嵌套删失数据期望最大化的高斯混合聚类算法

余海燕[1,2] 陈京京[1] 邱 航[2,3] 王 永[1] 王若凡[4]

摘 要 针对聚类问题中的非随机性缺失数据, 本文基于高斯混合聚类模型, 分析了删失型数据期望最大化算法的有效性, 并揭示了删失数据似然函数对模型算法的作用机制. 从赤池弘次信息准则、信息散度等指标, 比较了所提出方法与标准的期望最大化算法的优劣性. 通过删失数据划分及指示变量, 推导了聚类模型参数后验概率及似然函数, 调整了参数截尾正态函数的一阶和二阶估计量. 并根据估计算法的有效性理论, 通过关于得分向量期望的方程得出算法估计的最优参数. 对于同一删失数据集, 所提出的聚类算法对数据聚类中心估计更精准. 实验结果证实了所提出算法在高斯混合聚类的性能上优于标准的随机性缺失数据期望最大化算法.

关键词 高斯混合聚类, 删失数据, 期望最大化算法, 截尾正态函数, 二阶估计量

引用格式 余海燕, 陈京京, 邱航, 王永, 王若凡. 嵌套删失数据期望最大化的高斯混合聚类算法. 自动化学报, 2021, **47**(6): 1302−1314

DOI 10.16383/j.aas.c190081

Adapted Expectation Maximization Algorithm for Gaussian Mixture Clustering With Censored Data

YU Hai-Yan[1,2] CHEN Jing-Jing[1] QIU Hang[2,3] WANG Yong[1] WANG Ruo-Fan[4]

Abstract To provide a solution for clustering with data of missing not at random, this paper provided the efficiency analysis on the adapted expectation-maximization (EM) algorithm for Gaussian mixture clustering model with censored data. We also revealed the impact mechanism of the likelihood function of censored data on the clustering model and its estimation algorithm. With Akaike´s information criterion and Kullback-Leibler divergence, the performance of the proposed algorithm was compared with the standard EM algorithm. Based on data partition and the indicating variables of the censored data set, the paper proposed derived the posterior and likelihood function of the parameters, and adjusted its first and second moments of the truncated normal functions. According to the principles of efficient influence function, the optimal parameters of the algorithm are obtained by the equation of the expectation of the score vector. For the censored data, the proposed clustering algorithm is more accurate in estimating its centroids. The experimental results demonstrated that the proposed algorithm in Gaussian mixture clustering outperformed the standard EM algorithm, which was designed for the data of missing at random.

Key words Gaussian mixture clustering, censored data, expectation-maximization, truncated normal function, second order moment

Citation Yu Hai-Yan, Chen Jing-Jing, Qiu Hang, Wang Yong, Wang Ruo-Fan. Adapted expectation maximization algorithm for Gaussian mixture clustering with censored data. *Acta Automatica Sinica*, 2021, **47**(6): 1302−1314

高斯混合聚类[1−2] 作为统计机器学习、模式识别和阵列数据分析等的重要模型, 广泛用于健康医疗[3−4]、故障诊断[5−6] 等领域. 然而, 常因诸如截断的数据、传感器故障或传输错误等造成数据不完整问题[1], 引起推断偏差并使得聚类精度下降. 例如在医疗决策智能支持中[7−8], 需要依据患者的各项生理指标信息进行智能推理[9−10], 然而由于记录数据删失或截断等导致数据不完整, 从而给数据分析带来困难.

收稿日期 2019-02-11 录用日期 2019-07-30
Manuscript received February 11, 2019; accepted July 30, 2019
国家自然科学基金 (71601026, 61601331, 71571105), 重庆市产业类重大主题专项 (cstc2017zdcy-zdzxX0013), 四川省重点研发项目 (2018SZ0114, 2019YFS0271), 天津市自然科学基金青年项目 (18JCQNJC04700) 资助
Supported by National Natural Science Foundation of China (71601026, 61601331, 71571105), Chongqing Science and Technology Commission (cstc2017zdcy-zdzxX0013), Key Research and Development Program of Sichuan Province (2018SZ0114, 2019YFS0271), and Tianjin Natural Science Foundation Youth Project (18JCQNJC04700)
本文责任编委 胡清华
Recommended by Associate Editor HU Qing-Hua
1. 重庆邮电大学电子商务与现代物流重庆市重点实验室 重庆 404615 2. 电子科技大学计算机科学与工程学院 成都 611731 3. 电子科技大学大数据研究中心 成都 611731 4. 天津职业技术师范大学信息技术工程学院 天津 300222
1. Chongqing Key Laboratory of Electronic Commerce and Modern Logistics, Chongqing University of Posts and Telecomms., Chongqing 404615 2. School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu 611731 3. Big Data Research Center, University of Electronic Science and Technology, Chengdu 611731 4. School of Information Technology Engineering, Tianjin University of Technology and Education, Tianjin 300222

在恶性淋巴瘤等疾病诊断[11] 中, 流式细胞仪记录的数据因测量信号强度范围有限而使得数据记录在一个固定范围内 (如 0 到 1 023 之间), 如果测量值超出这一范围, 则该值将替换为最接近的值, 小于 0 的值将被删失记为 0, 大于 1 023 的值将被删失记为 1023. 类似的删失数据还包括保险费理赔计算中, 因一定数量免赔额的存在使得记录成为删失数据等. 这类删失数据处理不当会影响分析结果的可靠性, 甚至使得聚类模型参数推断出现较大偏差. 又因这类数据的分布参数的精确估计, 为处理变量或治疗方案对观察结果的因果效应分析[12] 提供基础, 甚至影响到后续的决策方案选择. 高斯聚类算法因能够提供分布参数的估计, 故而删失数据的参数估计已成为高斯混合聚类的一个重要热点问题.

删失数据的处理方法常基于缺失数据的处理机理. 因数据缺失机制不同, 处理方法也不尽相同. 数据缺失可以分为随机缺失 (Missing at random, MAR) 和非随机缺失 (Missing not at random, MNAR) 两大类[12]. 大多数传统的缺失数据处理方法主要集中于使用样本抽样推断、贝叶斯推断和似然法推断[13]. 其中贝叶斯推断和似然法在实际数据中的应用更为普遍. 当评估项目的长期性能数据随机缺失且观测数据也随机缺失时, 使用样本抽样估计数据集分布参数可以忽略缺失机制. 当数据属于随机缺失且缺失机制参数不同于数据集分布参数时, 使用贝叶斯推断和似然法也可以忽略缺失机制. 文献 [12] 对非随机缺失问题的探索, 还包括不可忽略性无响应问题、不可忽略性缺失性问题, 甚至被称为有信息缺失的问题等. 文献 [14] 认为存在解决非随机缺失的方法, 但是通常难以检验, 为此提出了惩罚验证标准, 通过惩罚未知参数过多的模型来防止模型过拟合. 删失数据作为一种非随机性缺失数据[15−16], 因其缺失机制 (如删失) 的特殊性而不能直接使用一般的非随机缺失方法直接计算[11].

删失数据常包括右删失和区间删失等类型. 对于右删失数据, 文献 [17] 基于一类广义概率测度的误差一致性, 提供了适用于删失数据的分类支持向量机并应用于删失数据平均值、中位数、分位数的估计以及分类问题. 针对区间删失数据, 文献 [15] 提出一种贝叶斯非参数化方法进行概率拟合. 文献 [18] 基于左截断右删失数据构造了分位差的经验估计, 并提出了分位数差的核光滑估计. 针对删失混合数据, 文献 [19] 提出了一个加权最小二乘估计的一般族, 并证明了现有的一致非参数方法属于这个族, 识别其估计量并分析其渐近性质. 而在高斯混合聚类模型算法中, 一般假设观测值的特征向量对聚类有相同的权重[20]. 然而文献 [1] 认为高斯混合聚类模型的每一个特征向量的权重并不一样, 提出竞争性惩罚期望最大化算法. 该算法将特征选择模型和高斯混合聚类模型结合在一起, 使用马尔科夫毯滤波器消除多余的特征项, 找到最小的相关特征子集, 同时确定高斯混合模型的混合成分个数. 文献 [21] 提出了一种基于高斯混合聚类和模型平均的算法. 对于缺失值, 该方法将每一组成成分得出的估计值作为线性组合的概率估计权重, 最终结果是混合成分的估计值的平均值. 文献 [2] 讨论高斯混合聚类分析的过拟合问题. 该文献改变了以往认为不相关变量必须通过线性回归方程依赖整个相关变量的做法, 认为相关变量并不一定要解释所有的不相关变量. 该模型可以有效地提高聚类算法的性能且变量选择的实现基于一个向后逐步算法. 标准期望最大化 (Expectation-maximization, EM) 算法作为高斯混合模型中常用的缺失数据处理方法[22], 更适用于处理随机缺失数据. 本文在标准 EM 的高斯混合聚类算法 (EMGM) 基础上, 提出了嵌套删失数据期望最大化的高斯混合聚类算法 (cenEMGM).

本文主要解决非随机缺失下的删失数据因利用率不高而导致聚类准确度不高的问题. 本文的主要贡献是: 利用高斯混合模型聚类算法独有的特性, 在标准 EM 算法的基础上提出改进算法 cenEMGM, 并揭示了删失率对模型算法的作用机制. 将删失数据和高斯混合模型聚类算法结合, 更加准确地处理删失数据. 通过调整删失数据的分布函数, 使得删失数据最大期望算法不断更新均值、协方差和混合系数的估计值, 从而使得聚类簇中心不断接近真实的簇中心. cenEMGM 算法在标准 EMGM 算法的基础上进行改进, 该方法更加灵活, 对删失和未删失数据采取不同的处理方式. 删失数据 EM 算法和高斯混合聚类相结合, 使得该方法比原方法聚类效果更好, 准确性更高. 后续章节结构如下: 第 1 节引入高斯混合聚类模型. 第 2 节论述删失型缺失数据的相关概念. 第 3 节构建高斯混合聚类的参数估计算法, 包括标准 EMGM 算法和 cenEMGM 两种算法, 以及两个模型校验准则. 第 4 节使用数值实验验证算法. 第 5 节得出结论.

