arXiv:2301.02422v1 [stat.ME] 6 Jan 2023

非参数多划分聚类

Marie du Roy de Chaumaray 和 Vincent Vandewalle

2023年1月9日

摘要

在基于模型的聚类框架下，研究者提出了一种称为多划分聚类的模型，该模型允许存在多个潜在类别变量。该模型假设观测数据的分布可分解为若干个相互独立的变量块，每个变量块都遵循自身的混合模型。本文假设每个变量块都遵循非参数潜在类别模型，$i.e.$ 即混合模型中各成分内的变量相互独立，且其类别条件分布不受任何参数化假设约束。本文旨在根据样本观测结果推断变量块的数量、变量在各块中的划分，以及每个变量块中成分的数量，这些特征共同构成所提出的模型。遵循近期关于非参数混合模型中模型与变量选择的研究，我们提出将数据离散化为分箱。这样便可采用经典的、针对参数化多项分布的多划分聚类方法，该方法基于惩罚似然法（$e.g.$ 贝叶斯信息准则）。本文证明了该方法的一致性，并提出了一种高效的优化方法。我们还在模拟数据上考察了该模型的性能。

1 引言

有限混合模型通过对变量的分布建模，并将混合模型的每个成分识别为一个簇，从而实现聚类（McLachlan 和 Peel，2000；McNicholas，2016；Bouveyron 等人，2019）。这使得人们能够通过针对各类数据调整类别条件模型，对多种数据（连续型、类别型、函数型和混合型数据）进行聚类。混合模型的大部分发展都基于参数框架，在该框架下可以进行一致的模型选择，例如选择簇的数量（Keribin，2000）。然而，假设模型为参数模型可能过于严格，难以涵盖各种簇形状。因此，人们开发了非参数模型（综述见 Chauveau 等人（2015））；这类模型通常只假设簇内变量满足类别条件独立性，而不对类别条件单变量分布作任何参数假设。随后，可以通过类似 EM 的算法（Benaglia 等人，2009）或最大化平滑对数似然（Levine 等人，2011）来进行估计。在这种情况下，簇数量的选择是一个难题，Du Roy de Chaumaray 和 Marbac（2021）最近提出了一种基于变量离散化的一致解决方案。

有限混合模型的一个局限是，它们试图仅用一个类别型变量来概括数据的异质性。然而，在海量数据领域，个体可能由数千个变量描述，数据中的全部异质性无法仅用一个潜变量来描述。例如，与某一关注主题相关的变量可能比与另一主题相关的变量更为常见。因此，人们开发了基于模型的聚类方法，以便处理多个潜在类别变量；我们将其称为多重划分聚类（关于该主题的最新综述，参见 Rodriguez-Sanchez 等人（2022））。在模型中引入多个潜变量主要有三种方式。第一种方式是假设分布可分解为相互独立的变量块，每个块中的异质性由一个潜在聚类变量解释（Galimberti 和 Soffritti，2007；Marbac 和 Vandewalle，2018）。第二种方式是假设变量存在多个用于分类的线性投影，每个投影由一个潜在簇变量解释（Attias，1999；Vandewalle，2020）。最后，也可以在更复杂的依赖结构中考虑多个潜变量，例如树（Poon 等人，2013）、贝叶斯网络（Rodriguez-Sanchez 等人，2022），或多层（可能是深层的）离散潜在结构（Gu 和 Dunson，2021）。与标准聚类相比，多重划分聚类会带来额外的复杂度，因为

需要学习大量结构参数（块的数量、变量在各块中的划分、每个潜变量的模态数量、网络结构）。由于上述多重划分模型属于参数框架，所有这些参数都可以通过惩罚似然准则（例如贝叶斯信息准则）或基于积分似然的准则（例如 MICL）进行选择。搜索难度取决于所假设模型的复杂度。例如，在连续情形下，Galimberti 等人（2018）提出了 Galimberti 和 Soffritti（2007）方法的扩展，其中需要考虑变量可能承担的多种角色，因此即便只重新分配一个变量，也需要大量计算。相反，Marbac 和 Vandewalle（2018）提出了一个非常简单的模型：变量被分组成相互独立的块，并假设每个变量块都服从具有类条件独立假设的参数化混合模型。他们据此提出了一种修正EM算法，能够在算法的每一步独立更新变量所属块的分配。值得注意的是，该模型扩展了 Marbac 和 Sedki（2017）以及 Marbac 等人（2018，2020）在聚类中的变量选择框架下提出的方法，其中仅考虑两个块，即一个假设条件独立的分类变量块，以及一个假设完全独立的非分类变量块。

本文提出了 Marbac 和 Vandewalle（2018）所提出方法的非参数扩展。为了解决模型选择的难题，我们首先对数据进行离散化，其中离散化的粒度取决于数据量，这与 Du Roy de Chaumaray 和 Marbac（2021）的方法相同。在这种情况下，非参数多重划分混合模型变为参数模型，其中每个块都服从多项分布乘积分布的混合。该模型是 Marbac 和 Vandewalle（2018）所提出模型的一个特例，因此我们能够通过优化惩罚似然，高效地进行模型搜索和参数估计。此外，遵循 Du Roy de Chaumaray 和 Marbac（2021）的思路，我们能够证明该方法的一致性。解决模型选择问题后，我们提出在每个变量块上重新拟合非参数潜在类别模型，并固定簇的数量，以限制变量离散化造成的信息损失，从而更准确地估计划分。

本文结构如下：第 2 节介绍非参数多重分区模型，第 3 节说明如何对数据进行离散化，第 4 节介绍模型的估计量及其一致性，并说明用于模型选择的修正EM算法。最后，第 5 节通过模拟数据展示我们的方法具有良好的性能。

