)	0.049 (0.22)
	$\hat{\mu}_2$	0.228 (0.79)	0.226 (0.77)	0.117 (0.50)	0.120 (0.49)
	$\hat{\sigma}_1^2$	0.103 (0.69)	0.030 (0.56)	0.030 (0.45)	0.006 (0.41)
	$\hat{\sigma}_2^2$	0.293 (1.07)	0.229 (0.93)	0.073 (0.68)	0.059 (0.65)
	$\hat{\lambda}_1$	6.385 (39.0)	0.304 (1.95)	0.434 (4.03)	0.178 (1.29)
	$\hat{\lambda}_2$	3.522 (33.0)	0.095 (2.13)	0.197 (1.99)	0.091 (1.25)
	$\hat{\pi}_1$	0.003 (0.03)	0.001 (0.03)	0.002 (0.02)	0.001 (0.02)

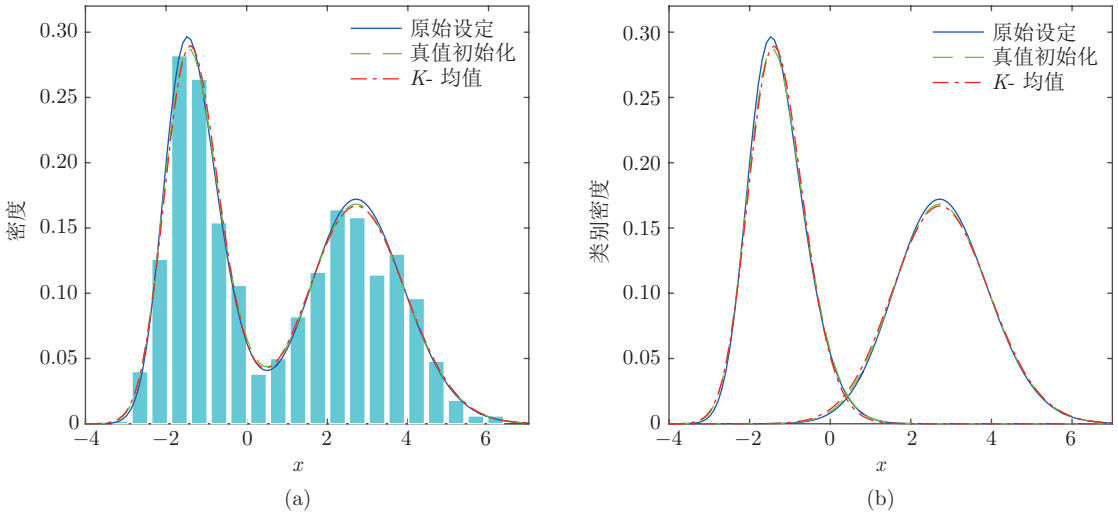

图 3　(a) 直方图及真实和估计的密度; (b) 真实和估计的类别密度

修正估计和极大似然估计的唯一区别在于 λ_1 和 λ_2 的估计表现. 表 4 表明在两种初始化方案下, 修正估计都能够有效消除偏度参数的发散估计. 而对比表 3 可进一步看出, 所提出的惩罚极大似然估计比修正估计的效果更好.

注 5.1　然而, 直接将 DiCiccio 和 Monti[24] 的方法拓展至偏正态混合模型缺乏严格的理论基础. 这是因为当 $\breve{\sigma}^2 = \hat{\sigma}^2 \to 0$ 时, 原假设 $H_0 : \Lambda = \breve{\Lambda}$ 下, 参数位于在参数空间的边界上. 所以, 在考虑的混合模型中, 基本的正则条件并不满足, 似然比检验统计量的渐近 χ^2 理论不再成立. 尽管在密度函数分离很好的情形下, 由于并不存在 $\hat{\sigma}^2 \to 0$, 他们的方法看起来是可使用的, 但是, 他们的方法实际上在偏正态混合模型中是无效的, 这也体现了我们所提出的惩罚估计的另一个优势.

在下文中, 考虑一个更复杂情形的模拟, 此时, 混合模型中的两个类别密度相互间非常接近. 此时存在一个有意思的现象, 尽管混合分布中的参数的估计效果并不好, 但总的密度函数可以获得精确的估计.

模型 II　模型 II 具有方差齐性, 且其密度函数明显是单峰的且分不开的. 该模型下的模拟结果列于表 5 和 6 中.

表 4　模型 I 中修正估计的偏差和均方根误差 (括号中)

参数	真值初始化		K- 均值初始化	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\breve{\lambda}_1$	1.247 (3.87)	0.443 (1.74)	0.972 (3.91)	0.317 (1.76)
$\breve{\lambda}_2$	0.967 (3.19)	0.336 (1.52)	0.446 (3.75)	0.183 (1.64)

表 5　模型 II 的极大似然估计结果

参数	真值初始化		K- 均值初始化	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
n $(\hat{\sigma}^2 < 10^{-10})$	62	12	3	0
n $(\lvert\hat{\lambda}\rvert > 100)$	533	91	638	75

表 6 模型 II 的偏差和均方根误差 (括号中)

	参数	$n=100$		$n=200$	
		MLE	PMLE	MLE	PMLE
真值初始化	$\hat{\mu}_1$	−0.123 (0.74)	−0.094 (0.72)	−0.081 (0.57)	−0.063 (0.54)
	$\hat{\mu}_2$	0.154 (0.76)	0.125 (0.73)	0.089 (0.58)	0.075 (0.56)
	$\hat{\sigma}_1^2$	−0.901 (5.88)	−0.579 (1.09)	−0.279 (2.28)	−0.280 (0.72)
	$\hat{\sigma}_2^2$	−0.793 (5.15)	−0.595 (1.12)	−0.319 (2.63)	−0.288 (0.74)
	$\hat{\lambda}_1$	12.14 (52.1)	0.303 (1.32)	2.707 (19.8)	0.286 (1.19)
	$\hat{\lambda}_2$	−13.02 (54.4)	−0.350 (1.36)	−2.497 (18.5)	−0.269 (1.13)
	$\hat{\pi}_1$	0.003 (0.23)	0.005 (0.25)	0.003 (0.19)	0.004 (0.19)
K- 均值初始化	$\hat{\mu}_1$	0.871 (1.18)	0.846 (1.20)	0.947 (1.13)	0.935 (1.14)
	$\hat{\mu}_2$	−0.808 (1.14)	−0.784 (1.17)	−0.951 (1.13)	−0.940 (1.13)
	$\hat{\sigma}_1^2$	−0.719 (1.52)	−0.745 (1.13)	−0.562 (0.87)	−0.582 (0.89)
	$\hat{\sigma}_2^2$	−0.920 (10.0)	−0.796 (1.19)	−0.543 (0.82)	−0.564 (0.83)
	$\hat{\lambda}_1$	−15.27 (50.5)	−2.752 (3.57)	−4.648 (15.7)	−2.414 (3.29)
	$\hat{\lambda}_2$	14.64 (50.1)	2.699 (3.52)	4.447 (16.8)	2.411 (3.29)
	$\hat{\pi}_1$	0.010 (0.24)	0.009 (0.25)	0.001 (0.20)	0.001 (0.20)