1　高斯混合聚类模型

对 d 维数据空间 $\mathbf{R}^d$ 中, 随机变量 $\boldsymbol{y}$ 的观察值为一个由 n 个样本构成的数据集, $D=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$, 其中 $\boldsymbol{y}_i$ 为其第 i 个样本. 并将第 j 维数据记为 $\boldsymbol{y}^{(j)}$. 假设样本生成过程由包含 K 个成分的高斯混合分布确定. 第 k 个成分 f_k 的参数为 $\Theta_k=(\pi_k,\mu_k,\Sigma_k)$; 其中, π_k 为其混合系数, μ_k 为均值, Σ_k 为方差. 全部参数 $\Theta=\{\Theta_1,\Theta_2,\cdots,\Theta_K\}$. $\boldsymbol{y}^{(j)}$ 为其第 j 维观测值. 对于 $\boldsymbol{y}$, 定义高斯混合分布[20] 如下:

$$p(\boldsymbol{y})=\sum_{k=1}^{K}\pi_k\cdot p(\boldsymbol{y}\,|\mu_k,\Sigma_k) \tag{1}$$

其中, K 为混合成分数量, 且每个混合成分对应一个高斯分布 $\mathrm{N}(\mu_k,\Sigma_k)$, 相应的 "混合系数" $\pi_k>0$, $\sum_{k=1}^{K}\pi_k=1$.

样本生成过程中, 记 $\pi=\{\pi_1,\cdots,\pi_K\}$, 首先根据 π 定义的先验分布选择高斯混合成分, 且选择第 k 个混合成分的概率为 π_k; 然后, 根据被选择的混合成分的概率密度函数进行采样, 从而生成相应的样本.

在高斯混合聚类模型中, 类似地存在 K 个簇, $\boldsymbol{C}=\{C_1,C_2,\cdots,C_K\}$. 将 $\boldsymbol{y}_i$ 是否被划分到簇 C_k 中的随机变量记为 $z_i^{(k)}$, 簇指示变量 $z_i^{(k)}\in\{0,1\}$. 当 $\boldsymbol{y}_i$ 被划分到簇 C_k 时, $z_i^{(k)}=1$, 意味着 $\boldsymbol{y}_i$ 由 f_k 生成; 否则 $z_i^{(k)}=0$. 对于 N个样本总体, $z^{(k)}=\left\{z_1^{(k)},z_2^{(k)},\cdots,z_N^{(k)}\right\}$ 表示第 k 个 $(k=1,2,\cdots,K)$ 高斯混合成分生成样本 $\boldsymbol{y}$ 的指示变量值. 因此, 对于 $i=1,2,\cdots,N$, $z_i^{(k)}=1$ 的概率 $p\left(z_i^{(k)}\right)$ 对应于 π_k. 根据贝叶斯定理, $z_i^{(k)}$ 的后验分布对应于

$$p\left(z_i^{(k)}=1\,|\boldsymbol{y}_i\right)=\frac{p\left(z_i^{(k)}=1\right)\cdot p\left(\boldsymbol{y}_i\,\middle|z_i^{(k)}=1\right)}{p(\boldsymbol{y}_i)}=\frac{\pi_k\cdot p(\boldsymbol{y}_i\,|\mu_k,\Sigma_k)}{\sum\limits_{l=1}^{K}\pi_l\cdot p(\boldsymbol{y}_i\,|\mu_l,\Sigma_l)}$$

换言之, $p\left(z_i^{(k)}=1\,|\boldsymbol{y}_i\right)$ 给出了样本 $\boldsymbol{y}_i$ 由第 k 个高斯混合成分生成的后验概率. 为方便叙述, 将其简记为 $\left\langle z_i^{(k)}\right\rangle(k=1,2,\cdots,K)$.

当高斯混合分布 (1) 已知时, 高斯混合聚类将把样本集 D 划分为 K 个簇, 样本 $\boldsymbol{y}_i$ 的簇标记 λ_i.

$$\lambda_i=\underset{k\in\{1,2,\cdots,K\}}{\arg\max}\left\langle z_i^{(k)}\right\rangle \tag{2}$$

可见, 高斯混合聚类的本质是采用概率模型 (高斯分布) 对原型进行刻画, 簇划分则由原型对应后验概率确定. 因一个簇对应一个中心点, 隶属于每一个簇 C 的数据样本将聚类在簇中心点附近. 高斯混合聚类模型效果越好, 所估计的簇中心点与实际簇的中心点之间距离将越小甚至重合.

2 删失型数据缺失机制

2.1 数据缺失机制

依据文献 [12] 将数据缺失机制分为四种类型, 包括随机缺失、完全随机缺失、取决于未被观测因素的缺失 (可以通过未被观察或记录的数据进行预测的) 以及和仅依赖于缺失值自身的缺失机制. 后两种缺失机制即为这里将定义的非随机缺失.

在数据空间 $\mathbf{R}^d$ 中, 令 A 为一个实数集合, 设 $\boldsymbol{1}_{\boldsymbol{A}}\left(\boldsymbol{y}_i^{(j)}\right)$ 为一个指示变量, 表示 $\boldsymbol{y}$ 的元素 $\boldsymbol{y}_i^{(j)}$ 在集合 A 中是否存在观察值. 若 $\boldsymbol{y}_i^{(j)}\in A$,则 $\boldsymbol{1}_{\boldsymbol{A}}\left(\boldsymbol{y}_i^{(j)}\right)=1$, 否则 $\boldsymbol{1}_{\boldsymbol{A}}\left(\boldsymbol{y}_i^{(j)}\right)=0$. 这里 $\boldsymbol{y}_i$ 不区分变量及其真实值, 而将其观测值记为 $\boldsymbol{y}_i^*$. 令 $\boldsymbol{y}_i^{(ob)}$ 作为 $\boldsymbol{y}_i$ 中不存在缺失的部分, $\boldsymbol{y}_i^{(mi)}$ 表示 $\boldsymbol{y}_i$ 中存在缺失值的部分, 那么 $\boldsymbol{y}_i^*=\left[\boldsymbol{y}_i^{(ob)},\boldsymbol{y}_i^{(mi)}\right]^{\mathrm{T}}$.

定义 1. 如果对所有 $\boldsymbol{y}_i^{(ob)}$ 和参数 Θ,

$$p\left(\boldsymbol{1}_{\boldsymbol{A}}(\boldsymbol{y})\,\middle|\boldsymbol{y}_i^{(ob)},\boldsymbol{y}_i^{(mi)},\Theta\right)=p\left(\boldsymbol{1}_{\boldsymbol{A}}(\boldsymbol{y})\,\middle|\boldsymbol{y}_i^{(ob)},\Theta\right)$$

则缺失数据机制为随机缺失.

定义 2. 如果对所有 $\boldsymbol{y}_i^{(ob)}$ 和参数 Θ,

$$p\left(\boldsymbol{1}_{\boldsymbol{A}}(\boldsymbol{y})\,\middle|\boldsymbol{y}_i^{(ob)},\boldsymbol{y}_i^{(mi)},\Theta\right)\neq p\left(\boldsymbol{1}_{\boldsymbol{A}}(\boldsymbol{y})\,\middle|\boldsymbol{y}_i^{(ob)},\Theta\right)$$

则缺失数据机制为非随机缺失.

可见, 对于随机缺失数据, 其样本数据及指示变量满足交换性, 而非随机缺失数据不满足这一性质[12]. 当缺失数据是随机缺失时, 可直接使用标准 EM 算法、多值插补、回归等方法揭示缺失机制. 下面引入一类非随机性缺失数据, 即删失数据, 并研究其缺失机制和参数估计方法.

2.2 删失数据的似然函数

这里给出删失数据的定义, 并详细阐述删失数据的缺失机制和似然函数. 在数据空间 $\mathbf{R}^d$ 中, $[\boldsymbol{a},\boldsymbol{b}]^d$ 为一个超矩阵[11], 其中上边界 $\boldsymbol{b}=(b^{(1)},\cdots,b^{(d)})^{\mathrm{T}}$, 下边界 $\boldsymbol{a}=\left(a^{(1)},\cdots,a^{(d)}\right)^{\mathrm{T}}$.

定义 3. 删失数据 (Censored data) 是指 $\boldsymbol{y}_i$ 的观测值满足分段函数:

$$\boldsymbol{y}_i^*=\begin{cases}\boldsymbol{a}, & \boldsymbol{y}_i\leq\boldsymbol{a}\\ \boldsymbol{y}_i, & \boldsymbol{a}<\boldsymbol{y}_i<\boldsymbol{b}\\ \boldsymbol{b}, & \boldsymbol{y}_i\geq\boldsymbol{b}\end{cases}$$

其中, $\boldsymbol{a}<\boldsymbol{y}_i<\boldsymbol{b}$, 是指 $\boldsymbol{y}_i$ 在所有 d 个维度上, 其对应的元素都存在于超矩阵的两个边界元素之间, 此时 $\boldsymbol{y}_i^*=\boldsymbol{y}_i$, 意为观测值等于真实值; 若 $\boldsymbol{y}_i\leq\boldsymbol{a}$, 是指 $\boldsymbol{y}_i$ 在所有 d 个维度上, 其对应的元素都小于超矩阵的下边界元素, 则 $\boldsymbol{y}_i^*=\boldsymbol{a}$, 意为观测值被赋予区间下界值, 此时数据类型为左删失数据; 若 $\boldsymbol{y}_i\geq\boldsymbol{b}$, 是指 $\boldsymbol{y}_i$ 在所有 d 个维度上, 其对应的元素都大于超矩阵的上边界元素, 则 $\boldsymbol{y}_i^*=\boldsymbol{b}$, 意为观测值被赋予区间上界值, 此时数据类型为右删失数据.

换言之, $\boldsymbol{y}_i$ 中的缺失部分 $\boldsymbol{y}_i^{(mi)}$ 被分别赋予 $\boldsymbol{a}$ 或 $\boldsymbol{b}$ 对应维度上的元素值. 为分析概率密度和估计参数, 假设 $\boldsymbol{y}_i^{(ob)}$ 的元素个数为 $J_1, \boldsymbol{y}_i^{(mi)}$ 的元素个数为 J_2, 且 $J_1+J_2=d$. 不妨进一步假设, $\boldsymbol{y}_i^{(ob)}=\left(\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}, \cdots, \boldsymbol{y}_i^{(J_1)}\right), \boldsymbol{y}_i^{(mi)}=\left(\boldsymbol{y}_i^{(J_1+1)}, \boldsymbol{y}_i^{(J_1+2)}, \cdots, \boldsymbol{y}_i^{(d)}\right)$. 对于删失数据, $A=[\boldsymbol{a}, \boldsymbol{b}]^d$. 为简化, 令 $\delta_{ij}=1-\mathbf{1}_{\boldsymbol{A}}\left(\boldsymbol{y}_i^{(j)}\right)$, 当 $\delta_{ij}=1$ 时, 表示 $\boldsymbol{y}_i^{(j)}$ 因删失而存在缺失数据, 其对应观测值被赋予边界值; 相应地, $\delta_{ij}=0$, 表示 $\boldsymbol{y}_i^{(j)}$ 不存在缺失数据, 即观测值等同于真实值. $\boldsymbol{y}$ 观测值的样本删失率 $p_{ce}=(\sum_i \sum_j \delta_{ij})/nd$. 对于一维数据, 删失率 $p_{ce}=n_{ce}/n$, 其中 n_{ce} 是存在删失的样本数.