2 多重分区混合模型

2.1 基础模型

待聚类的数据 $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ 由 n 个观测 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ 构成，这些观测由 d 个可能类型不同的变量描述（即，每个变量可以是连续型或类别型）。假设观测独立地来自一个多重分区模型（MPM），该模型认为变量被分组为 B 个独立区块。变量块由 $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ 定义，其中 $\omega_j = b$ 表示变量 j 属于块 b。属于同一块 b 的变量索引集合记为 $\Omega_b = \{j : \omega_j = b\}$。此外，MPM 认为块 b 中的变量服从一个具有 G_b 个成分的混合模型，并假设成分内独立性。该假设使得这些变量的条件密度可以写成 $|\Omega_b|$ =card(Ω_b) 个单变量密度的乘积。因此，模型 $\boldsymbol{m}$ 由块的数量 B、各块中变量的划分 $\boldsymbol{\omega}$，以及驱动每个块的混合模型的成分数 $\boldsymbol{G} = (G_1, \ldots, G_B)$ 决定。令 $\boldsymbol{x}_{i\{b\}} = (x_{ij}; j \in \Omega_b)$ 表示块 b 的观测变量向量。那么，对于模型 $\boldsymbol{m} = (B, \boldsymbol{G}, \boldsymbol{\omega})$，$\boldsymbol{x}_i$ 的概率密度函数（pdf）为

$$p(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} p_b(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}) \text{ with } p_b(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}) = \sum_{g=1}^{G_b} \pi_{bg} \prod_{j \in \Omega_b} \eta_{gj}(x_{ij}), \tag{1}$$

其中，$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\eta})$ 对模型参数进行分组，$\boldsymbol{\pi} = (\boldsymbol{\pi}_b; b = 1, \ldots, B)$ 是每个混合模型中各成分的比例，$\boldsymbol{\pi}_b = (\pi_{b1}, \ldots, \pi_{bG_b})$、$\pi_{bg} > 0$ 和 $\sum_{g=1}^{G_b} \pi_{bg} = 1$，以及 $\boldsymbol{\eta}$ 将单变量密度 η_{gj} 分组；对于连续型变量，这些密度是无限维参数。我们用 $\Theta_{\boldsymbol{m}}$ 表示与给定模型 $\boldsymbol{m}$ 相关的所有可能参数的集合 $\boldsymbol{\theta}$。我们用 p_0 表示真实模型 $\boldsymbol{m}_0$ 和真实参数 $\boldsymbol{\theta}_0$ 下的概率密度函数，即 $p_0 = p(.|\boldsymbol{m}_0, \boldsymbol{\theta}_0)$。

2.2 数据的所得划分

MPM 对观测进行 B 个划分（每个变量块对应一个划分）。块 b 的划分记为 $\mathbf{z}_b = (\boldsymbol{z}_{1b}, \ldots, \boldsymbol{z}_{nb}) \in \mathcal{Z}_{G_b}$，其中 $\mathcal{Z}_{G_b}$ 是将 n 个元素划分为 G_b 个簇的所有划分的集合；$\boldsymbol{z}_{ib} = (z_{ib1}, \ldots, z_{ibG_b})$，其中，当观测 i 属于块 b 的簇 g 时，$z_{ibg} = 1$ 取值为 1，否则为 0。因此，模型 $\boldsymbol{m}$ 的多个划分 $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_B)$ 属于 $\boldsymbol{\mathcal{Z}}_{\boldsymbol{m}} = \mathcal{Z}_{G_1} \times \ldots \times \mathcal{Z}_{G_B}$。

2.3 模型和参数的可辨识性

在本部分中，我们解释为什么模型（1）在成分标签的切换和块顺序的改变意义下是可辨识的。我们需要以下假设。

假设1。 *(i)* 每个块中至少有三个变量，并且混合模型中的所有比例均不为零，即 对于任意 $b = 1、\ldots B、\ |\Omega_b| \geq 3$，以及任意 $g = 1、\ldots、\ G_b、\ \pi_{bg} > 0$。

(ii) 对于每个 $b = 1、\ldots、\ B$，存在 $\Upsilon_b \subseteq \Omega_b$，使得 $|\Upsilon_b| = 3$，并且对于任意 $j \in \Upsilon_b$，单变量密度 $\{\eta_{gj};\ g = 1、\ \ldots、\ G_b\}$ 是线性无关的。

(iii) 对于每个 $b = 1、\ldots、\ B、\ p_b$ 不能分解为密度乘积。

假设 B 和 $\boldsymbol{\omega}$ 固定且已知，也就是说变量的块结构已给定，那么在假设1(i)和(ii)下，模型参数 $\boldsymbol{\theta}$ 和成分数量 $\boldsymbol{G}$ 是可辨识的（见Allman等（2009））。

我们将 B 定义为使 $p(\cdot|\boldsymbol{m},\ \boldsymbol{\theta})$ 能够分解为乘积 $\prod_{b=1}^{B} p_b(\cdot|\boldsymbol{m},\ \boldsymbol{\theta})$ 的最大整数。因此，假设1(iii)确保了 B 以及变量的相应划分 $\boldsymbol{\omega}$ 的可辨识性。