从表 5 可以看出, (i) 即使真实分布用于作为迭代算法的初始值, 极大似然估计依然受到参数 σ^2 和 λ 变差的困扰; (ii) 惩罚似然方法同时解决了 σ^2 和 λ 变差的问题, 此时所有的参数估计值都控制的很好; (iii) 经典的聚类方法 K- 均值算法具有一定的阻止 $\hat{\Psi}$ 中变差的类别方差出现的效果, 由于我们发现相比于真值初始化, K- 均值初始化降低了方差变差的比率.

表 5 也显示了 $\hat{\lambda}$ 变差的频率要显著地高于 $\hat{\sigma}^2$ 变差的频率. 因此, 偏正态混合模型中形状参数的发散问题在实际中必须引起更多的重视. 这种现象也能部分解释我们在 (2.7) 中使用一个随着 n 上升, 以速率 $(\log n)^{-1}$ 缓慢减弱的惩罚项 $p_{2n}(\lambda)$, 而它显然比 $p_{1n}(\sigma)$ 的衰减速率 n^{-1} 慢得多.

表 6 报告了模型 II 下 MLE 和 PMLE 的偏差和均方根误差. 为了彰显变差的 $\hat{\sigma}^2$ 与 σ_0^2 之间的差异, 使得比较方案更合理, 我们计算 $\log(\hat{\sigma}_i^2)$ 的偏差和均方根误差, 这个指标与 Chen 等 [18] 定义的相对指标成正比. 当 Ψ_0 作为初始值时, PMLE 的偏差和均方根误差随着 n 的增长快速下降, 并且 PMLE 相对于 MLE 在 $\hat{\lambda}$ 和 $\hat{\sigma}^2$ 上具有明显的优势.

在 K- 均值初始化的情形下, 尽管 $\hat{\sigma}^2 \to 0$ 被极大地阻止了, 但是 Ψ 的估计并不准确, 两种估计方法甚至都失去了在 $\hat{\mu}$ 上的相合性. 我们采用一个单独的模拟来研究这种现象, 此实验的模拟数据从模型 II 产生且样本量 $n = 1{,}000$. 表 7 表明 MLE 和 PMLE 在相同的初始化方式下是几乎相等的, 但是不同的初始化下结果差异非常大. 同时, 真值初始化下获得的估计得到的 $p\ell_n(\hat{\Psi})$ 明显小于 K- 均值初始化下的结果. 也就是说, 当 Ψ_0 作为初始值时, EM- 型算法收敛到了一个局部最大值, 而 K- 均值初始化下的估计才是一个全局的最大解, 这应该是导致表 6 中模拟结果表现很差的原因.

图 4 形象地总结了表 7 的结果. 我们可以看出, 尽管基于 K- 均值方法的拟合密度和真值下的估计非常接近, 但是两者的子密度差异非常大. 换句话说, 这种现象并未对有限混合模型的可识别性提出挑战 (参见文献 [8, 26]), 但是显示了偏正态混合模型中当两个类靠得很近时存在一个所谓的 “过度灵活” (over-flexibility) 的缺点.

在基于 K- 均值方法的估计中, 尽管表 6 显示了其不良表现, 但另一个有意思的发现是, 它在

表 7　模型 II 的参数估计 ($n = 1{,}000$)

	方法	μ_1	μ_2	σ_1^2	σ_2^2	λ_1	λ_2	π_1	$p\ell_n(\hat{\Psi})$
真值初始化	MLE	−1.018	1.440	1.940	1.873	1.018	−1.016	0.504	−1617
	PMLE	−1.018	1.439	1.940	1.873	1.018	−1.015	0.504	−1617
K- 均值初始化	MLE	−0.037	0.273	1.136	1.399	−0.846	1.814	0.525	−1616
	PMLE	−0.038	0.273	1.136	1.398	−0.844	1.804	0.525	−1616

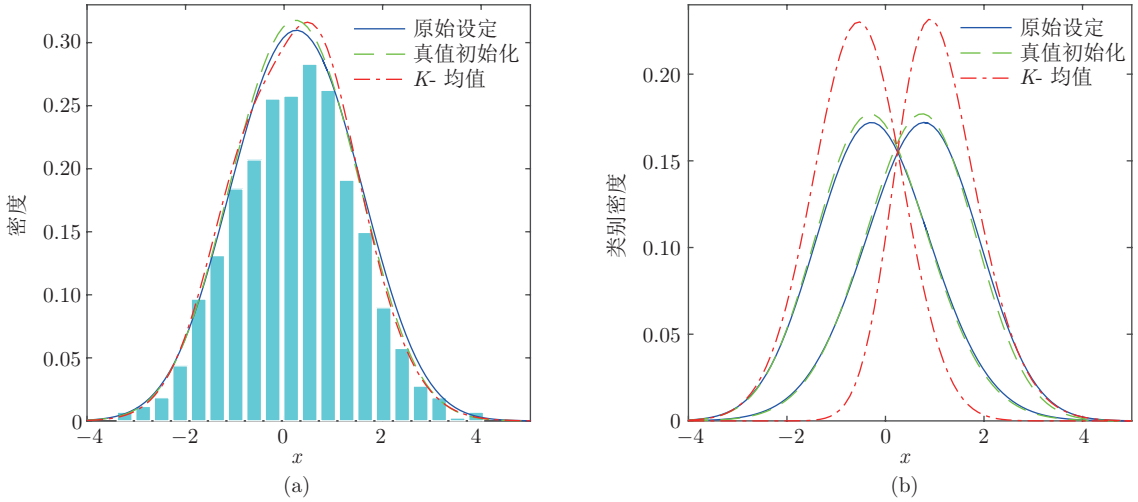

图 4　(a) 直方图及真实和估计的密度; (b) 真实和估计的类别密度

$\hat{\sigma}^2 \to 0$ 的发生次数远低于真值初始化的估计结果, 这在表 5 中列举的偏正态混合模型的估计结果得到了体现. 为了进一步分析这种现象可能的原因, 考虑文献 [18, 例 2], 从模型 $0.5N(0,1)+0.5N(1.5,3)$ 中产生. 同时, 我们采用两种初始化策略来计算估计, 并同时拟合了两个类的 Gauss 混合模型和偏正态混合模型. 重复次数依然取 5,000.