根据删失数据的定义, $\boldsymbol{y}_{1:n}$ 的部分真实值 (如序数为 $n_1+1, \cdots, n$ 的值) 被修改. 那么, 其被修改后的数据 (不存在缺失部分的值、和缺失部分的修改值) 构成新数据集, 记为 $\boldsymbol{x}_{1:n}$. 对于 $\forall i, \forall j$, 有

$$\boldsymbol{x}_i^{(j)}=\boldsymbol{y}_i^{(j)}\mathbf{1}_{[\boldsymbol{a}^{(j)}, \boldsymbol{b}^{(j)}]}\left(\boldsymbol{y}_i^{(j)}\right)+\boldsymbol{a}^{(j)}\mathbf{1}_{(-\infty, \boldsymbol{a}^{(j)})}\left(\boldsymbol{y}_i^{(j)}\right)+\boldsymbol{b}^{(j)}\mathbf{1}_{(\boldsymbol{b}^{(j)}, \infty)}\left(\boldsymbol{y}_i^{(j)}\right)$$

其中, 当 $\boldsymbol{y}_i^{(j)} \in A, 1_A\left(\boldsymbol{y}_i^{(j)}\right)=1$, 否则 $1_A\left(\boldsymbol{y}_i^{(j)}\right)=0$. 且 $(-\infty, \boldsymbol{a}^{(j)})$ 表示小于 a_i 的真实值所在的超矩阵, (b_i, ∞) 表示大于 b_i 的真实值所在的超矩阵. 因此,

$$\boldsymbol{a}^{(j)} \leq \boldsymbol{x}_i^{(j)} \leq \boldsymbol{b}^{(j)}, i=1, \cdots, n, j=1, \cdots, d$$

与缺失数据机制对应, 但因每一个样本 $\boldsymbol{y}_i$ 的删失模式会不一样, 而使用 i_m 和 i_o 分别表示删失和未删失数据的坐标序号集, 故 $\boldsymbol{y}_{i\in i_m}$ 和 $\boldsymbol{x}_{i\in i_m}$ 分别指删失部分的缺失值 (缺失时的真实值) 和删失后的改写值 (简称删失值), $\boldsymbol{y}_{i\in i_o}$ 和 $\boldsymbol{x}_{i\in i_o}$ 分别指原数据不存在缺失的部分与删失型数据对应的部分值, 尽管没有删失时它们值等同. 那么 $\boldsymbol{y}_i^*=\left[\boldsymbol{y}_i^{(ob)}, \boldsymbol{y}_i^{(mi)}\right]^{\mathrm{T}}=[\boldsymbol{y}_{i\in i_o}, \boldsymbol{y}_{i\in i_m}]^{\mathrm{T}}$. 同时, $\boldsymbol{x}_i=[\boldsymbol{x}_{i\in i_o}, \boldsymbol{x}_{i\in i_m}]^{\mathrm{T}}$.

为简化, 将 $\boldsymbol{y}$ 的数据空间划分为 $\{\mathcal{Y}_t | t=0, 1, \cdots, T\}$, 其中当 $\boldsymbol{y}_i^{(j)} \in \mathcal{Y}_0=\Pi_{i=1}^d\left[\boldsymbol{a}^{(j)}, \boldsymbol{b}^{(j)}\right]$, 此时数据不存在删失; 而当 $\boldsymbol{y}_i^{(j)} \in \mathcal{Y}_t, t>0$ 时, 数据发生删失. 将删失部分调整后的观测值 $\boldsymbol{x}$ 的数据空间划分为 $\{\mathcal{X}_t | t=1, \cdots, T\}$, 注意, 这里没有涵盖不存在删失的部分, 即 $\boldsymbol{x}$ 的数据空间划分不涵盖 $\mathcal{X}_0$. 对于 $\boldsymbol{y}_i \in \mathcal{Y}_0$, 观测值 $\boldsymbol{x}_i$ 的似然函数如下:

$$f(\boldsymbol{x}_i)=f(\boldsymbol{y}_i) \tag{3}$$

而对于 $\boldsymbol{y}_i$ 缺失机制, 有 $\boldsymbol{y}_i \in \mathcal{Y}_{t_i}, t_i>0$, 其似然函数如下:

$$f(\boldsymbol{x}_i)=\int_{\mathcal{X}_{t_i}} f(\boldsymbol{y}_{i_m}, \boldsymbol{y}_{i_o})\,\mathrm{d}\boldsymbol{y}_{i_m}=f(\boldsymbol{x}_{i_o})\int_{\mathcal{X}_{t_i}} f(\boldsymbol{y}_{i_m}|\boldsymbol{x}_{i_o})\,\mathrm{d}\boldsymbol{y}_{i_m} \tag{4}$$

式中将 $f(\boldsymbol{y}_{i_m}, \boldsymbol{y}_{i_o})$ 分解为 $f(\boldsymbol{x}_{i_o})$ 和 $f(\boldsymbol{y}_{i_m}|\boldsymbol{x}_{i_o})$ 两部分. $\mathcal{X}_{t_i}$ 为 $\boldsymbol{x}_i$ 所属的数据空间划分, 因每一个向量删失模式会不一样. 只对删失数据坐标序数进行积分, $\mathcal{X}_{t_i}(t_i>0)$ 表示相应的积分范围. 例如, 当 $\boldsymbol{x}_i^{(1)}=a_1, \boldsymbol{x}_i^{(2)}=b_2$, 同时其他元素严格的在 a_i 和 b_i 之间, 那么

$$\mathcal{Y}_{t_n}=(-\infty, a_1)\times(b_2, \infty)\times\Pi_{i=3}^d[a_i, b_i]$$

$$\mathcal{X}_{t_n}=(-\infty, a_1)\times(b_2, \infty)$$

并且关于 $f(\boldsymbol{x}_i)$ 推导式 (4) 的右边部分转化为:

$$f(\boldsymbol{x}_{i_o})\int_{-\infty}^{a_1}\int_{b_2}^{\infty} f(\boldsymbol{y}_{i_{m1}}, \boldsymbol{y}_{i_{m2}}|\boldsymbol{x}_{i_o})\,\mathrm{d}\boldsymbol{y}_{i_{m1}}\mathrm{d}\boldsymbol{y}_{i_{m2}}$$

其中, $\boldsymbol{y}_{i_{m1}}, \boldsymbol{y}_{i_{m2}} \in \boldsymbol{y}_{i_m}$ 为数据 $\boldsymbol{y}$ 存在删失型缺失的两个维度, 且 $\boldsymbol{y}_{i_{m1}} \in (-\infty, a_1)$, $\boldsymbol{y}_{i_{m2}} \in (b_2, \infty)$.

3　高斯混合聚类的参数估计

高斯混合聚类参数估计主要包括成分的期望、方差和对应的混合系数. 嵌套标准 EM 的高斯混合聚类算法, 这里简记为 EMGM. 并将针对删失数据所提出的改进算法, 即嵌套删失型数据期望最大化的高斯混合聚类算法, 简记为 cenEMGM 算法.

3.1　基于高斯混合聚类的标准算法 EMGM

对于独立观测变量集合 $\boldsymbol{y}_{1:n}$, 参数空间 Θ, 第 k 个成分 f_k 和簇指示变量 $z_i^{(k)}$, 对数似然函数为:

$$\mathcal{L}(\Theta)=\sum_i\sum_k z_i^{(k)}\left[\ln\pi_k+\ln f_k(\boldsymbol{y}_i)\right]=\sum_i\sum_k z_i^{(k)}\left[\ln\pi_k-\frac{1}{2}\ln|\Sigma_k|-\frac{1}{2}\mathrm{tr}\left((\Sigma_k)^{-1}(\boldsymbol{y}_i-\mu_k\boldsymbol{I}_{n_1})(\boldsymbol{y}_i-\mu_k\boldsymbol{I}_{n_1})^{\mathrm{T}}\right)\right]+Const$$

其中, $\Theta_k=(\pi_k, \mu_k, \Sigma_k)$ 表示第 k 个成分的参数, $(\Sigma_k)^{-1}$ 表示 Σ_k 的倒数, $Const$ 表示常数, $tr(\cdot)$ 表示矩阵的迹, $\boldsymbol{I}_n$ 表示值全为 1 的 $1\times n$ 向量.

根据标准的期望最大化算法[23], 其假设为数据存在随机缺失. 对于独立观测变量集合 $\boldsymbol{y}_{1:n}, \Theta, \Theta^{old}$ 和 Θ^{new} 分别为参数空间, 算法中更新前的参数及更新后的参数.

算法第一步 (步骤 E): 计算期望函数 $Q\left(\Theta; \Theta^{old}\right)=\mathrm{E}\left[\mathcal{L}(\Theta)\left|\boldsymbol{y}_{1:n}; \Theta^{old}\right.\right]$, 步骤 E 可以简化为计算条件

概率:

$$\left\langle z_i^{(k)}\right\rangle = p\left(z_i^{(k)}=1\left|\boldsymbol{y}_{1:n};\Theta^{old}\right.\right)=\frac{\pi_k f_k\left(\boldsymbol{y}_{1:n}\right)}{\sum\limits_l \pi_l f_l\left(\boldsymbol{y}_{1:n}\right)} \tag{5}$$

第二步 (步骤 M): 寻找新的参数集 Θ^{new}, 使得 $\Theta^{new}=\arg\max_\Theta Q\left(\Theta;\Theta^{old}\right)$. 更新后的参数 $\Theta^{new}=(\widehat{\pi}_k,\widehat{\mu}_k,\widehat{\Sigma}_k)$, 形成一个更新的闭环形式:

$$\widehat{\pi}_k=\frac{1}{n}\sum_i\left\langle z_i^{(k)}\right\rangle \tag{6}$$

$$\widehat{\mu}_k=\frac{\sum\limits_i\left\langle z_i^{(k)}\right\rangle \boldsymbol{y}_{1:n}}{\sum\limits_i\left\langle z_i^{(k)}\right\rangle} \tag{7}$$

$$\widehat{\Sigma}_k=\frac{\sum\limits_i\left\langle z_i^{(k)}\right\rangle \left(\boldsymbol{y}_{1:n}-\widehat{\mu}_k\right)\left(\boldsymbol{y}_{1:n}-\widehat{\mu}_k\right)^{\mathrm{T}}}{\sum\limits_i\left\langle z_i^{(k)}\right\rangle} \tag{8}$$

该算法不断迭代 E 步和 M 步, 直至收敛. 以最后获得的更新参数作为 Θ 的最优估计值.