请注意，假设1(iii)意味着属于同一块的变量并非相互独立。

3 通过分箱估计进行模型选择

3.1 用于估计的离散化模型

若假设所有变量均为连续型，该方法基于将每个变量 j 离散化为 R 个互不重叠的箱 $I_{Rj1}, \ldots\ldots, I_{RjR}$，使得 $\cup_{r=1}^{R} I_{Rjr} = \mathcal{X}_j$，且对于任意 $(r,\ r')$，若 $r \neq r'$，则 $I_{Rjr} \cap I_{Rjr'} = \emptyset$。我们用 σ_{Rjr}（其中 $r \in \{1, \ldots, R\}$）表示每个箱的示性函数，其定义为：若 $x_{ij} \in I_{Rjr}$，则 $\sigma_{Rjr}(x_{ij}) = 1$；若 $x_{ij} \notin I_{Rjr}$，则 $\sigma_{Rjr}(x_{ij}) = 0$。我们用 $|I_{Rjr}|$ 表示箱 I_{Rjr} 的勒贝格测度。另一种做法是为不同变量使用不同数量的箱，但为便于阅读，以下不采用这种做法。箱的数量及其位置均已预先确定。下一节将给出确保该方法具有良好性质的条件，以及选择这些项目的方法。

块 b 的离散化变量遵循潜在类别模型，其中每个成分都是 $|\Omega_b|$ 个多项分布的乘积，且每个多项分布有 R 个水平。因此，主体 i 在块 b 中各变量的离散化概率密度函数为

$$p_{Rb}(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}_{Rb}) = \sum_{g=1}^{G_b} \pi_{bg} \prod_{j \in \Omega_b} \prod_{r=1}^{R} \left(\frac{\alpha_{Rgjr}}{|I_{Rjr}|} \right)^{\sigma_{Rjr}(x_{ij})}, \tag{2}$$

其中，$\boldsymbol{\theta}_{Rb} = (\boldsymbol{\pi}_b, \boldsymbol{\alpha}_{Rb})$ 将成分比例 π_{bg} 与概率 α_{Rgjr} 分组；该概率表示当变量 j 离散化为 R 个分箱时，来自成分 g 的一个个体取值为 r 的概率。除以 $|I_{Rjr}|$ 表示对类别条件单变量密度进行直方图近似。参数空间由单纯形的乘积 $S_{G_b} \times S^{G_b|\Omega_R}{}^{b|}$ 给出。注意，p_{Rb} 是 p_b 的近似，并且随着 R 趋于无穷大，该近似会更加准确。由此可推得个体 i 的概率密度函数（pdf）p 的离散化版本 p_R，其表达式为

$$p_R(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}_R) = \prod_{b=1}^{B} p_{Rb}(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}_{Rb}),$$

其中， p_{Rb} 由方程（2）给出， $\boldsymbol{\theta}_R$ 汇总了每个块 $\boldsymbol{\theta}_{R1},\ldots,\boldsymbol{\theta}_{RB}$ 的所有参数。真实密度 $p(.|\boldsymbol{m}_0,\boldsymbol{\theta}_0)$ 的离散化版本记作 $p_{0,R}$。注意，根据构造，离散化会导致参数 α_{Rgjr} 的可辨识性损失。

3.2 处理混合类型数据

该模型也适用于处理混合类型数据，其中一些变量是类别型的。实际上，我们只对连续型变量进行离散化。

4 基于惩罚似然的模型选择

竞争模型集合 $\mathcal{M}$ 定义如下

$$\mathcal{M}=\{\boldsymbol{m}=(B,\boldsymbol{G},\boldsymbol{\omega});B\leq B_{\max},\ \forall b,j\ G_b\leq G_{\max},\omega_j\in\{1,\ldots,B\},|\Omega_b|\geq 3\},\tag{3}$$

其中， $B_{\max}$ 是块的最大数量， $G_{\max}$ 是每个块中成分的最大数量。对于给定的模型 $\boldsymbol{m}$，参数 $\boldsymbol{\theta}_R$ （由所有 π_{bg} 和所有 α_{Rgjr} 构成）是未知的，必须根据观测进行估计。给定模型 $\boldsymbol{m}$ 的参数空间记为 $\Theta_{R,\ \boldsymbol{m}}$ ，它是以下乘积空间

$$\Theta_{R,\boldsymbol{m}}=\prod_{b=1}^{B}S_{G_b}\times S_R^{G_b|\Omega_b|},\tag{4}$$

其中， $S_K=\{\boldsymbol{u}\in[0,1]^K:\ \ \sum_{k=1}^K u_k=1\}$ 表示大小为 K 的单纯形。我们决定选择使离散化版本的惩罚对数似然达到最大值的模型，其中惩罚项 $a_{n,\ \boldsymbol{m},\ R}$ 经过了适当选择。例如，当 $a_{n,\ \boldsymbol{m},\ R}$ 等于 $\nu_{\boldsymbol{m}}\ln(n)/2$ 时，它对应于 BIC（Schwarz，1978年），其中 $\nu_{\boldsymbol{m}}$ 是模型 $\boldsymbol{m}$ 的复杂度。根据方程（4），

$$\nu_{\boldsymbol{m}}=\sum_{b=1}^{B}(G_b-1)+(R-1)G_b|\Omega_b|.$$

4.1 模型推断

对于样本 $\mathbf{x}$ 和模型 $\boldsymbol{m}$，观测数据对数似然定义为

$$\ell(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x})=\sum_{i=1}^{n}\sum_{b=1}^{B}\ln p_{Rb}(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m},\boldsymbol{\theta}_{Rb}),\tag{5}$$

从中减去惩罚项，即可得到其惩罚版本 ℓ_{pen}

$$\ell_{\text{pen}}(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x})=\ell(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x})-a_{n,\boldsymbol{m},R}\tag{6}$$

基于惩罚似然的模型选择，首先是关于 $\boldsymbol{\theta}_R$ 最大化 ℓ_{pen} ，然后再关于 $\boldsymbol{m}$最大化。因此，选出的模型为

$$\widehat{\boldsymbol{m}}_{n,R}=\underset{\boldsymbol{m}\in\mathcal{M}}{\arg\max}\ \underset{\boldsymbol{\theta}_R\in\Theta_{R,\boldsymbol{m}}}{\max}\ \ell_{pen}(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x}).\tag{7}$$