表 8 的主要结果总结如下: 对于 Gauss 混合模型, 当样本量 $n = 100$ 时, 相比于真值初始化下的估计结果, 基于 K- 均值方法的估计存在更多 $\hat{\sigma}^2$ 变差的结果. 而在偏正态混合模型中, 基于 K- 均值的估计比基于真值的估计更有效地抑制了 $\hat{\sigma}^2 \to 0$ 的出现. 但此时基于 K- 均值的估计比基于真值的估计遭受更频繁的 $\hat{\lambda}$ 的发散问题的困扰, 表明基于 K- 均值的估计总的估计效果更差.

综上所述, 模型 II 中抑制 $\hat{\sigma}^2 \to 0$ 的效果因此来自于偏正态混合的建模形式和 K- 均值初始化方法的组合. 对于这种现象一个推断的解释是, 偏正态混合模型中 K- 均值初始值不会落进奇异值附近的吸引域 (attraction domain), 这个吸引域的存在性已被 Biernacki 和 Chrétien [30] 证明.

表 8　模型 $0.5N(0,1)+0.5N(1.5,3)$ 的估计结果

	参数	真值初始化		K- 均值初始化	
		$n = 100$	$n = 200$	$n = 100$	$n = 200$
GMIX	n ($\hat{\sigma}^2 < 10^{-10}$)	30	3	40	3
SNMIX	n ($\hat{\sigma}^2 < 10^{-10}$)	22	2	3	2
	n ($\lvert\hat{\lambda}\rvert > 100$)	0	0	588	116

5.3 $p > p_0$ 时的模拟分析

为方便起见, 在 $p_0 < p < \infty$ 时, 模拟数据依然从模型 I 中产生, 此时考虑样本量 $n = \{100, 200, 500\}$. 由于 $p_0 = 2$, 对于每一个模拟数据, 我们都计算当 $p = \{2, 3, 4, 5\}$ 时的极大似然估计和惩罚极大似然估计. 此时, 模拟规模为 1,000 次.

由于 $p > p_0$, 我们无法得到 $\hat{\Psi}$ 的每个元素都收敛到 Ψ_0 中相应的部分的结论. 为了处理此问题, Chen 等[18] 及 Chen 和 Tan[20] 采用真值 Ψ_0 附近的 10 个参数值作为迭代算法的初始值. 本节采用轻微扰动 Ψ_0 中的 μ_{0j} 的方式来获取 10 个初始值, 详细的扰动方案如下:

$$\mu_i = \mu_{0j} + N(0, 0.1^2), \quad \pi_i = \frac{\pi_{0j}}{\omega_j}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, p_0,$$

其中 ω_j 是 μ_i 来自于 μ_{0j} 的总类别数目, 同时 $\sum_{j=1}^{p_0} \omega_j = p$. 给定这 10 个初始值后, 最大的目标函数对应的估计被确定为最终的估计值.

在这种情形下, 研究 (3.1) 中定义的距离 $D(\hat{\Psi}, \Psi_0)$ 具有非常重要的意义. 但是, 直接用 $D(\hat{\Psi}, \Psi_0)$ 来测度 $\hat{\Psi}$ 与 Ψ_0 之间的差异并不是一个明智的选择. 为改善这种局面, 我们采用一个修正的距离

$$D^*(\hat{\Psi}, \Psi_0) = \int_{\Theta^*} |\hat{\Psi}(\theta) - \Psi_0(\theta)| d\theta,$$

其中

$$\theta = \left(\mu, \frac{\log(\sigma^2)}{5}, \frac{\log(\lambda)}{2}\right),$$

且 $\Theta^* = [-5, 10] \times [-15, 1] \times [-10, 5]$. 实际上, 两种估计方法的所有参数估计值都很好地限定在区域 Θ^* 内.

表 9 显示了极大似然方法中参数变差的频率. 很显然, 随着样本量 n 的增加, σ^2 和 λ 变差的频率在不断下降, 同时该频率随着推断的类别个数 p 的增加而增加. 此外, 我们也可以观察到 λ 变差的频率要显著高于 σ^2, 这个现象与此前表 2 和 5 的结论一致. 我们也注意到, 在所有的情形下, 惩罚估计都没有出现变差的结果.

表 10 报告了两种估计方案下距离 $D^*(\hat{\Psi}, \Psi_0)$ 的平均值. 在每一种情形中, $D^*(\hat{\Psi}, \Psi_0)$ 的平均值都随着样本量 n 的增加而下降, 但是这个平均距离下降速率非常慢. 这种现象可以用 Chen[31] 的结论来解释, 他证明了在 $p > p_0$ 时, 估得分布的最优收敛速率最高只能达到 $n^{-1/4}$. 此外, 在 $n = 100$ 时, 我们可以发现 PMLE 的平均距离比 MLE 的值显著要小, 且增长的更慢, 这反映了 $p > p_0$ 时 PMLE 的优势. 然而, 两种估计方法在 $D^*(\hat{\Psi}, \Psi_0)$ 的平均值上的差异随着 n 的增加逐渐消失.

表 9 模型 I 极大似然估计中参数变差的频率

$p_0 = 2$	$n = 100$		$n = 200$		$n = 500$	
	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\sigma}^2$	$\hat{\lambda}$
$p = 2$	0	40	0	2	0	0
$p = 3$	12	492	3	134	0	1
$p = 4$	79	901	12	290	0	12
$p = 5$	166	1196	30	437	0	25

表 10 模型 I 两种估计方法 $D^*(\Psi, \Psi_0)$ 的平均

$p_0 = 2$	MLE			PMLE		
	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
$p = 2$	9.46	6.84	4.35	8.17	6.53	4.32
$p = 3$	15.56	10.83	6.59	10.37	9.00	6.23
$p = 4$	19.04	13.39	7.95	12.29	10.83	7.48
$p = 5$	22.28	15.44	9.08	13.82	12.19	8.48

6 实例分析

6.1 Iris 数据

本节采用 Iris 数据来说明方法的应用. 如第 1 节中所描述的, 该数据包含了 3 种 Iris 亚种的 150 个观测样本, 每个亚种包含 50 个样本. 观测变量总共是 4 个, 分别是萼片长度 (sepal length)、萼片宽度 (sepal width)、花瓣长度 (petal length) 和花瓣宽度 (petal width).