3.2 估计算法的有效性

对于数据的真实参数 Θ, $\hat{\Theta}_n^I$ 为 Θ 的初始估计参数, 全数据 $D=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$, 其概率密度为 $p_D(D,\Theta)$. $\boldsymbol{y}_i$ 对应的删失型缺失数据 $\boldsymbol{y}_i^*$, 可观测部分数据的概率密度为 $f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{x}\right.\right)$. 根据缺失数据的半参数模型推理相关理论[24], 其得分向量 (Score function) 记为 $S^F(D,\Theta),S^F(D,\Theta)=\dfrac{\partial\ln p_D(D,\Theta)}{\partial\Theta}$. 在合适的正则性条件下有以下引理.

引理 1. 通过最大似然估计方法获得全数据的参数 $\hat{\Theta}_n^F$, 即求解全数据得分向量方程 $\sum_{i=1}^n S^F(\boldsymbol{y}_i,\Theta)=0$, 得到

$$\sqrt{n}(\hat{\Theta}_n^F-\Theta)\to N(0,\{I^F(\Theta)\}^{-1}) \tag{9}$$

其中, $I^F(\Theta)$ 为全数据信息矩阵, $I^F(\Theta)=\mathrm{E}[S^F(D,\Theta)S^{F^{\mathrm{T}}}(D,\Theta)]$.

通过正则渐近线性法 (Regular and asymptotically linear, RAL)[24] 获得全数据的参数记为 $\hat{\Theta}_n^I$, 即求解全数据得分向量方程 $\sum_{i=1}^n S^F(\boldsymbol{y}_i,\Theta)=0$.

引理 2. 对于 RAL 方法估计的参数, $\hat{\Theta}_n^I$ 应满足:

$$\sqrt{n}(\hat{\Theta}_n^I-\Theta)=\frac{\sum\limits_{i=1}^n\left[\psi_{eff}(\boldsymbol{y}_i^*)+h(\boldsymbol{y}_i^*)\right]}{\sqrt{n}}+\mathrm{o}_p(1)$$

其中, $\psi_{eff}(\boldsymbol{y}_i^*)$ 为有效影响函数, $\psi_{eff}(\boldsymbol{y}_i^*)=\left\{\mathrm{E}[S(\boldsymbol{y}_i^*)S^T(\boldsymbol{y}_i^*)]\right\}^{-1}S(\boldsymbol{y}_i^*)$, 且将 RAL 的影响函数记为 $q(\boldsymbol{y}_i^*)$, 则 $\mathrm{E}[q(\boldsymbol{y}_i^*)S^T(\boldsymbol{y}_i^*)]=I^{q\times q}$, $I^{q\times q}$ 为单位矩阵; q 维度的随机变量函数 $h(\boldsymbol{y}_i^*)$ 的均值是 0, 且 $h(\boldsymbol{y}_i^*)S^T(\boldsymbol{y}_i^*)=0^{q\times q},\mathrm{o}_p(1)$ 为 1 的高阶无穷小.

对于 $\arg\max_\Theta Q\left(\Theta;\Theta^{old}\right)$, 根据全数据参数估计的引理, 存在关于期望最大化算法估计删失型缺失数据的定理.

定理 1. 令全数据 $D=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$, 对应的删失型缺失数据 $\boldsymbol{y}_i^*$, 对缺失数据使用逐步更新的 EMGM 算法估计参数, $\hat{\Theta}_n^{EM}$ 可通过以下方程求解.

$$\sum_{i=1}^n\mathrm{E}\left\{S^F(D,\hat{\Theta}_n^{EM})|\boldsymbol{y}_i^*,\hat{\Theta}_n^I\right\}=0 \tag{10}$$

其中, $\hat{\Theta}_n^I$ 为 Θ 的初始估计参数, $\mathrm{E}[S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I]$ 为在给定缺失数据 $\boldsymbol{y}_i^*$ 和 RAL 估计参数 $\hat{\Theta}_n^I$ 下的原数据的得分向量, 对于 EMGM 算法 $E[S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I]=\mathrm{E}\left[\dfrac{\partial\mathcal{L}(\Theta)}{\partial\Theta}\right]$.

依据第 3.2 节给定删失数据及其似然函数, cen-EMGM 算法首先计算完全对数似然函数的期望:

$$Q_c\left(\Theta;\Theta^{old}\right)=\mathrm{E}\left\{\mathcal{L}(\Theta)\left|\boldsymbol{x}_{1:n};\Theta^{old}\right.\right\}=\mathrm{E}\left\{\sum_i\sum_k z_i^{(k)}\left[\ln\pi_k-\frac{1}{2}\ln|\Sigma_k|-\frac{1}{2}\mathrm{tr}\left((\Sigma_k)^{-1}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\right)\right]\left|\boldsymbol{x}_{1:n};\Theta^{old}\right.\right\}$$

其中, $\boldsymbol{V}=(\boldsymbol{y}_i^*-\mu_k\boldsymbol{I}_n)$, $\boldsymbol{I}_n$ 表示值全为 1 的 $1\times n$ 向量.

将 $z_i^{(k)}$ 在给定 $\boldsymbol{x}_i$ 时的后验概率 $p\left(z_i^{(k)}=1\left|\boldsymbol{x}_i\right.\right)$ 简记为 $p_{z|x}$. 计算关于条件分布的期望

$$\mathrm{E}\left[\left(z_i^{(k)}\boldsymbol{y}_{i_m}\right)\left|\boldsymbol{x}_i\right.\right]=p_{z|x}\mathrm{E}\left[\boldsymbol{y}_{i_m}\left|\boldsymbol{x}_i,\,z_i^{(k)}=1\right.\right]$$

$$\mathrm{E}\left[z_i^{(k)}\boldsymbol{y}_{i_m}\boldsymbol{y}_{i_m}^{\mathrm{T}}\left|\boldsymbol{x}_i\right.\right]=p_{z|x}\mathrm{E}\left[\boldsymbol{y}_{i_m}\boldsymbol{y}_{i_m}^{\mathrm{T}}\left|\boldsymbol{x}_i,\,z_i^{(k)}=1\right.\right]$$

故而推导出 $z_i^{(k)}$ 的后验概率为:

$$\left\langle z_i^{(k)}\right\rangle=p\left(z_i^{(k)}=1\left|\boldsymbol{x}_i\right.\right)=\frac{\pi_k f_k\left(\boldsymbol{x}_i\right)}{\sum_l\pi_l f_l\left(\boldsymbol{x}_i\right)} \tag{11}$$

该式子可以由式 (4) 进一步推导出结果.

结合高斯混合分布定义 (1), 针对 $\boldsymbol{y}^{(mi)}$ 的条件概率分布, $f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{x}\right.\right)$, 推导其条件分布期望. 因为 $f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{y}^{(ob)}\right.\right)$ 是正态密度函数且满足

$$f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{x}\right.\right)=f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{y}^{(ob)},\boldsymbol{y}\in\mathcal{Y}_c\right.\right)=\frac{f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{y}^{(ob)}\right.\right)}{\int_{\mathcal{X}_{t_n}}f_k\left(\boldsymbol{y}^{(mi)}\left|\boldsymbol{y}^{(ob)}\right.\right)\mathrm{d}\boldsymbol{y}^{(mi)}}\mathbf{1}_{\mathcal{X}_c}\left(\boldsymbol{y}^{(mi)}\right)$$

其中, $\mathbf{1}_{\mathcal{X}_c}\left(\boldsymbol{y}^{(mi)}\right)$ 表示 $\boldsymbol{y}^{(mi)}$ 在集合 $\mathcal{X}_c$ 中是否存在观察值, 若存在则为 1, 否则为 0.

条件密度 $f_k\left(\boldsymbol{y}^{(mi)}|\boldsymbol{x}\right)$ 是在 $\mathcal{X}_c$ 上的截尾正态密度函数, 那么计算关于 Q_c 的充分统计量:

$$\begin{aligned}\langle \boldsymbol{y}_{i_m}|k\rangle =&\mathrm{E}\left[\boldsymbol{y}_{i_m}|\boldsymbol{x}_i,\, z_i^{(k)}=1\right]=\\ &\mathrm{E}\left[\boldsymbol{y}_{i_m}|\boldsymbol{y}_{i_o},\boldsymbol{y}_i\in\mathcal{Y}_{t_n},\, z_i^{(k)}=1\right]=\\ &\mathcal{M}^1\left(\mu_{k,i_m|i_o},\Sigma_{k,i_m|i_o};\mathcal{X}_{t_n}\right)\end{aligned}$$

$$\begin{aligned}\left\langle \boldsymbol{y}_{i_m}\boldsymbol{y}_{i_m}^{\mathrm{T}}|k\right\rangle =&\mathrm{E}\left[\boldsymbol{y}_{i_m}\boldsymbol{y}_{i_m}^{\mathrm{T}}|\boldsymbol{x}_i,\, z_i^{(k)}=1\right]=\\ &\mathrm{E}\left[\boldsymbol{y}_{i_m}\boldsymbol{y}_{i_m}^{\mathrm{T}}|\boldsymbol{y}_{i_o},\,\boldsymbol{y}_i\in\mathcal{Y}_{t_n},z_i^{(k)}=1\right]=\\ &\mathcal{M}^2\left(\mu_{k,i_m|i_o},\Sigma_{k,i_m|i_o};\mathcal{X}_{t_n}\right)\end{aligned}$$

其中, $\mu_{k,i_m|i_o}$ 和 $\Sigma_{k,i_m|i_o}$ 分别是 $f_k\left(\boldsymbol{y}_i^{(mi)}\middle|\boldsymbol{y}_i^{(ob)}\right)$ 的均值和方差. $\mathcal{M}^1(\cdot)$ 和 $\mathcal{M}^2(\cdot)$ 分别表示截尾正态分布的一阶估计量和二阶估计量. 关于估计量的计算详见文献 [11].

定理 2. 全数据 $D=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$, 对应的删失型缺失数据 $\boldsymbol{y}_i^*$, 在给定缺失数据 $\boldsymbol{y}_i^*$ 和 RAL 估计参数 $\hat{\Theta}_n^I$ 下的原数据的得分向量 $\mathrm{E}[S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I]$, 对缺失数据使用 cenEMGM 算法估计参数, $\hat{\Theta}_n^{cEM}$ 满足

$$\sqrt{n}(\hat{\Theta}_n^{cEM}-\Theta)=\frac{\sum\limits_{i=1}^{n}\mathrm{E}[S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I]}{\sqrt{n}I^F(\Theta)}+\mathrm{o}_p(1)\tag{12}$$

其中, $I^F(\Theta)$ 为全数据信息矩阵, Θ 为数据的真实参数, 对于 cenEMGM 算法 $\mathrm{E}[S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I]=\mathrm{E}\left[\dfrac{\partial Q_c\left(\Theta;\hat{\Theta}_n^I\right)}{\partial\Theta}\right]$.