注意， (5)， 这意味着 得益于条件独立性，我们可以在对 b 求和之前，先对 i 求和 n 可以分别在每个块中进行最大化。

在实际应用中，为避免数值问题，我们引入一个阈值 ε ，使参数空间变为 $\Theta_{R,\boldsymbol{m},\varepsilon}=\prod_{b=1}^{B}S_{G_b,}$ $_{\varepsilon}\times S^{G_b|}{}_{\Omega R\,\varepsilon}{}^{b|}$.，其中 $\varepsilon>0$ 是单纯形中所有元素的最小值，即 $S_{K,\varepsilon}=\{\boldsymbol{u}\in]\varepsilon,1]^K:\ \ \sum_{k=1}^K u_k=1\}$。请注意，在这一框架下使用此类阈值十分常见（参见 Toussile 和 Gassiat（2009）； Bontemps 和 Toussile（2013）； Du Roy de Chaumaray 和 Marbac（2021））。在 $R\varepsilon$ 趋于零、 R 趋于无穷且 ε 趋于零的条件下，参数空间 $\Theta_{R,\boldsymbol{m},\varepsilon}$ 随着 ε 趋于零而收敛至整个参数空间 $\Theta_{R,\boldsymbol{m}}$ 。根据下一节将由假设 4(i)给出的关于 B 增长速率的假设，只需针对某个 $\alpha>0$设定 $\varepsilon^{-1}=O(n^{\alpha+1})$。

4.2 概率意义下的渐近收敛

Du Roy de Chaumaray 和 Marbac（2021）在四组假设下证明了估计量的一致性。为完整起见，我们在此回顾这些假设。假设1已在第 2.3 节给出，以确保基础模型的可辨识性。假设2规定了成分分布的约束。假设3给出了关于惩罚项的一些条件。最后，假设4给出了关于离散化的一些条件。

假设2。 *(i)* 存在某个函数 τ 属于 $L_1(p_0\nu)$，使得对于任意模型 $\boldsymbol{m} \in \mathcal{M}$ 和任意参数 $\boldsymbol{\theta} \in \Theta_{\boldsymbol{m}}$，

$$|\ln p(.|\boldsymbol{m}, \boldsymbol{\theta})| < \tau \ \ \nu\text{-}a.e..$$

(ii) 每个变量 j 都定义在一个紧空间 $\mathcal{X}_j$ 上，并且其各个分量 g的密度（记为 η_{gj}）除勒贝格测度为零的集合外均严格为正。

(iii) 存在某个正常数 $L < \infty$，对于任意块 b 和任意变量 $j \in \Omega_b$，它都能界定密度 η_{gj} 在 $\mathcal{X}_j$上的导数：

$$\forall x_j \in \mathcal{X}_j, \ \ |\eta'_{gj}(x_j)| \leq L.$$

假设3。 *(i)* 对于任意模型 $\boldsymbol{m}$， $a_{n,\ \boldsymbol{m},\ R}$是关于 R、 G_1、…、 G_B、 $|\Omega_1|$、…以及 $|\Omega_B|$的递增函数。

(ii) 对于任意模型 $\boldsymbol{m}$， $a_{n,\ \boldsymbol{m},\ R}/n$ 在 n 趋于无穷大时趋于 *0*。

(iii) 对于任意模型 $\boldsymbol{m}$, $R/a_{n,\ \boldsymbol{m},\ R}$ 在 n 趋于无穷时趋于 *0*。

(iv) 对于任意满足 $\boldsymbol{m} \subset \widetilde{\boldsymbol{m}}$ 的模型 $\boldsymbol{m}$ 和 $\widetilde{\boldsymbol{m}}$，当 n 趋于无穷时，$a_{n,\ \widetilde{\boldsymbol{m}},\ R}/a_{n,\ \boldsymbol{m},\ R}$ 的下极限严格大于 *1*。

假设4。 *(i)* 分箱数量 R 随 n 按以下方式趋于无穷：$\lim_{n\to\infty} R = \infty$ 且 $\lim_{n\to\infty} R(\ln^2 n)/n^{1/2} = 0$。

(ii) 每个 分箱的长度不为零，并且满足：对于任意变量 j 和任意分箱 r， $|I_{Rjr}|^{-1} = O(R \qquad)$.

(iii) 对于任意变量 j，令 $\mathcal{I}_{jR}$ 为 R 区间上界的集合，则对于任意 $x_j \in \mathcal{X}_j$，$d(x_j, \mathcal{I}_{jR})$ 在 R 趋于无穷时趋于零。

定理 1. 假设独立数据来自(1)，其真实模型为 $\boldsymbol{m}_0 = \{B_0, \boldsymbol{G}_0, \boldsymbol{\omega}_0\}$，且竞争模型集合 $\mathcal{M}$ 由(3)给出。此外，若假设 *1*、*2*、*3* 和 *4* 均成立，则由(7) 定义的估计量 $\widehat{\boldsymbol{m}}_{n,R}$ 在 n 趋于无穷时依概率收敛到 $\boldsymbol{m}_0$。

证明与 Du Roy de Chaumaray 和 Marbac（2021）中的定理 1 类似，因为我们注意到需要分别处理每个块，并且必须区分变量集合中哪些变量涉及所考虑的块，哪些变量不涉及。