正如第 1 节中提到的, 我们考虑萼片长度和花瓣宽度这两个边际特征, 采用类别个数为 3 的 GMIX 模型和 SNMIX 模型来拟合这两个变量. 这两个模型的极大似然估计值列于表 11 中. 计算上, 我们利用 K- 均值初始化方法, 20 个初始值下计算的目标函数表现最优者被选为最终的估计值. 同时, 目标函数的相对变化达到 10^{-6} 作为停止 ECM 算法的收敛准则.

从表 11 中可以看出, 在两个变量中, 由于存在估计的类别方差等于 0, GMIX 的 MLE 都遇到了无界的似然函数的问题. 当 SNMIX 模型用于拟合萼片长度时, 似然函数的缺陷依然存在. 同时由于 Azzalini 和 Arellano-Valle [14] 建议将 $|\lambda| > 100$ 作为偏度参数发散的标志, 因此, 我们得到 λ_1 的 MLE 是发散的. 图 5(a) 中数据直方图的左半部分揭示了数据的偏斜特征 [15], 这导致 $\hat{\lambda}_1$ 的缺陷. 从表 11 的花瓣宽度的结果中, 我们也可以看出 SNMIX 模型中 λ_2 的 MLE 也发散, 但其似然函数不存在发散的问题. 这一现象与表 8 的结果一致, 反映了 SNMIX 模型和 K- 均值初始化在阻止 $\hat{\sigma}^2 \to 0$ 上的优势.

上述结果事实上证实了, 即使在类别个数已知的 Iris 数据中, 极大似然方法是无效的. 此外, 由于无界的似然函数和形状参数的发散, SNMIX 模型的 MLE 并未被正确定义.

惩罚极大似然估计的结果列于表 12 中, 其计算的初始化策略和收敛准则与表 11 一致. 相对于表 11 中极大似然估计不能令人满意的表现, 惩罚极大似然方法的所有估计值都得到了很好的限制. 这个结果表明了惩罚极大似然估计在克服 σ^2 和 λ 的估计缺陷上的有效性.

图 5 显示了两个变量的直方图和 SNMIX 模型两种估计下的拟合密度. 从图形可视化的角度, 惩

表 11 Iris 数据的极大似然估计结果

	模型	π_1	π_2	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2	λ_1	λ_2	λ_3
萼片长度	GMIX	0.25	0.72	4.93	6.10	7.70	0.08	0.43	0			
	SNMIX	0.21	0.76	5.11	6.12	7.70	0.11	0.45	$< 10^{-29}$	-136	-0.14	4.21
花瓣宽度	GMIX	0.03	0.28	0.10	0.25	1.65	0.00	0.01	0.20			
	SNMIX	0.33	0.30	0.13	1.50	1.98	0.02	0.08	0.08	3.62	-192	0.08

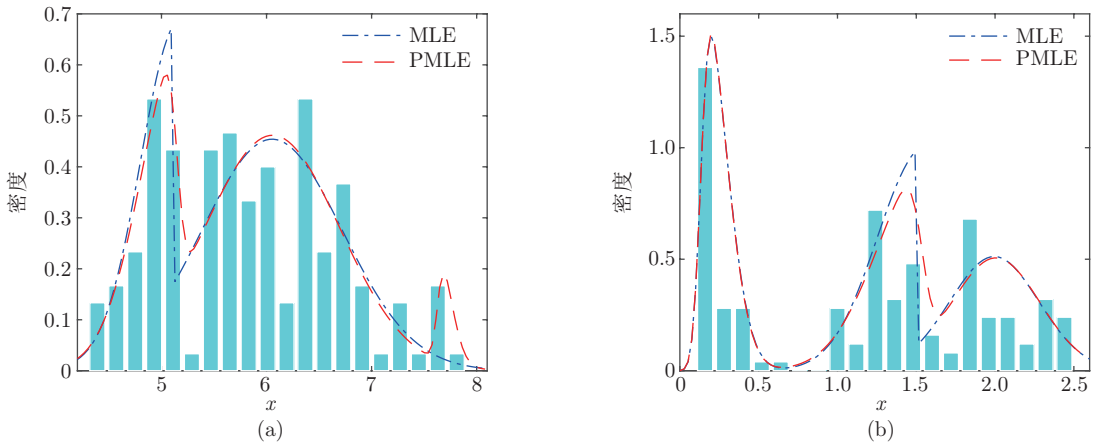

图 5　(a) 萼片长度的直方图和 MLE 及 PMLE 的估计密度; (b) 花瓣宽度的直方图和 MLE 及 PMLE 的估计密度

表 12　Iris 数据的惩罚极大似然估计结果

	模型	π_1	π_2	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2	λ_1	λ_2	λ_3	$p\ell_n(\hat{\Psi})$
萼片长度	GMIX	0.27	0.70	4.93	6.10	7.71	0.09	0.38	0.01				−174.4
	SNMIX	0.22	0.75	5.15	6.33	7.63	0.13	0.50	0.02	−5.85	−0.58	2.84	−171.9
花瓣宽度	GMIX	0.33	0.39	0.24	1.37	2.08	0.01	0.06	0.06				−101.3
	SNMIX	0.33	0.32	0.13	1.54	1.96	0.02	0.09	0.08	3.52	−5.07	0.22	−95.0

罚似然方法比极大似然方法提供了一种更为合理的估计. 注意到图 5(b) 中, 数据的直方图并未表现出极大似然估计所表示的极端模式, 而惩罚似然方法取得了显著更优的结果.

此外, 惩罚似然方法为进一步统计推断提供了可能, 例如, 通过检验 $H_0 : \lambda_1 = \lambda_2 = \lambda_3 = 0$ 来验证正态性假设. 显然在原假设下, 惩罚似然比检验统计量 (penalized likelihood ratio test statistic, PLRT) 依分布收敛于 $\chi^2(3)$. 尤其在花瓣宽度中, 我们得到 PLRT $= 2(101.3 - 95.0) = 12.6$, 同时相应的 p 值为 0.006. 因此, 我们有强有力的证据反对正态假设.

在图 5(a) 和 5(b) 中, 我们也注意到在 (2.4) 的条件不成立时, SNMIX 模型中 λ 的发散估计也会出现. 在实践中, 我们发现基于 EM 算法的极大似然估计经常存在这种问题, 这一现象也揭示了应用惩罚估计的必要性.

6.2　Faithful 数据

Faithful 数据收集于黄石国家公园的老实泉, 本节研究所提出的惩罚估计在该数据中的准确性. Faithful 数据获得了大量研究人员的关注 (参见文献 [32, 33]). 该数据有 272 个样本组成, 以时间为单位观测了两个变量: 喷发长度和喷发持续时间. Lin 等 [1] 和 Prates 等 [34] 分别采用一元和二元的两类别 SNMIX 模型拟合了该数据, 并都发现 SNMIX 模型较相应的 GMIX 模型具有更优的表现.