证明. 因 cenEMGM 算法中删失数据的对数似然函数期望为 $Q_c\left(\Theta;\hat{\Theta}_n^I\right)$, 那么其得分向量的期望

$$\mathrm{E}[S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I]=\mathrm{E}\left[\frac{\partial Q_c\left(\Theta;\hat{\Theta}_n^I\right)}{\partial\Theta}\right]$$

其估计参数 $\hat{\Theta}_n^{cEM}$ 通过以下方程求解.

$$\sum_{i=1}^{n}\mathrm{E}\left\{S^F(D,\hat{\Theta}_n^{cEM})|\boldsymbol{y}_i^*,\hat{\Theta}_n^I\right\}=0$$

故而有

$$\begin{aligned}&\frac{\sum\limits_{i=1}^{n}\mathrm{E}\left\{S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I\right\}}{\sqrt{n}}+\\ &\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial S^F(D,\Theta)}{\partial\Theta^T}|\boldsymbol{y}_i^*,\Theta\right]\right\}\sqrt{n}\left(\hat{\Theta}_n^{cEM}-\Theta\right)+\\ &\mathrm{o}_p(1)=0\end{aligned}$$

又因为

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial S^F(D,\Theta)}{\partial\Theta^T}|\boldsymbol{y}_i^*,\Theta\right]\to\\ &\mathrm{E}\left\{\mathrm{E}\left[\frac{\partial S^F(D,\Theta)}{\partial\Theta^T}|\boldsymbol{y}_i^*,\Theta\right]\right\}=\\ &\mathrm{E}\left[\frac{\partial S^F(D,\Theta)}{\partial\Theta^T}\right]=-I^F(\Theta)\end{aligned}$$

所以有

$$\begin{aligned}&\frac{\sum\limits_{i=1}^{n}\mathrm{E}\left\{S^F(D,\Theta)|\boldsymbol{y}_i^*,\hat{\Theta}_n^I\right\}}{\sqrt{n}}-\\ &I^F(\Theta)\sqrt{n}\left(\hat{\Theta}_n^{cEM}-\Theta\right)+\mathrm{o}_p(1)=0\end{aligned}$$

□

3.3　针对删失数据的算法 cenEMGM

根据定理 2 获得对数似然函数的期望 Q_c 关于 Θ 最大化的解, 即得到了 $\Theta^{(t)}=\arg\max_\Theta Q_c\left(\Theta;\Theta^{(t-1)}\right)$ 的优化解, $\hat{\Theta}_n^{cEM}=\left(\hat{\pi}_k,\hat{\mu}_k,\hat{\Sigma}_k\right)$.该算法的步骤 $t\geq 1$, 并且 $\Theta^{(0)}$ 表示初始值, 可通过 K-means 聚类方法获得赋值. 求解的高斯混合聚类的混合系数 π_k 为:

$$\widehat{\pi}_k=\frac{1}{n}\sum\nolimits_i\left\langle z_i^{(k)}\right\rangle\tag{13}$$

同时, μ_k 和 Σ_k 关于 $\arg\max_\Theta Q_c\left(\Theta;\Theta^{(t-1)}\right)$ 的优化解分别为:

$$\hat{\mu}_k=\frac{\sum_i\left\langle z_i^{(k)}\right\rangle\cdot\varphi}{\sum_i\left\langle z_i^{(k)}\right\rangle}\tag{14}$$

$$\widehat{\Sigma}_k=\frac{\sum_i\left\langle z_i^{(k)}\right\rangle S_i^{(k)}}{\sum_i\left\langle z_i^{(k)}\right\rangle}\tag{15}$$

其中, $\varphi=\left[\boldsymbol{y}_i^{(ob)},\langle\boldsymbol{y}_{i_m}|k\rangle\right]^{\mathrm{T}}$, $S_i^{(k)}=\hat{V}\hat{V}^{\mathrm{T}}+\begin{bmatrix}0&0\\0&R_i^{(k)}\end{bmatrix}$. 且 $\hat{V}=\varphi-\widehat{\mu}_k$, $R_i^{(k)}=\left\langle\boldsymbol{y}_{i_m}\boldsymbol{y}_{i_m}^{\mathrm{T}}|k\right\rangle-\langle\boldsymbol{y}_{i_m}|k\rangle\langle\boldsymbol{y}_{i_m}|k\rangle^{\mathrm{T}}$.

式 (13)~(15) 作为标准 EM 算法式 (6)~(8) 针对删失型缺失数据的改进. 式 (13) 与 (6) 在形式上没有变化, 从理论上论证了删失型算法 cenEMGM 与标准算法 EMGM 在混合系数上一致. 式 (14) 与 (7) 相比较发现, 在删失数据算法 cenEMGM 中, $\boldsymbol{y}_{1:n}$ 的删失部分被条件均值 $\langle\boldsymbol{y}_{i_m}|k\rangle$ 代替. 式 (15) 与 (8) 相比较发现, 删失数据算法 cenEMGM 的 $(\boldsymbol{y}_{1:n}-\widehat{\mu}_k)(\boldsymbol{y}_{1:n}-\widehat{\mu}_k)^{\mathrm{T}}$ 被样本校正协方差 R_k^n 所替代. 标准算法 EMGM 即为算法 cenEMGM 处理不存在删失数据时的特定情形.

3.4 模型检验准则

为了防止算法出现过拟合并计算估计值和真实值之间的距离, 需要设定模型检验准则. 这里引入信息散度 (Kullback-Leibler divergence, KLD) 和赤池弘次信息准则 (Akaike′s information criterion, AIC)[20, 25]. 信息散度 KLD 公式[25] 为:

$$KLD\left(p\,\|q\right)=\sum_{\boldsymbol{y}}p(\boldsymbol{y})\log_2\left(\frac{p(\boldsymbol{y})}{q(\boldsymbol{y})}\right)$$

其中, $p(\boldsymbol{y})$ 是 $\boldsymbol{y}$ 真实分布的概率密度函数, $q(\boldsymbol{y})$ 是 $\boldsymbol{y}$ 估计分布的概率密度函数. 本文中 $\boldsymbol{y}$ 的概率密度函数由高斯混合分布 (1) 确定. $p(\boldsymbol{y})=\sum_{k=1}^{K}\pi_k\cdot p\left(\boldsymbol{y}\,|\mu_k,\Sigma_k\right),q(\boldsymbol{y})=\sum_{k=1}^{K}\hat{\pi}_k\cdot p\left(\boldsymbol{y}\,\middle|\hat{\mu}_k,\hat{\Sigma}_k\right)$.在算法 EMGM 中, $p(\boldsymbol{y})$ 由式 (6)~(8) 确定; 在算法 cenEMGM 中, $q(\boldsymbol{y})$ 由式 (13)~(15) 确定.

对于 AIC 准则, 其值最小的模型即为最佳模型. 假设模型的误差服从独立正态分布, AIC 可表示为:

$$AIC=2N(\Theta)-2\ln\left(\mathcal{L}\left(\Theta\right)\right)=2[(d-1)+K(d+\frac{d(d+1)}{2})]-2\ln\left(\mathcal{L}\left(\Theta\right)\right)\tag{16}$$

其中, $N(\Theta)$ 是模型算法参数的数量, d 为 D 数据维度, K 为高斯混合模型的成分数量, $\mathcal{L}\left(\Theta\right)$ 是参数集 Θ 的似然函数.

3.5 cenEMGM 算法及分析

嵌套删失型数据期望最大化的高斯混合聚类算法 (cenEMGM) 主要由高斯混合聚类和针对删失数据的期望最大化算法构成, 如算法 1 所示. 第 1) 步初始化参数, 常使用 k-means 算法. 第 2)~10) 步, 运行直至满足停止条件, 跳出循环. 其中第 3)~4) 步, cenEMGM 算法的 E 步, 计算后验概率; 第 5)~9) 步, cenEMGM 算法的 M 步, 计算新的模型参数. 第 11)~13) 步, 划分簇. 算法流程的停止条件是 $\left\|\Theta^{(t)}-\Theta^{(t-1)}\right\|<\varepsilon$, 其中 ε 是一个小的正数 (如 1.0×10^{-6}). 其中, $\left\|\Theta^{(t)}-\Theta^{(t-1)}\right\|:=\max\{\hat{\pi}_k^{(t)}-\hat{\pi}_k^{(t-1)},\hat{\mu}_k^{(t)}-\hat{\mu}_k^{(t-1)},\hat{\Sigma}_k^{(t)}-\hat{\Sigma}_k^{(t-1)}\},k=1,2,\cdots,K$. cenEMGM 算法的计算复杂度 (时间复杂度) 受到样本规模 n 和参数规模 $\left[(d-1)+K\left(d+\frac{d(d+1)}{2}\right)\right]$ 影响, 其中 d 为 D 数据维度, K 为高斯混合模型的成分数量.

算法 1. 嵌套删失型数据期望最大化的高斯混合聚类算法 cenEMGM

输入: $D=\{\boldsymbol{y}_1,\,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\},\,K,\,\varepsilon$;

输出: 簇划分 $\boldsymbol{C}=\{C_1,\,C_2,\cdots,C_K\}$;

1) $C_k=\phi(k=1,\,2,\cdots,K)$, 使用 K-means 算法, 初始化高斯混合聚类模型参数 $\Theta^{(0)}:\left\{\left(\pi_k^{(0)},\,\mu_k^{(0)},\Sigma_k^{(0)}\right)\right\}$

2) **do**

3) **for** $n=1,\,2,\cdots,n$ **do**

4) 用式 (5) 得出 $\left\langle z_i^{(k)}\right\rangle$;

5) **for** $k=1,\,2,\cdots,K$ **do**

6) 用式 (14) 计算新均值向量: $\hat{\mu}_k$;

7) 用式 (15) 计算新协方差矩阵: $\hat{\Sigma}_k$;

8) 用式 (13) 计算新混合系数: $\hat{\pi}_k$;

9) 更新参数 $\Theta^{(t)}:\left\{\left(\hat{\pi}_k^{(t)},\,\hat{\mu}_k^{(t)},\hat{\Sigma}_k^{(t)}\right)\right\}$;

10) **while** $(\left\|\Theta^{(t)}-\Theta^{(t-1)}\right\|<\varepsilon)$

11) **for** $n=1,\,2,\cdots,n$ **do**

12) 根据式 (2) 确定 $\boldsymbol{y}_{1:n}$ 的簇标记 λ_i;

13) 将 $\boldsymbol{y}_{1:n}$ 划入相应的簇: $C_{\lambda_i}=C_{\lambda_i}\cup\{\boldsymbol{y}_{1:n}\}$

return 簇划分 $\boldsymbol{C}=\{C_1,\,C_2,\cdots,\,C_K\}$

cenEMGM 算法的核心步骤主要基于式 (13)~(15). 与之对应的标准 EMGM 算法, 其核心是式 (6)~(8). cenEMGM 算法是针对删失型缺失数据的改进算法, 先根据新均值向量 $\hat{\mu}_k$ 计算新样本规模, 然后计算新混合系数 $\hat{\pi}_k$. 因为样本规模改变, 所以样本方差、删失率、观测数据均值等参数同步做出改变. 针对删失数据修改的这些内容, 使 cenEMGM 算法更灵活, 更能适应含有删失数据的高斯混合聚类.