4.3 用于模型选择的 EM 算法

为了计算 $\widehat{\boldsymbol{m}}_{n,R}$，我们需要同时对 $\boldsymbol{\theta}_R$ 和 $\boldsymbol{m}$最大化惩罚对数似然。我们将使用完全数据对数似然，它基于成分归属$\mathbf{z}$ 被观测到这一假设，因此定义为

$$\ell(\boldsymbol{\theta}_R|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) = \sum_{b=1}^{B} \ln p(\mathbf{z}_b|G_b, \boldsymbol{\pi}_b) + \sum_{b=1}^{B} \sum_{j\in\Omega_b} \ln p(\mathbf{x}_j|G_b, \mathbf{z}_b, \boldsymbol{\alpha}_{Rbj}), \tag{8}$$

其中 $\mathbf{x}_j$ =(x_{1j}，……，x_{nj})，$\boldsymbol{\alpha}_{Rbj} = \{\alpha_{Rgjr}, g = 1, \ldots, G_b, r = 1, \ldots, R\}$，

$$\ln p(\mathbf{z}_b|G_b, \boldsymbol{\pi}_b) = \sum_{i=1}^{n} \sum_{g=1}^{G_b} z_{ibg} \ln \pi_{bg},$$

and

$$\ln p(\mathbf{x}_j|G_b,\mathbf{z}_b,\boldsymbol{\alpha}_{Rj})=\sum_{i=1}^{n}\sum_{g=1}^{G_b}z_{ibg}\sum_{r=1}^{R}\sigma_{Rjr}(x_{ij})\ln\left(\frac{\alpha_{Rgjr}}{|I_{Rjr}|}\right).$$

最大似然估计（MLE）可通过EM算法获得（Dempster 等，1977；McLachlan 和 Krishnan，1997）。变量块之间的独立性使得我们能够分别在每个块上最大化观测数据对数似然。

由于候选模型数量众多，逐一计算每个候选模型的 BIC 这一穷举方法在实际中不可行。我们将采用 Marbac 和 Vandewalle (2018) 的相同思路来规避组合问题。固定 $(B,\ \boldsymbol{G})$ 后，使用 BIC 进行模型选择并进行最大似然推断，意味着要针对 $(\boldsymbol{\omega},\ \boldsymbol{\theta}_R)$ 最大化惩罚似然。利用经过修改的 EM算法版本（Green, 1990; Marbac 等，2018），可以同时完成这一最大化过程；随后，对 $(B,\ \boldsymbol{G})$ 的每个允许取值运行该算法，即可找到 $\widehat{\boldsymbol{m}}_{n,\ R}$。因此，EM算法的调用次数应少于 $\sum_{B=1}^{B_{\max}}G_{\max}^{B}$。请注意，EM算法的调用次数不取决于变量的数量；如果 $B_{\max}$ 较小（即 $B_{\max}<5$），那么计算负担并不重。这看起来可能有所限制，但请注意，经典聚类方法会考虑 $B_{\max}=1$。此外，如果希望 $B_{\max}$ 大于五，那么该模型仍然定义良好，但所提出的模型选择方法会遇到组合问题。因此，在这种情况下，应使用其他算法（如前向/后向搜索）进行模型估计。

为实　现这一修正EM算法，我们引入惩罚完全数据对数似然　　d

$$\ell_{pen}(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x},\mathbf{z})=\ell(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x},\mathbf{z})-a_{n,\boldsymbol{m},R}.$$

我们需要假设惩罚项可以分解为与各类参数相关的惩罚项之和，如下所示：$a_{n,\ \boldsymbol{m},\ R}=\sum_{b=1}^{B}\left(a_{n,\boldsymbol{\pi}_b,R}+\sum_{j\in}\right.$ $\left._{\Omega_b}a_{n,\ \boldsymbol{\alpha}_{Rbj},\ R}\right)$. 这一假设并不苛刻，例如，贝叶斯信息准则满足此假设，其中 $2a_{n,\ \boldsymbol{\pi}_b,\ R}=(G_b-1)\ln\ n$，且 $2a_{n,\ \boldsymbol{\alpha}_{Rbj},\ R}=(R-1)(G_b-1)\ln n$。因此，惩罚完全数据对数似然可改写为

$$\ell_{pen}(\boldsymbol{\theta}_R|\boldsymbol{m},\mathbf{x},\mathbf{z})=\sum_{b=1}^{B}\left(\ln p(\mathbf{z}_b|\boldsymbol{\pi}_b)-a_{n,\boldsymbol{\pi}_b,R}\right)+\sum_{b=1}^{B}\sum_{j\in\Omega_b}\left(\ln p(\mathbf{x}_j|\mathbf{z}_b,\boldsymbol{\alpha}_{Rbj})-a_{n,\boldsymbol{\alpha}_{Rbj},R}\right).\qquad(9)$$

固定 $(B,\ \boldsymbol{G})$，并从 $(\boldsymbol{\omega}^{[0]},\ \boldsymbol{\theta}_R^{[0]})$ 开始，算法的第 $[s]$ 次迭代由以下三个步骤组成：

E步 计算模糊划分 $t_{ibg}^{[s]}:=\mathbb{E}[Z_{ibg}|\boldsymbol{x}_i,\boldsymbol{m}^{[s-1]},\boldsymbol{\theta}_R^{[s-1]}]$，因此对于 $b=1,\ \ldots,\ B$，对于$g=1,\ \ldots,\ G_b$，对于 $i=1,\ \ldots,\ n$

$$t_{ibg}^{[s]}=\frac{\pi_{bg}^{[s-1]}\prod_{j\in\Omega_b^{[s-1]}}\left(\alpha_{Rgjr}^{[s-1]}\right)^{\sigma_{Rjr}(x_{ij})}}{\sum_{k=1}^{G_b}\pi_{bk}^{[s-1]}\prod_{j\in\Omega_b^{[s-1]}}\left(\alpha_{Rkjr}^{[s-1]}\right)^{\sigma_{Rjr}(x_{ij})}}.$$