本节采用两类别 SNMIX 模型拟合喷发长度, 并将结果列于表 13 中. 惩罚估计的所有参数值与极大似然估计值基本完全一致. PMLE 和 MLE 的相似性也可通过图 6 中的密度函数曲线和累积分布函数曲线得出, 两者都密不可分.

基于表 13 和图 6 的结果, 可看出我们所提出的惩罚似然方法能提供一个非常精确的估计, 即使在小样本的情形下也具有充分好的表现.

表 13　Faithful 数据的参数估计结果

方法	μ_1	μ_2	σ_1^2	σ_2^2	λ_1	λ_2	π_1	$p\ell_n(\hat{\Psi})$
MLE	1.73	4.80	0.15	0.46	5.82	−3.40	0.35	−258
PMLE	1.73	4.79	0.14	0.46	5.56	−3.36	0.35	−258

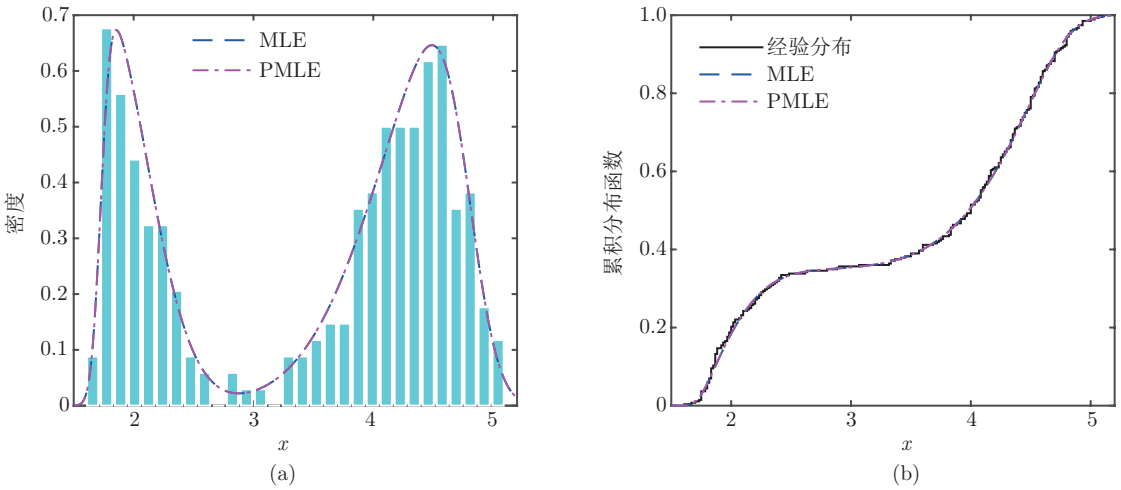

图 6　(a) Faithful 数据的直方图和 MLE 及 PMLE 的估计密度; (b) Faithful 数据的经验累积分布函数和 MLE 及 PMLE 估计的累积分布函数

7　结论

本文针 SNMIX 模型的 MLE 中 σ^2 变差和 $|\lambda|$ 发散提出了一种惩罚极大似然估计. 在推断的类别个数 p 大于或等于 p_0 时, 我们都给出了惩罚估计的强相合性的严格证明. 这种方法能有效地重获估计的相合性和有效性, 并具有在参数空间上不添加额外限制的优点. 这种方法可以广泛地应用于其他类型的有限混合模型, 如多元混合模型和混合回归模型, 相应的情形更为复杂并值得进一步研究.

此外, 由于模型的复杂性, 估计过程相比一般的参数模型更加复杂. 一个自然的问题是, 我们是否应该使用一个假设模型结构未知的非参数估计. 理论上, 本文讨论的方法较任意的非参数方法都更加有效. 此外, 我们进行的一些模拟表明, 在有限样本下, 核密度估计的均方误差要大得多. 因此, 惩罚估计在偏正态混合模型中是更合理的.

致谢　感谢 Adelchi Azzalini 提出的许多优秀建议.

参考文献

1　Lin T I, Lee J C, Yen S Y. Finite mixture modelling using the skew normal distribution. Statist Sinica, 2007, 17: 909–927
2　Basso R M, Lachos V H, Cabral C R B, et al. Robust mixture modeling based on scale mixtures of skew-normal distributions. Comput Statist Data Anal, 2010, 54: 2926–2941
3　Frühwirth-Schnatter S, Pyne S. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics, 2010, 11: 317–336
4　Liu M, Lin T. A skew-normal mixture regression model. Educ Psychol Meas, 2014, 74: 139–162
5　Zeller C B, Cabral C R B, Lachos V H. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST, 2016, 25: 375–396

6 Lindsay B G. Mixture Models: Theory, Geometry, and Applications. Bethesda: Institute of Mathematical Statistics; Alexandria: American Statistical Association, 1995
7 MacLahlan G, Peel D. Finite Mixture Models. New York: John Wiley & Sons, 2000
8 Kiefer J, Wolfowitz J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Statist, 1956, 27: 887–906
9 Day N E. Estimating the components of a mixture of normal distributions. Biometrika, 1969, 56: 463–474
10 Fisher R A. The use of multiple measurements in taxonomic problems. Ann Eugenics, 1936, 7: 179–188
11 Basford K E, McLachlan G J. Estimation of allocation rates in a cluster analysis context. J Amer Statist Assoc, 1985, 80: 286–293
12 Ray S, Lindsay B G. The topography of multivariate normal mixtures. Ann Statist, 2005, 33: 2042–2065
13 Boldea O, Magnus J R. Maximum likelihood estimation of the multivariate normal mixture model. J Amer Statist Assoc, 2009, 104: 1539–1549
14 Azzalini A, Arellano-Valle R B. Maximum penalized likelihood estimation for skew-normal and skew-t distributions. J Statist Plann Inference, 2013, 143: 419–433
15 Azzalini A, Capitanio A. Statistical applications of the multivariate skew normal distribution. J R Stat Soc Ser B Stat Methodol, 1999, 61: 579–602
16 Redner R. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann Statist, 1981, 9: 225–228
17 Hathaway R J. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Statist, 1985, 13: 795–800
18 Chen J, Tan X, Zhang R. Inference for normal mixtures in mean and variance. Statist Sinica, 2008, 18: 443–465
19 Ciuperca G, Ridolfi A, Idier J. Penalized maximum likelihood estimator for normal mixtures. Scand J Statist, 2003, 30: 45–59
20 Chen J, Tan X. Inference for multivariate normal mixtures. J Multivariate Anal, 2009, 100: 1367–1383
21 Alexandrovich G. A note on the article 'Inference for multivariate normal mixtures' by J. Chen and X. Tan. J Multivariate Anal, 2014, 129: 245–248
22 Chen J. Consistency of the MLE under mixture models. Statist Sci, 2017, 32: 47–63
23 Chen J, Li S, Tan X. Consistency of the penalized MLE for two-parameter Gamma mixture models. Sci China Math, 2016, 59: 2301–2318
24 DiCiccio T J, Monti A C. Inferential aspects of the skew exponential power distribution. J Amer Statist Assoc, 2004, 99: 439–450
25 Azzalini A. A class of distributions which includes the normal ones. Scand J Statist, 1985, 12: 171–178
26 Wald A. Note on the consistency of the maximum likelihood estimate. Ann Math Statist, 1949, 20: 595–601
27 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
28 Ridolfi A, Idier J. Penalized maximum likelihood estimation for univariate normal mixture distributions. AIP Conf Proc, 2001, 568: 229
29 Celeux G, Chauveau D, Diebolt J. Stochastic versions of the em algorithm: An experimental study in the mixture case. J Stat Comput Simul, 1996, 55: 287–314
30 Biernacki C, Chrétien S. Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM. Statist Probab Lett, 2003, 61: 373–382
31 Chen J. Optimal rate of convergence for finite mixture models. Ann Statist, 1995, 23: 221–233
32 Silverman B W. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986
33 Azzalini A, Bowman A W. A look at some data on the old faithful geyser. J R Stat Soc Ser C Appl Stat, 1990, 39: 357–365
34 Prates M O, Cabral C R B, Lachos V H. Mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. J Statist Softw, 2013, 54: 1–20