高斯混合分布中, π_k 是选择第 k 个混合成分的概率, 由式 (8) 和式 (13) 可以看出, 样本删失率间接地通过样本容量影响着 π_k, 所以 p_{ce} 对 π_k 产生影响. 数据质量可以衡量采样机制产生的选择偏差程度[26], 其不仅和估计准确度 ($\hat{\mu}_k-\mu_k$) 有关, 更是与删失率有关. 为了提高模型的准确性, 可以根据删失率调整并确定样本规模 n. 关于样本规模在实验设计中已有讨论[27]. 这里给出样本方差未知时删失率 p_{ce} 与样本规模 n 的结论. 根据统计推断理论, 检验水准 α 时, 预测能力 $(1-\beta)$ 表示, 当所考虑的总体与原假设 H_0 确有差别时, 按照检验水准 α 能够发现拒绝它的概率. 总体方差未知时, 在删失数据缺失率为 p_{ce} 的情况下, 估计样本容量大小如下: $n_0=\frac{p_{ce}\left(1-p_{ce}\right)}{\delta^2}t^2_{\frac{\alpha}{2},n-1}$, 其中 δ 表示估计精度 (即允许误差), $\delta=|\hat{p}_{ce}-p_{ce}|$, $\hat{p}_{ce}$ 为数据分布中的真实缺失率, t 为检验统计量. 对于一定规模的同一数据集, 随着样本删失率 p_{ce} 上升, 参数估计模型的估计能力下降, 导致准确性也降低. 因此, 数据分析中要求样本容量不小于 n_0. 随着数据感知和收集成本下降, 数据可得性变高, 统计机器学习模型使用的数据规模选取常会超过模型的测试能力要求, 且通常会考虑数据的缺失机制[12].

4　数值实验分析

这里使用人工数值实验与真实数据分析, 验证方法的有效性.

4.1　人工数值实验分析

实验从预设分布生成数据集, 并对数据进行删失处理. 在删失数据上, 分别采用嵌套标准 EM 的高斯混合聚类算法 EMGM 和嵌套删失型数据 cenEM 的高斯混合聚类算法 cenEMGM 进行实验分析. 实验结果通过聚类的真实参数与估计参数比较、KL 散度等统计指标进行比较分析.

为在多变量上比较算法, 这里设计两个含有三个成分的二元高斯混合模型的实验. 在两个实验中, 实验数据集 DS-a 的观测值 (Y_1, Y_2) 被设置在 $[10, 50] \times [5, 45]$ 的矩形窗中, 用于右删失型数据和双边删失型数据在 EMGM 算法和 cenEMGM 算法上的实验; 实验数据集 DS-b 的观测值 (Y_1, Y_2) 被设置在 $[-20, 60] \times [-10, 60]$ 的矩形窗中, 用于左删失型数据和双边删失型数据在 EMGM 算法和 cenEMGM 算法上的实验. 右 (左) 删失型缺失是指在变量值域范围内, 设定了观测值上 (下) 界, 且大 (小) 于该上 (下) 界的其他值被赋予该上 (下) 界值, 但并无给定的下 (上) 界. 双边删失型缺失是指在变量值域范围内, 同时设定了观测值上界和下界值, 大于该上界的其他值被赋予该上界值, 且小于该下界的其他值被赋予该下界值. 这里生成的两组数据分别采用了两种删失机制, 并非只讨论一组数据的左删失、右删失及双边删失, 以便体现删失数据边界的多样性和实验的可重复性.

在实验数据集 DS-a 中, 三个分量的中心都在对应的矩形窗内, 参数设置如下: 成分权重为 $\pi = (0.25, 0.40, 0.35)$; 均值为 $\mu_1 = (23.50, 23.50)$, $\mu_2 = (33.50, 23.50)$, $\mu_3 = (40.50, 40.50)$; 方差中, 成分 1 与成分 2 在两个变量之间不存在相关性:

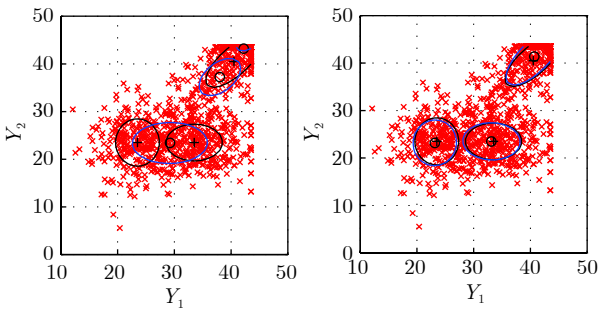
$$\Sigma_1 = \begin{bmatrix} 15 & 0 \\ 0 & 25 \end{bmatrix}; \Sigma_2 = \begin{bmatrix} 25 & 0 \\ 0 & 15 \end{bmatrix}$$

成分 3 的两个变量之间存在相关性:

$$\Sigma_3 = \begin{bmatrix} 25 & 20 \\ 20 & 30 \end{bmatrix}$$

在实验数据集 DS-b 中, 虽然三个成分的中心都在对应的矩形窗内, 但有两个成分的中心落在了下界之外. 参数设置如下: 成分权重和方差分别与实验数据集 DS-a 对应一致. 但它们的均值分别为 $\mu_1 = (-3.50, 23.50)$, $\mu_2 = (33.50, -3.50)$, $\mu_3 = (40.50, 40.50)$.

在每种情形下绘制 1 000 个数据点后, 根据删失型缺失的预设边界, 边界外的所有数据都删失. 在 DS-a 中, 针对右删失缺失型数据, 其上界值设为 43.5, 表明删失类型的 (超) 矩形窗为 $[10, 43.5] \times [5, 43.5]$, 其中 10 和 5 为小于其观测值最小值的一个数, 来源于观测值的矩形窗下界, 并不表示删失数据的下界, 并观察到约 862 个数据点未删失, 并使用 EMGM 算法和 cenEMGM 算法进行实验, 如图 1 所示; 若其还存在左删失, 如将其下界值设为 15, 形成双边删失型缺失数据, 表明删失类型的 (超) 矩形窗为 $[15, 43.5] \times [15, 43.5]$, 约 818 个数据点未删失, 如图 2 所示. 类似地, 在 DS-b 中, 针对左删失缺失型数据, 其下界值设为 0, 表明删失类型的 (超) 矩形窗为 $[0, 60] \times [0, 60]$, 其中 60 为大于其观测值最大值的一个数, 来源于观测值的矩形窗上界, 并不表示删失数据的上界, 约 484 个数据点未删失, 如图 2 所示; 若其还存在右删失, 例如其上界值设为 40, 形成双边删失型缺失数据, 表明删失类型的 (超) 矩形窗为 $[0, 40] \times [0, 40]$, 约 241 左右的数据点未删失, 如图 3 所示. 图中小十字表示删失后的数据点, 'o' 和实心椭圆是每个成分在算法估计后的聚类中心和距离为 1 的等高曲线. 其距离使用成对马氏 (Mahalanobis) 距离计算. '+' 和虚线椭圆表示高斯混合模型成分的真实聚类中心和等高曲线.

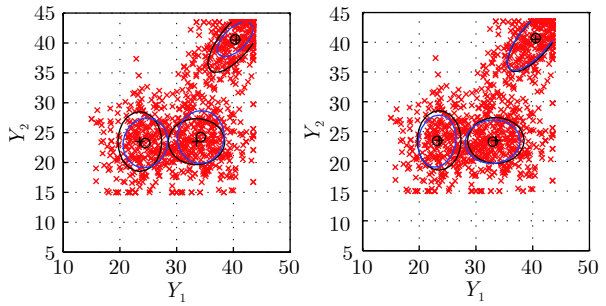


(a) EMGM　　(b) cenEMGM

图 1　在数据集 DS-a 右删失上的两种算法比较

Fig. 1　Comparison of the two algorithms on the dataset DS-a with right censoring

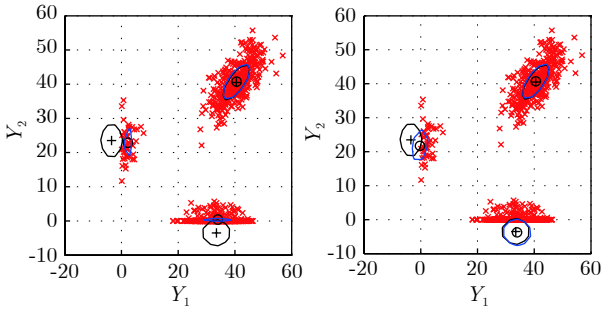
图 1 显示 EMGM 算法和 cenEMGM 算法在二维合成数据 DS-a 右删失上的实验结果. EMGM 算法在该数据集上的结果 (图 1 (a)) 显示, 'o' 和实心椭圆所表示的估计的聚类中心和距离为 1 的等高曲线与 '+' 和虚线椭圆表示高斯混合模型成分的真实聚类中心和等高曲线之间存在显著差异. 而 cenEMGM 算法在该数据集上的结果 (图 1(b)) 显示, cenEMGM 算法估计的聚类中心和等高曲线与



(a) EMGM (b) cenEMGM

图 2 在数据集 DS-a 双边删失上的两种算法比较

Fig. 2 Comparison of the two algorithms on the dataset DS-a with double-side censoring



(a) EMGM (b) cenEMGM

图 3 在数据集 DS-b 左删失上的两种算法比较

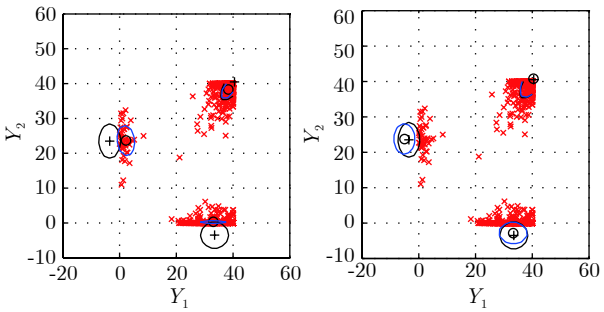
Fig. 3 Comparison of the two algorithms on the dataset DS-b with left censoring

真实聚类中心和等高曲线之间的差异明显减小, 其结果明显优于 EMGM 算法.