M 步骤1 更新变量到块的分配，对于 $j=1,\ldots,d$,

$$\omega_j^{[s]}=\underset{b\in\{1,\ldots,B\}}{\arg\max}\left(\sum_{g=1}^{G_b}\max_{\boldsymbol{\alpha}_{Rgj}\in S_R}Q(\boldsymbol{\alpha}_{Rgj}|\mathbf{x}_j,\mathbf{t}_{bg}^{[s]})-a_{n,\boldsymbol{\alpha}_{Rbj},R}\right),$$

其中 $\boldsymbol{\alpha}_{Rgj}=(\alpha_{Rgj1},\ldots,\alpha_{RgjR})$ 和 $Q(\boldsymbol{\alpha}_{Rgj}|\mathbf{x}j,\mathbf{t})=\sum_{i=1}^{n}t_i\sum_{r=1}^{R}\sigma_{Rjr}(x_{ij})\ln\ \alpha_{Rgjr}$。因此 $\Omega_b^{[s]}=\{j:\omega_j^{[s]}=b\}$。

M 步骤2 更新模型参数：对于 $b=1,\ldots,B$，对于 $g=1,\ldots,G_b$,

$$\pi_{bg}^{[s]}=\frac{1}{n}\sum_{i=1}^{n}t_{ibg}^{[s]}$$

且对于 $j\in\Omega_b^{[s]}$，

$$\boldsymbol{\alpha}_{Rgj}^{[s]}=\underset{\boldsymbol{\alpha}_{Rgj}\in S_R}{\arg\max}\,Q(\boldsymbol{\alpha}_{jg}|\mathbf{x}_j,\mathbf{t}_{\omega_j^{[s]}g}^{[s]}),$$

这意味着，对于任意 $r = 1, \ldots, R$,

$$\alpha_{Rgjr}^{[s]} = \frac{\sum_{i=1}^{n} t_{i\omega_j^{[s]}g} \sigma_{Rjr}(x_{ij})}{\sum_{i=1}^{n} t_{i\omega_j^{[s]}g}}.$$

与标准 EM算法类似，目标函数 $\ell_{pen}(\boldsymbol{\theta}_R|\boldsymbol{m}, \mathbf{x}, \mathbf{z})$ 在每次迭代时都会增加，但通常无法达到全局最优解。因此，必须进行不同的随机初始化。最后请注意，该算法可能会返回空块。这是因为 M 步骤 1 执行时并未限制每个块至少包含一个变量。因此，每个 $\omega_j^{[s]}$ 都可以独立获得。

该算法可用于估计底层模型的密度。然而，已有研究普遍证明，基于分箱的密度估计器通常不如基于核的密度估计器。因此，我们建议仅使用所提出的方法来选择模型，然后针对选定的模型使用基于核的密度估计，例如类 EM 算法（Benaglia等人，2009）或 MM算法（Levine等人，2011）。数值实验将展示这一点。

5 数值实验

5.1 模拟设置

数据由一个包含三个变量块的混合模型生成（$B = 3$），每个块都是由三个成分构成的混合模型（$\boldsymbol{G} =$ (3,3,3) ），簇的比例均衡（即，比例相等 $\pi_{bg} = 1/3$），且涉及的变量数量相同（即， $|\Omega_1| = |\Omega_2| = |\Omega_3|$）。在块 b中，给定成分归属时， X_i 的边缘密度是 $|\Omega_b|$ 个单变量密度的乘积，其中

$$X_{ij} = \sum_{g=1}^{G_b} z_{ibg}\delta_{gj} + \xi_{ij}$$

其中所有 ξ_{ij} 均相互独立，且若 j 模 G_b 等于 g，则 $\delta_{gj} = \tau$；否则 $\delta_{gj} = \tau$ 为 $\delta_{gj} = 0$。对 τ 的取值进行调整，以达到选定的理论误分类率。

我们考虑 ξ_{ij} 的三种分布（标准高斯分布、自由度为 3 的 Student分布和拉普拉斯分布）、四种样本量（$n = 50$,100, 200, 400），以及每个块中不同数量的变量（每个块中 b， $|\Omega_b|$ 依次等于 6、9 和 12）；同时，设定 τ 的取值，使理论误分类率分别为 5% 和 10%。对于每种情形（样本量、变量数量、分布和误分类率），生成 100 个重复样本。

离散化步骤采用由经验分位数确定的 R 个分箱。对于每个先前的重复样本，我们考察四种分箱数： $R = [n^{1/4}]$, $[n^{1/5}]$, $[n^{1/6}]$ 和 $[n^{1/7}]$。

采用所提出的修正EM算法进行模型选择，设置为 $B_{\max} = 3$ 和 $G_{\max} = 3$，并使用BIC惩罚。随后，对选定的模型，使用mixtools软件包进行参数估计和密度估计，从而得到更准确的估计划分。

5.2 所提方法的性能。

在每种情形下，我们计算得到的变量块与真实变量块之间的调整兰德指数（ARI），以及得到的个体划分与真实划分在每个块中的调整兰德指数（ARI）。

图1展示了在每种情形下，基于100个重复样本得到的真实变量块与估计变量块之间调整兰德指数（ARI）的箱线图。误分类率固定为5%。结果表明，只要样本量和分箱数足够大，对于每类分布，所提方法都能完美恢复不同的变量块。我们提醒读者，分箱数相对于样本量不应过大，以满足假设4。随着 n 增大，该方法的准确性会提高；而当 n较小时，随着每个块中的变量数增加，方法性能会略有下降。这并不意外，因为这会使分类问题更加复杂。后续几幅图将进一步说明这一点。