附录 A 定理的证明

定理 3.1 的证明 定义关于样本观测的指标集

$$A_{(k)} = \{i : |x_i - \mu_{(k)}| < |\sigma_{(k)} \log \sigma_{(k)}|\},$$

其中 $k=1,\ldots,p$. 对于任意的集合 S, 令 $n(S)$ 表示 S 中元素的个数. 同时定义

$$\ell_n(\Psi;S)=\sum_{i\in S}\log f(x_i;\Psi).$$

对于 $\Psi\in\Gamma_\sigma^p$, 且 ϵ_0 足够小, 当样本 $i\in A_{(k)}$ 时, 混合模型的密度满足

$$f(x_i;\Psi)\leqslant\frac{1}{\sigma_{(k)}}.$$

显然, $n(\bigcap_{t=1}^{k-1}A_{(t)}^c\cap A_{(k)})\leqslant n(A_{(k)})$. 回顾引理 2.2 中 $n(A_{(k)})$ 的上界, 那么几乎必然可以得到

$$\begin{aligned}\ell_n\Big(\Psi;\bigcap_{t=1}^{k-1}A_{(t)}^c\cap A_{(k)}\Big)&\leqslant -n(A_{(k)})\log\sigma_{(k)}\\&\leqslant 4Mn\sigma_{(k)}\log^2\sigma_{(k)}-10\log\sigma_{(k)}\log n.\end{aligned}\tag{A.1}$$

引入满足条件 (C1) 和 (C2) 的惩罚函数 $p_{1n}(\sigma_{(k)})$, (A.1) 可进一步拓展为

$$\begin{aligned}&\ell_n\Big(\Psi;\bigcap_{t=1}^{k-1}A_{(t)}^c\cap A_{(k)}\Big)+p_{1n}(\sigma_{(k)})\\&\leqslant 4Mn\sigma_{(k)}\log^2\sigma_{(k)}-(10\log n-\log^2 n)\log\sigma_{(k)}\\&\leqslant 4Mn\sigma_{(k)}\log^2\sigma_{(k)}\\&\leqslant 4Mn\epsilon_0\log^2\epsilon_0.\end{aligned}\tag{A.2}$$

此外, 对于任意的 $i\in\bigcap_{t=1}^p A_{(t)}^c$, 由于 $|x_i-\mu_{(k)}|>|\sigma_{(k)}\log\sigma_{(k)}|$, 故容易得到

$$\begin{aligned}\log f(x_i;\Psi)&\leqslant\log\Big\{\sum_{k=1}^p\frac{2\pi_k}{\sigma_k}\phi\Big(\frac{x_i-\mu_k}{\sigma_k}\Big)\Big\}\\&\leqslant\log\Big\{\sum_{k=1}^p\frac{2\pi_k}{\sigma_k}\phi(-\log\sigma_k)\Big\}\\&\leqslant-\log\epsilon_0-\frac{\log^2\epsilon_0}{2}<0.\end{aligned}$$

根据 $4pM\epsilon_0\log^2\epsilon_0\leqslant 1$, 若 ϵ_0 足够小, $-4pM\epsilon_0\log\epsilon_0\leqslant\frac{p-1}{p}$ 几乎必然成立, 这进一步表明

$$n\Big(\bigcap_{t=1}^p A_{(t)}^c\Big)\geqslant n-\sum_{t=1}^p n(A_{(t)})\geqslant\frac{n}{p}.$$

因此, $\bigcap_{t=1}^p A_{(t)}^c$ 中所有样本观测对数似然贡献的总和具有如下上界:

$$\ell_n\Big(\Psi;\bigcap_{t=1}^p A_{(t)}^c\Big)\leqslant-\frac{n}{p}\Big\{\log\epsilon_0+\frac{(\log\epsilon_0)^2}{2}\Big\}.\tag{A.3}$$

因此, 当 $\Psi\in\Gamma_\sigma^p$ 且所选择的 ϵ_0 充分小时, 根据 (A.2) 和 (A.3) 的结论和条件 (C3), 惩罚对数似然具有如下上界:

$$p\ell_n(\Psi)=\sum_{k=1}^p\Big\{\ell_n\Big(\Psi;\bigcap_{t=1}^{k-1}A_{(t)}^c\cap A_{(k)}\Big)+p_{1n}(\sigma_{(k)})\Big\}+\ell_n\Big(\Psi;\bigcap_{t=1}^p A_{(t)}^c\Big)+o(n)$$

$$\begin{aligned}&\leqslant 4pMn\epsilon_0\log^2\epsilon_0-\frac{n}{p}\left\{\log\epsilon_0+\frac{(\log\epsilon_0)^2}{2}\right\}+o(n)\\&\leqslant n+n(K_0-2)+o(n)\\&=n(K_0-1)+o(n).\end{aligned}$$