图 2 显示 EMGM 算法和 cenEMGM 算法在二维合成数据 DS-a 双边删失上的实验结果. EMGM 算法在该数据集上的结果 (图 2 (a)) 显示, 聚类中心和距离为 1 的等高曲线比 EMGM 算法 (图 1 (a)) 明显更接近于真实值. 因为这里除了存在右删失外, 还存在左删失. 尽管缺失率更高, 但观测到的数据 (未删失部分) 的均值更接近真实值. 同时可见, cenEMGM 算法估计 (图 2 (b)) 的聚类中心和真实聚类中心之间的差异也明显更小, 其结果进一步表明 cenEMGM 算法在处理删失数据聚类问题上明显优于 EMGM 算法.

图 3 显示 EMGM 算法和 cenEMGM 算法在二维合成数据 DS-b 左删失上的实验结果. '+' 表示高斯混合模型成分的真实聚类中心, 其中两个已落在了值域的下界之外. EMGM 算法的结果 (图 3 (a)) 显示, 其估计的聚类中心 ('o') 和等高曲线 (实心椭圆) 没有超出值域的下界, 表明估计值与对应的真实值之间存在显著差异. 而 cenEMGM 算法的估计结果 (图 3 (b)) 显示, 其估计的聚类中心和等高曲线与真实值之间的差异明显更小. 对于图 3 (b) 图中靠近 Y_2 坐标轴的成分, 尽管其估计值与真实值之间尚存在一些差异, 但这一差异与 EMGM 算法所表现出的差异已经小很多, 且另外两个成分的估计值与真实值之间几乎无差异, 因此这些结果进一步表明 cenEMGM 算法在这类数据聚类上更优于 EMGM 算法.

图 4 显示 EMGM 算法和 cenEMGM 算法在二维合成数据 DS-b 双边删失上的实验结果. 三个成分的聚类中心真实值 ('+') 都在下界或上界之外. EMGM 算法在该数据集上的结果 (图 4 (a)) 显示, 三个成分的估计的聚类中心和距离为 1 的等高曲线与真实值之间都存在显著差异. 与此相反, cenEMGM 算法在该数据集上的结果 (图 4 (b)) 显示, 其估计值也可以位于上下界之外, 更接近真实聚类中心和等高曲线, 即估计值与真实值之间的差异明显变小. 结果表明 cenEMGM 算法在处理这类删失数据聚类时明显优于 EMGM 算法.



(a) EMGM (b) cenEMGM

图 4 在数据集 DS-b 双边删失上的两种算法比较

Fig. 4 Comparison of the two algorithms on the dataset DS-b with double-side censoring

此外, 进行 100 次重复实验, 记录多次实验结果在 KLD 值与 AIC 值上的平均值和方差. 实验合成数据集真实分布和估计分布之间的 KLD 值见表 1, 对于参数估计的两种算法 AIC 值比较见表 2. 结果表明, 对于两种算法在同一数据集上的表现, 不论是 KLD 值还是 AIC 值, cenEMGM 算法的值都小于对应 EMGM 算法的值, 说明在删失型缺失数据参数估计上 cenEMGM 算法优于 EMGM 算法. 对于同一算法在不同数据集上的表现, 因双边删失比对应的单边删失因缺失而拥有更少的样本数据, 双边删失的 AIC 值小于对应的单侧删失的 AIC 值.

4.2 真实数据分析

数据来源于某大型医院信息系统中的临床数据[4].

表 1　实验合成数据集真实分布和估计分布之间的 KLD 值

Table 1　Kullback-Leibler divergence (KLD) between the true densities and the estimated densities of the synthetic data set

数据集	观测值 (删失)	EMGM	cenEMGM
DS-a 右删失	0.072 ± 0.011	0.261 ± 0.016	0.051 ± 0.003
DS-a 双边删失	0.226 ± 0.017	10.602 ± 1.966	0.028 ± 0.009
DS-b 左删失	4.362 ± 0.393	32.263 ± 4.193	22.583 ± 3.392
DS-b 双边删失	4.219 ± 0.381	30.321 ± 4.128	29.655 ± 3.938

表 2　实验合成数据集参数估计的两种算法 AIC 比较

Table 2　AIC comparison of the two estimation algorithms on the synthetic data set

数据集	EMGM	cenEMGM
DS-a 右删失	12852 ± 594	12349 ± 481
DS-a 双边删失	12782 ± 436	12323 ± 417
DS-b 左删失	9435 ± 317	8815 ± 305
DS-b 双边删失	8759 ± 293	7152 ± 264

这些数据样本包括 554 个相关属性, 其中有 106 个建档属性、23 个检验数据属性、157 个来自实验室信息系统的试验结果属性以及 268 个电子健康档案中病案首页的属性. 根据医学领域专家意见和文献进行属性筛选, 经过数据清理后所得数据集包括 50 个属性, 具体包括年龄、婚龄、孕妇体重指数、红细胞计数、谷氨酰转肽酶、空腹血糖水平值等属性. 根据验证的目的, 这里所使用的数据集为原临床数据集中提取的包含 4 个属性的数据. 这些属性具体为关于孕妇在筛检妊娠期糖尿病过程中的血糖水平值和医生给出的诊断结果, 即是否患有妊娠期糖尿病. 其中包括关于血糖水平值的 3 个属性分别为口服糖耐量试验中的空腹血糖水平值 (Fasting blood sugar level, FBSL)、1 小时血糖水平值 (1h-blood sugar level, 1h-BSL) 和 2 小时后的血糖水平值. 根据国际妊娠合并糖尿病研究组织建议, 妊娠期糖尿病的诊断标准为[4], 空腹血糖水平值高于 5.1 mmol/L、1 小时血糖水平值高于 10 mmol/L 和 2 小时血糖水平值高于 8.5 mmol/L, 满足以上三项中的任一项即诊断为患有妊娠期糖尿病, 数据记录聚类为患病簇, 否则为正常簇. 在电子病历记录与数据联结整合中, 小于等于 10 mmol/L 的血糖水平值记录为原始测量值, 而高于 10 mmol/L 的空腹血糖水平值和 1 小时血糖水平值的数据被记录为 “> 10 mmol/L”型删失型数据. 虽然这些删失型数据能够为诊断结果提供直接的临床证据, 但是这些数据的删失对于进一步探索关于妊娠期糖尿病的风险因子, 以及这些因子对血糖水平值影响的因果关系研究构成困难. 又因妊娠期糖尿病的主要治疗方案包括膳食改变、增加锻炼甚至胰岛素等的药物治疗[28], 但这些治疗方案对以血糖水平值作为结果的影响作用大小是有差异的. 为后续研究这些影响作用, 在使用这些删失型的血糖水平值数据时, 需要对这些数据的分布参数进行较为精确的估计. 本文的聚类算法正是针对这些删失型数据提供分布参数的估计.

从原数据中选择了 917 例数据进行数值计算, 其中 756 例样本属于正常簇, 161 例样本属于患病簇. 在 917 例样本数据中, 以空腹血糖水平值和 1 小时血糖水平值进行分析, 发现 78 例样本数据属于删失型数据, 主要存在于 1 小时血糖水平值上. 对这一数据集, 分别采用 EMGM 算法和 cenEMGM 算法进行高斯混合聚类, 结果如图 5 所示.

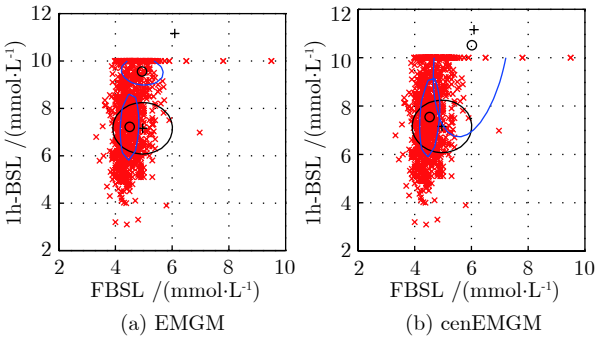


图 5　在血糖测试数据右删失上两种算法比较

Fig. 5　Comparison of the two algorithms on the dataset of blood sugar tests with right-side censoring

图 5 显示了 EMGM 算法和 cenEMGM 算法在删失型血糖水平值数据上的聚类结果. 横坐标为空腹血糖水平值, 纵坐标为 1 小时血糖水平值, 其样本数据关于 “>10 mmol/L”删失. 真实数据中一个成分的聚类中心真实值 (‘+’) 在样本数据所展示的范围内, 为 (4.96, 7.16); 另一个成分的聚类中心真实值 (‘+’) 在样本数据的上界之外, 为 (6.09, 11.16), 即中心值在 1 小时血糖水平值上 “>10 mmol/L”. 图 5 (a) 显示 EMGM 算法在该数据集上存在一个成分的估计聚类中心和距离为 1 的等高曲线与真实值之间存在显著差异, 即估计值所在的聚类中心在 1 小时血糖水平值以下, 而真实值所在的聚类中心在 1 小时血糖水平值以上. 不同的是, 图 5 (b) 显示 cenEMGM 算法在该数据集上的估计值也可以位于上界之外, 使得其更接近真实聚类中心, 说明估计值与真实值之间的差异明显变小. 在模型检验准则上, 对于这一真实数据集, EMGM 算法在真实

分布与估计分布之间的 KLD 值 (12.7) 高于 cenEMGM 算法的 KLD 值 (9.1), 同时后者的 AIC 值 (4 263) 低于前者的 AIC 值 (4 366). 因此, 这些结果说明 cenEMGM 算法在处理真实的删失数据聚类时优于 EMGM 算法.

此外, 为进一步验证方法的有效性, 对于真实数据调整删失率进行拓展, 动态改变删失率而进行计算, 并对聚类中心、AIC 与 KLD 值进行定量对比, 如表 3 所示.