图2展示了每个块中真实划分与估计划分之间调整兰德指数均值的箱线图。误分类率固定为 5%。由此也可以看出，该方法的表现良好。

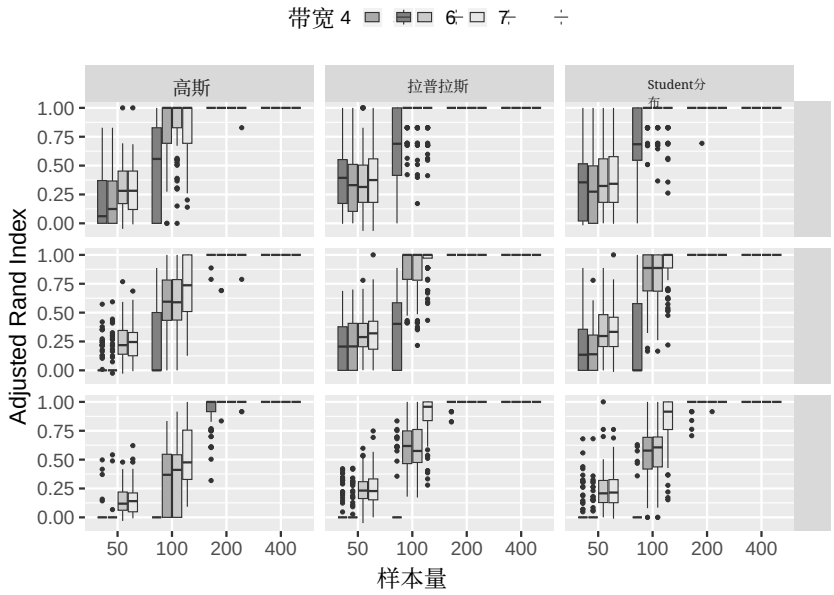


图1：不同样本量、不同成分族、不同变量数量以及不同分箱数量 $[n^{1/7}]$, $[n^{1/6}]$, $[n^{1/5}]$,$[n^{1/4}]$下真实变量块划分与估计变量块划分之间的调整兰德指数，灰度表示不同分箱数量。

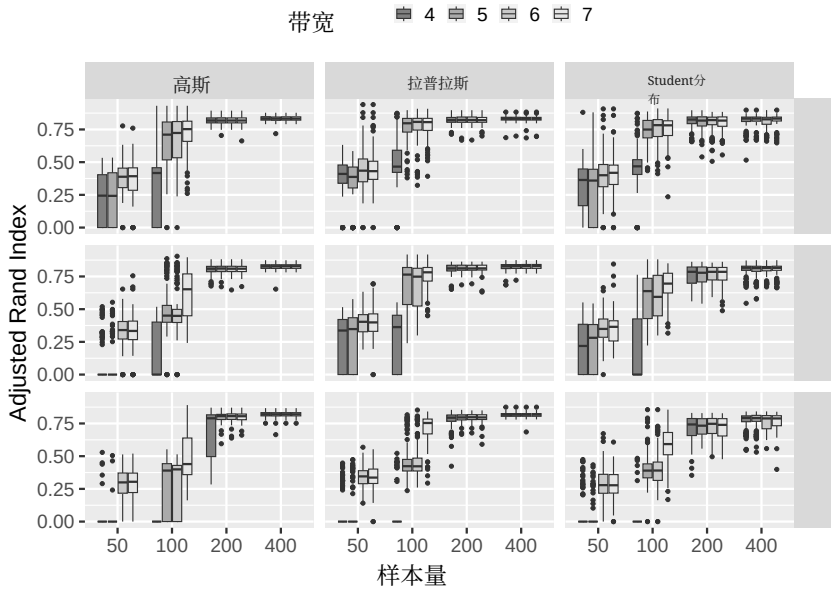


图2：不同样本量、不同成分族、不同变量数量以及不同分箱数量 $[n^{1/7}]$,$[n^{1/6}]$, $[n^{1/5}]$, $[n^{1/4}]$下真实个体划分与估计个体划分之间的平均调整兰德指数，灰度表示不同分箱数量。

在图 3 和图 4 中，我们研究了所提出方法在理论误分类率取不同值时的表现。分箱数量固定为 $[n^{1/6}]$。同样，我们注意到，随着 n 增大，该方法的准确性提高；而每个块中的变量数量增加时，方法的表现会变差。