根据强大数定理, 我们知道 $\frac{1}{n}p\ell_n(\Psi_0)\xrightarrow{\text{a.s.}}K_0$. 所以, 当 $n\to\infty$ 时,

$$\sup_{\Gamma_\sigma^p}p\ell_n(\Psi)-p\ell_n(\Psi_0)\leqslant -n+o(n)\to-\infty$$

几乎必然成立. □

定理 3.2 的证明 $\bar{\Gamma}_\sigma^\tau$ 表示紧致的 Γ_σ^τ, 其允许 $\sigma_{(1)}=\cdots=\sigma_{(\tau)}=0$. 对于 $\Psi\in\bar{\Gamma}_\sigma^\tau$, 定义如下连续函数:

$$g_\tau(x;\Psi)=\sum_{k=1}^{\tau}\frac{\pi_{(k)}}{\sqrt{2}}\phi\left(\frac{x-\mu_{(k)}}{\sqrt{2}\epsilon_0}\right)+\sum_{k=\tau+1}^{p}\pi_{(k)}f(x;\theta_{(k)}),$$

其中 $f(x;\theta_{(k)})$ 是第 k 个类的密度函数. 由于 $\sigma_{(p)}\geqslant\cdots\geqslant\sigma_{(\tau+1)}\geqslant\epsilon_0$, 所以, $g_\tau(x;\Psi)$ 在参数空间 $\bar{\Gamma}_\sigma^\tau$ 上是有界的. 因此, $\forall\,\Psi\in\bar{\Gamma}_\sigma^\tau$, 我们可以得到

$$\log\mathrm{E}_{\Psi_0}\left\{\frac{g_\tau(X;\Psi)}{f(X;\Psi_0)}\right\}=-\Delta_\tau(\epsilon_0)<0.$$

同样明显的是, $\Delta_\tau(\epsilon_0)$ 是一个递减函数, 且 $\lim_{\epsilon_0\to0}\Delta_\tau(\epsilon_0)\in(0,\infty)$. 所以, 对于足够小的 ϵ_0, 下列不等式成立:

$$8\tau M\epsilon_0\log^2\epsilon_0<\Delta_\tau(\epsilon_0).$$

在参数空间 $\bar{\Gamma}_\sigma^\tau$ 上定义如下函数:

$$l_n^\tau(\Psi)=\sum_{i=1}^{n}\log\{g_\tau(x_i;\Psi)\}.$$

根据强大数定律和 Jensen 不等式的上界可知, 几乎必然有

$$\sup_{\Psi\in\bar{\Gamma}_\sigma^\tau}n^{-1}\{l_n^\tau(\Psi)-\ell_n(\Psi_0)\}\to\mathrm{E}_{\Psi_0}\log\left\{\frac{g_\tau(X;\Psi)}{f(X;\Psi_0)}\right\}\leqslant-\Delta_\tau(\epsilon_0).\tag{A.4}$$

对于 $\Psi\in\Gamma_\sigma^\tau$ 和 $\tau\in\{1,\ldots,p-1\}$, 根据此前 $A_{(k)}$ ($k\in\{1,\ldots,\tau\}$) 的定义, 对于任意的 $i\in A_{(k)}$, 混合分布的密度

$$f(x_i;\Psi)\leqslant\frac{1}{\sigma_{(k)}}g_\tau(x_i;\Psi).$$

同时对于其他的样本观测, 由于 $|x_i-\mu_{(k)}|\geqslant|\sigma_{(k)}\log\sigma_{(k)}|$, 如果 $\sigma_{(k)}$ 足够小且满足

$$\sigma_{(k)}^{-1}=\exp\{-\log\sigma_{(k)}\}<\exp\left\{\frac{1}{4}\log^2\sigma_{(k)}\right\},$$

那么, 当 $\sigma_{(k)}\leqslant\epsilon_0$ 时, 下列不等式成立:

$$f(x;\theta_{(k)})\leqslant\frac{2}{\sigma_{(k)}}\phi\left(\frac{x-\mu_{(k)}}{\sigma_{(k)}}\right)$$

$$\leqslant \frac{1}{\sqrt{2}}\phi\left(\frac{x-\mu_{(k)}}{2\sigma_{(k)}}\right)$$
$$\leqslant \frac{1}{\sqrt{2}}\phi\left(\frac{x-\mu_{(k)}}{2\epsilon_0}\right).$$

等价地, 我们可以得到 $f(x_i;\Psi)\leqslant g_\tau(x_i;\Psi)$.

总结上述结论可知, 样本观测 x_i 的对数似然贡献具有如下上界:

$$\log f(x_i;\Psi)\leqslant\begin{cases}-\log\sigma_{(k)}+\log g_\tau(x_i;\Psi), & i\in A_{(k)},\\ \log g_\tau(x_i;\Psi), & \text{其他}.\end{cases}$$

进一步地, 对数似然函数的上界可表示为

$$\ell_n(\Psi)\leqslant l_n^\tau(\Psi)-\sum_{k=1}^{\tau}n(A_{(k)})\log\sigma_{(k)}.$$

根据 (A.2) 和 (A.4) 的结论, 对于我们所选择的 ϵ_0, 我们可以得到如下结论:

$$\begin{aligned}\sup_{\Gamma_\sigma^\tau}p\ell_n(\Psi)-p\ell_n(\Psi_0)&\leqslant\sup_{\Gamma_\sigma^\tau}\{l_n^\tau(\Psi)-\ell_n(\Psi_0)\}+\sup_{\Gamma_\sigma^\tau}\sum_{k=1}^{\tau}\{-n(A_{(k)})\log\sigma_{(k)}+p_{1n}(\sigma_{(k)})\}+o(n)\\&\leqslant-n\Delta_\tau(\epsilon_0)+4\tau Mn\epsilon_0\log^2\epsilon_0+o(n)\\&\leqslant-\frac{\Delta_\tau(\epsilon_0)}{2}n+o(n).\end{aligned}$$

注意到 $\Delta_\tau(\epsilon_0)>0$, 因此, 对于任意的 $\tau\in\{1,\ldots,p-1\}$, 当 $n\to\infty$ 时, 几乎必然有

$$\sup_{\Gamma_\sigma^\tau}p\ell_n(\Psi)-p\ell_n(\Psi_0)\to-\infty.$$

所以, 定理得证. □

定理 3.3 的证明 当 $\Psi\in\Gamma_\sigma^c\cap\Gamma_\lambda$ 时, 模型标准差显然具有一个正的下界, 且发散的偏度参数值并不会导致密度趋向于无穷, 因此, 混合密度 $f(x;\Psi)$ 在参数空间 $\Gamma_\sigma^c\cap\Gamma_\lambda$ 上是有界的.