表 3 真实数据及其拓展数据的两种算法比较

Table 3 Comparison of the two algorithms with the real data and its extended data

		EMGM 算法	cenEMGM 算法
右边删失率 8.51%	聚类中心	(4.50, 7.22) (4.94, 9.55)	(4.53, 7.54) (6.01, 10.51)
	KLD	12.7	9.1
	AIC	4366	4263
右边删失率 11.67%	聚类中心	(4.50, 7.20) (4.81, 9.70)	(4.53, 7.54) (6.08, 9.85)
	KLD	11.35	9.08
	AIC	4 290	4 209
双边删失率 15.05%: 右边删失 8.51%, 左边删失 6.54%	聚类中心	(5.10, 7.43) (5.48, 8.56)	(5.10, 7.48) (5.48, 8.94)
	KLD	173.7	158.6
	AIC	2226	−24327

表 3 结果表明, 当右侧删失率从 8.51% 增加到 11.67% 时, 两种算法的聚类中心估计值与真实值 (4.96, 7.16) 和 (6.09, 11.16) 之间的差异增大, KLD 值与 AIC 值减小. cenEMGM 算法的 KLD 值与 AIC 值比 EMGM 算法的对应值小, 说明其在处理删失数据聚类时仍然优于 EMGM 算法. 当将数据拓展为双边删失型数据时, 即在右边删失的基础上增加左边删失 6.54 %, 总体上删失 15.05 % 时, 两种算法的聚类中心估计值与真实值之间的差异进一步增大, 且 KLD 值增大而 AIC 值减小. 总体上, 随着删失率的增加, 算法处理的能力在一定程度上逐渐减弱, 但是 cenEMGM 算法的聚类中心估计值与真实值相对更接近, 且 KLD 值与 AIC 值比 EMGM 算法的对应值更小, 进一步说明其通过聚类在处理删失数据的参数估计时仍然优于 EMGM 算法.

5 结论

删失型数据处理特别是在机器学习或数据挖掘等数据处理中, 作为工程实践和管理中数据处理的焦点问题. 由于删失数据处理的知识有限性, 需要根据删失模式制定合适的算法模型. 尽管当前数据智能处理所面临的数据规模较大, 但选取高价值的实验数据或稀有事件等所面临的删失数据处理仍然显得较为重要. 然而, 现有的缺失数据处理问题主要集中在随机缺失, 对非随机缺失下的删失型数据研究不深, 因此本文根据估计算法的有效性理论, 针对删失数据期望最大化的高斯混合聚类算法 (cenEMGM), 通过关于得分向量期望的方程得出算法估计的最优参数. 与嵌套标准 EM 的高斯混合聚类算法 (EMGM) 相比, 本方法根据删失数据的指示变量调整样本似然函数, 进而改进参数估计的期望最大化算法, 使得高斯混合聚类模型参数估计准确性更高, AIC 信息准则值更小, 聚类效果更好. 并通过数值实验论证了本方法相对于 EMGM 算法的优越性. 更多类型数据中的删失型缺失机制 (模式) 识别、不同删失情形下多种算法有效性分析及其高斯混合聚类算法拓展是下一步工作重点.

References

1 Scrucca L, Raftery A E. Clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R. *Journal of Statistical Software*, 2018: 84

2 O´Hagan A, Murphy TB, Gormley IC, McNicholas PD, Karlis D. Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics & Data Analysis*, 2016, **93**: 18−30

3 Xu M, Yu H Y, and Shen J. New approach to eliminate structural redundancy in case resource pools usingαmutual information. *Journal of Systems Engineering and Electronics*, 2013, **24**(4): 625−633

4 Qiu H, Yu H Y, Wang L Y, Yao Q, Wu S N, Yin C, Deng J. Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. *Scientific Reports*, 2017, **7**(1): 16417

5 Li Xiao-Qing, Tang Hao, Si Jia-Sheng, Miao Gang-Zhong. An improved semi-supervised FCM clustering method for mixed attribute datasets. *Acta Automatica Sinica*, 2018, **44**(12): 2259−2268
(李晓庆, 唐昊, 司加胜, 苗刚中. 面向混合属性数据集的改进半监督 FCM 聚类方法. 自动化学报, 2018, **44**(12): 2259−2268)

6 Xu M, Yu H Y, and Shen J. New algorithm for CBR-RBR fusion with robust thresholds. *Chinese Journal of Mechanical Engineering*, 2012, **25**: 1255−1263

7 Shen Jiang, Yu Hai-Yan, Xu Man. Heterogeneous evidence chains based fusion reasoning for multi-attribute group decision making. *Acta Automatica Sinica*, 2015, **41**: 832−842
(沈江, 余海燕, 徐曼. 实体异构性下证据链融合推理的多属性群决策. 自动化学报, 2015, **41**: 832−842)

8 Yu Hai-Yan, Shen Jiang, Xu Man. ECs-based reasoning for group decision analysis in the mislabeled classification context. *Systems Engineering and Electronic Technology*, 2015, (11): 2546−2553

(余海燕, 沈江, 徐曼. 类别误标下证据链推理的群决策分类方法. 系统工程与电子技术, 2015, (11): 2546−2553)

9 Yu H Y, Shen J, Xu M. Temporal case matching with information value maximization for predicting physiological states. *Information Sciences*, 2016, **367**: 766−782

10 Yu H Y, Shen J, Xu M. Resilient parallel similarity-based reasoning for classifying heterogeneous medical cases in mapreduce. *Digital Communications & Networks*, 2016, **2**(3): 145−150

11 Lee G, Scott C. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 2012, **56**(9): 2816−2829

12 Little R J, and Donald B R. Statistical Analysis with Missing Data. *John Wiley & Sons*, 2019.

13 Linero A R, Daniels M J. Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science*, 2018, **33**: 198−213

14 Fang F, Shao J. Model selection with nonignorable nonresponse. *Biometrika*, 2016, **103**(4): asw039

15 Wu Y J, Fang W Q, Cheng L H, et al. A flexible Bayesian nonparametric approach for fitting the odds to case II interval-censored data. *Journal of Statistical Computation and Simulation*, 2018, **88**(16): 3132−3150

16 Leão J, Leiva V, Saulo H, et al. A survival model with Birnbaum – Saunders frailty for uncensored and censored cancer data. *Brazilian Journal of Probability and Statistics*, 2018, **32**(4): 707−729

17 Goldberg Y, Kosorok M R. Support vector regression for right censored data. *Electronic Journal of Statistics*, 2017, **11**(1): 532−69

18 Xun Li, Zhou Yong. Estimators and their asymptotic properties for quantile difference with left truncated and right censored data. *Acta Mathematica Sinica (Chinese Series)*, 2017, **60**(3): 451−464
(荀立, 周勇. 左截断右删失数据分位差估计及其渐近性质. 数学学报, 2017, **60**(3): 451−464)

19 Ma Y, Wang Y. Estimating disease onset distribution functions in mutation carriers with censored mixture data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2014, **63**(1): 1−23

20 Zhou Zhi-Hua. *Machine Learning*, Beijing: Tsinghua University Press, 2016.
(周志华. 机器学习. 北京: 清华大学出版社, 2016.)

21 Cai T T, Ma J, Zhang L. CHIME: Clustering of highdimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 2019, **47**: 1234−1267

22 Chauveau D. A stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning & Inference*, 1995, **46**(1): 1−25

23 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Series B (Methodological)*, 1977: 1−38

24 Tsiatis A. Semiparametric Theory and Missing Data. Springer Science & Business Media, 2007.

25 Wang Yong, et al. A hybrid user similarity model for collaborative filtering. *Information Sciences*, 2017, **418**: 102−118

26 Yu H, Chen J, Wang J N, Chiu Y L, Qiu H, Wang L Y. Identification of the differential effect of city-level on the Gini coefficient of healthcare service delivery in online health community. *International Journal of Environmental Research and Public Health*, 2019, **16**: 2314

27 Luers B, Klasnja P, Murphy S. Standardized effect sizes for preventive mobile health interventions in micro-randomized trials. *Prevention Science*, 2019, **20**: 100−109

28 McIntyre H D, Catalano P, Zhang C, Desoye G, Mathiesen E R, Damm P. Gestational diabetes mellitus. *Nature Reviews Disease Primers*, 2019, **5**: 47

余海燕 重庆邮电大学副教授. 美国宾西法尼亚州立大学博士后访问学者. 2015 年获得天津大学博士学位. 主要研究方向为统计机器学习, 因果推断. 本文通信作者.
E-mail: yuhy@cqupt.edu.cn
(**YU Hai-Yan** Associate professor at Chongqing University of Posts and Telecommunications (CQUPT). Postdoctoral visiting scholar at The Pennsylvania State University. He received his Ph.D. degree from Tianjin University in 2015. His research interest covers statistical machine learning, causal inference. Corresponding author of this paper.)

陈京京 重庆邮电大学经济管理学院硕士研究生. 主要研究方向为聚类算法和数据缺失机制.
E-mail: chenjingjing_361@163.com
(**CHEN Jing-Jing** Master student at the School of Economics and Management, Chongqing University of Posts and Telecommunications. Her main research interest covers clustering algorithm and data missing mechanism.)

邱 航 电子科技大学计算机科学与工程学院副教授. 2011 年获得电子科技大学计算机应用技术博士学位. 2013 ~ 2014 年英国诺丁汉大学访问学者. 主要研究方向为机器学习和计算机图形学.
E-mail: qiuhang@uestc.edu.cn
(**QIU Hang** Associate professor at the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). He received his Ph. D. degree in computer application technology from UESTC in 2011. From 2013 to 2014, he was a visiting researcher at University of Nottingham, UK. His research interest covers machine learning and computer graphics.)

王 永 重庆邮电大学管理工程系教授. 2007 年于重庆大学获得计算机科学与技术专业博士学位. 主要研究方向为数据分析和信息安全.
E-mail: wangyong_cqupt@163.com
(**WANG Yong** Professor in the Department of Management Engineering, Chongqing University of Posts and Telecommunications. He received his Ph. D. degree in computer science and technology, from Chongqing University, in 2007. His research interest covers data analysis and information security.)

王若凡 天津职业技术师范大学讲师. 2015 年获得天津大学博士学位. 2018 ~ 2019 年美国宾夕法尼亚州立大学访问学者. 主要研究方向为神经影像数据分析, 机器学习.
E-mail: wangrf@tju.edu.cn
(**WANG Ruo-Fan** Lecturer at the School of Information Technology Engineering, Tianjin University of Technology and Education. She received her Ph. D. degree from Tianjin University in 2015. From 2018 to 2019, she was a visiting scholar in the Department of Biomedical Engineering, The Pennsylvania State University. Her research interest covers analysis of neuroimaging data and machine learning.)