根据 Jensen 不等式可知, 对于任意的 $\Psi\in\Gamma_\sigma^c\cap\Gamma_\lambda$, 下式成立:

$$\mathrm{E}_{\Psi_0}\log\left\{\frac{f(X;\Psi)}{f(X;\Psi_0)}\right\}<0.$$

同时, 我们也可以选择足够大的 η_0, 满足 $\Psi_0\notin\Gamma_\sigma^c\cap\Gamma_\lambda$. 那么, 与 Wald [26] 相似, 我们可以得到在 $n\to\infty$ 时, 几乎必然存在

$$\sup_{\Gamma_\sigma^c\cap\Gamma_\lambda}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(x_i;\Psi)}{f(x_i;\Psi_0)}\right)\right\}\to-\Delta(\eta_0)<0. \tag{A.5}$$

注意到 $\Delta(\eta_0)$ 是一个大于 0 的正数, 同时是一个关于 η_0 的递增函数. 利用 (A.5) 中获得的上界和条件 (C1)–(C3), 我们可以进一步得到

$$\sup_{\Gamma_\sigma^c\cap\Gamma_\lambda}p\ell_n(\Psi)-p\ell_n(\Psi_0)=\sup_{\Gamma_\sigma^c\cap\Gamma_\lambda}\sum_{i=1}^{n}\log\left(\frac{f(x_i;\Psi)}{f(x_i;\Psi_0)}\right)+\sup_{\Gamma_\sigma^c\cap\Gamma_\lambda}p_n(\Psi)-p_n(\Psi_0)$$

$$\leqslant -\frac{\Delta(\eta_0)}{2}n + o(n).$$

所以, 在 $n \to \infty$ 时,

$$\sup_{\Gamma_\sigma^c \cap \Gamma_\lambda} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty$$

是几乎必然成立的. □

定理 3.5 的证明 根据上述 $p = p_0$ 情形下的证明步骤, 在 $p > p_0$ 时, 我们建立一个简要的证明过程.

借助之前在 (3.1) 中定义的距离, 并假设常数 $\kappa > 0$, 定义一个如下的新参数空间:

$$\Omega(\kappa) = \{\Psi : \Psi \in \Gamma, D(\Psi, \Psi_0) \geqslant \kappa\}.$$

显而易见, 当 $\kappa > 0$ 时, $\Psi_0 \notin \Omega(\kappa)$.

对于任意的 $\Psi \in \Gamma_\sigma^p \cap \Omega(\kappa)$, 很容易证明, 若将 $\Psi \in \Gamma_\sigma^p$ 替换为 $\Psi \in \Gamma_\sigma^p \cap \Omega(\kappa)$, 定理 3.1 中的推导过程仍然是适用的. 因此, 我们可以快速地得到如下结论: 当 $n \to \infty$ 时,

$$\sup_{\Gamma_\sigma^p \cap \Omega(\kappa)} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty.$$

同时, 惩罚估计 $\tilde{\Psi}$ 以概率 1 不属于参数空间 $\Gamma_\sigma^p \cap \Omega(\kappa)$.

此外, 由于 $\Psi_0 \notin \Omega(\kappa)$, 对于 $1 \leqslant \tau \leqslant (p-1)$, 当 $\Psi \in \Gamma_\sigma^\tau \cap \Omega(\kappa)$ 和 $\Psi \in \Gamma_\sigma^c \cap \Gamma_\lambda \cap \Omega(\kappa)$ 时, 相应的不等式:

$$\mathrm{E}_{\Psi_0} \log\left\{\frac{g_\tau(X;\Psi)}{f(X;\Psi_0)}\right\} < 0,$$
$$\mathrm{E}_{\Psi_0} \log\left\{\frac{f(X;\Psi)}{f(X;\Psi_0)}\right\} < 0$$

依然分别成立.

因此, 对于恰当选择的 ϵ_0、η_0 和定义良好的 $g_\tau(x;\Psi)$, (A.4) 和 (A.5) 的结论可推广至如下情形:

$$\sup_{\Gamma_\sigma^\tau \cap \Omega(\kappa)} n^{-1}\{l_n^\tau(\Psi) - \ell_n(\Psi_0)\} \leqslant -\Delta_\tau(\epsilon_0) < 0,$$
$$\sup_{\Gamma_\sigma^c \cap \Gamma_\lambda \cap \Omega(\kappa)} \left\{\frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(X_i;\Psi)}{f(X_i;\Psi_0)}\right)\right\} \to -\Delta(\eta_0) < 0.$$

基于上述的两个结论, 对于 $\tau \in \{1, \ldots, (p-1)\}$, 当 $n \to \infty$ 时, 可以类似地得到

$$\sup_{\Gamma_\sigma^\tau \cap \Omega(\kappa)} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty,$$
$$\sup_{\Gamma_\sigma^c \cap \Gamma_\lambda \cap \Omega(\kappa)} p\ell_n(\Psi) - p\ell_n(\Psi_0) \to -\infty.$$

总结上述结论可知, 惩罚极大似然估计 $\tilde{\Psi}$ 必然以概率 1 落在参数空间 $\Gamma^* \cup \Omega^c(\kappa)$ 内. 考虑到 κ 取值的任意性, $\tilde{\Psi} \in \Omega^c(\kappa)$ 等价于 $D(\tilde{\Psi}, \Psi_0) \to 0$. 同时, 根据文献 [8], $\tilde{\Psi} \in \Gamma^*$ 也意味着 $D(\tilde{\Psi}, \Psi_0) \to 0$. 所以, 当 $p > p_0$ 时, 惩罚估计的强相合性得证. □

Penalized maximum likelihood estimation for skew normal mixtures

Libin Jin, Wangli Xu, Liping Zhu & Lixing Zhu

Abstract Skew normal mixture models provide a more flexible framework than the popular normal mixtures for modelling heterogeneous data with asymmetric behaviors. Due to the unboundedness of likelihood function and the divergency of shape parameters, the maximum likelihood estimators of the parameters of interest are often not well defined, leading to dissatisfactory inferential process. We put forward a proposal to deal with these issues simultaneously in the context of penalizing likelihood function. The resulting penalized maximum likelihood estimator is proved to be strongly consistent when the putative order of mixture is equal to or greater than the true one. We also provide penalized EM-type algorithms to compute penalized estimators. Finite sample performances are examined by simulations and the comparison to the existing methods. Two real examples including the famous Iris dataset are analysed for illustration.

Keywords **likelihood degeneracy, boundary estimator, skew normal mixtures, penalized MLE, strong consistency**

MSC(2010) **62F12**

doi: 10.1360/SCM-2018-0256