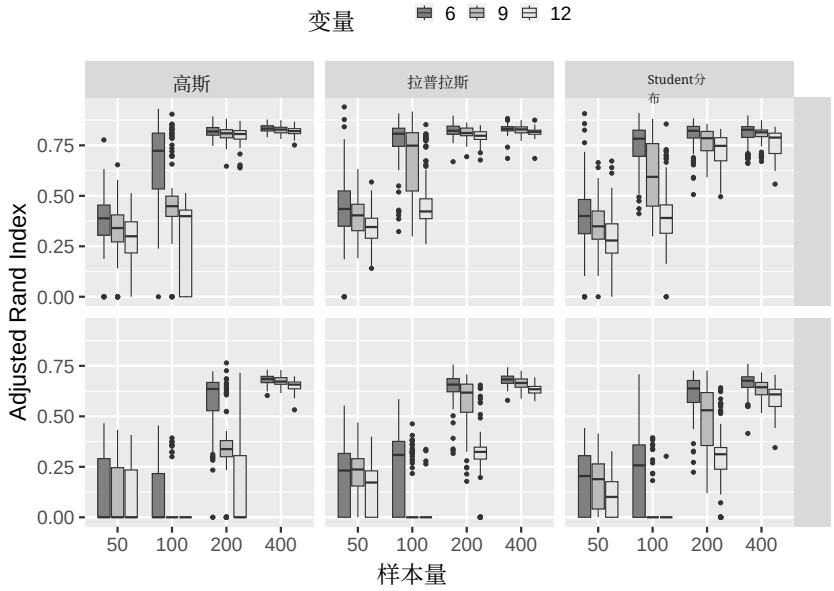


图3：不同样本量、不同成分族、不同误分类率以及不同变量数（以灰度表示）下，真实个体划分与估计个体划分之间的平均调整兰德指数。

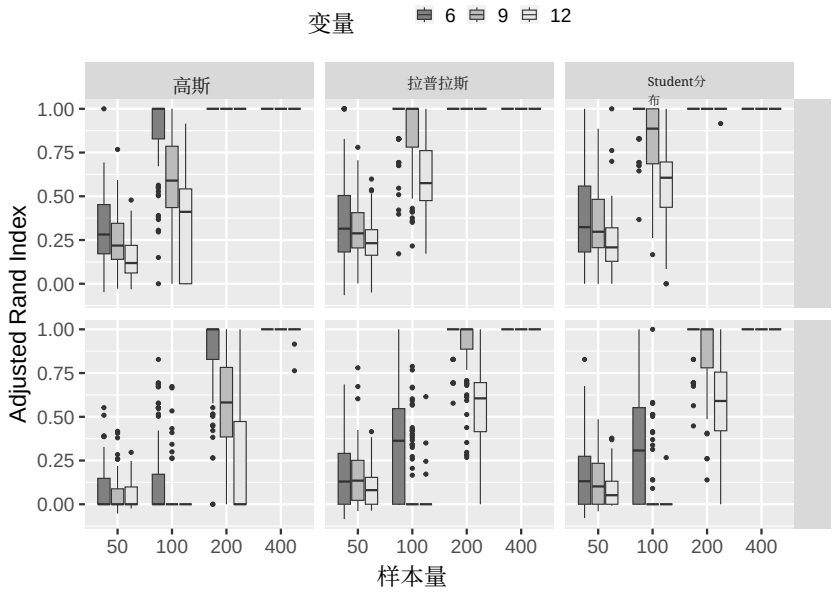


图4：不同样本量、不同成分族、不同误分类率以及不同变量数（以灰度表示）下，真实变量块划分与估计变量块划分之间的平均调整兰德指数。

6 结论

我们提出了一种新的方法，用于在不对给定成分归属时变量的条件分布作出参数假设的情况下，进行多个划分的聚类。该方法适用于连续型数据。它基于对连续型数据进行离散化，从而能够重用已有的、适用于多项分布的方法。对于混合类型数据，只需对其中的连续型数据进行离散化，即可直接扩展该方法。我们使用修正EM算法在离散化后的数据上进行模型选择，同时估计变量划分和模型参数。对于一系列惩罚项，该过程均具有一致性，其中包括经典的贝叶斯信息准则。

参考文献

参考文献

Allman, E.、Matias, C. 和 Rhodes, J.（2009）。具有许多观测变量的潜在结构模型中参数的可辨识性。《统计年刊》，37(6A):3099–3132。Attias, H.（1999）。独立因子分析。《神经计算》，11(4):803–851。Benaglia, T.、Chauveau, D. 和 Hunter, D. R.（2009）。用于多元混合模型半参数和非参数估计的类 EM 算法。《计算与图形统计学杂志》，18(2):505–526。Bontemps, D. 和 Toussile, W.（2013）。分类多元数据的聚类与变量选择。《电子统计学杂志》，7(none):2344–2371。Bouveyron, C.、Celeux, G.、Murphy, T. B. 和 Raftery, A. E.（2019）。基于模型的数据科学聚类与分类：*R*语言应用，第50卷。剑桥大学出版社。Chauveau, D.、Hunter, D. R. 和 Levine, M.（2015）。条件独立多元有限混合模型的半参数估计。《统计学综述》，9:1–31。Dempster, A.、Laird, N. 和 Rubin, D.（1977）。通过 EM 算法从不完整数据进行最大似然估计。《英国皇家统计学会会刊》。*B*辑（方法论），39(1):1–38。Du Roy de Chaumaray, M. 和 Marbac, M.（2021）。非参数多元有限混合模型的完整模型估计。Galimberti, G.、Manisi, A. 和 Soffritti, G.（2018）。建模变量在基于模型的聚类分析中的作用。《统计学与计算》，28(1):145–169。Galimberti, G. 和 Soffritti, G.（2007）。用于识别数据集中多个聚类结构的基于模型的方法。《计算统计学与数据分析》，52(1):520–536。Green, P. J.（1990）。关于将 EM 用于惩罚似然估计。《英国皇家统计学会会刊》。*B*辑（方法论），第443–452页。Gu, Y. 和 Dunson, D. B.（2021）。贝叶斯金字塔：适用于离散数据的可辨识多层离散潜在结构模型。*arXiv* 预印本 *arXiv:2101.10373*。Keribin, C.（2000）。混合模型阶数的一致估计。*Sankhy¯a*：《印度统计学杂志》*A*辑，第49–66页。Levine, M.、Hunter, D. R. 和 Chauveau, D.（2011）。多元混合模型的最大平滑似然。《生物统计学》，第403–416页。Marbac, M.、Patin, E. 和 Sedki, M.（2018）。混合数据聚类的变量选择：一种基于模型的方法。《分类学杂志》，待刊。Marbac, M. 和 Sedki, M.（2017）。使用积分完整数据似然进行基于模型的聚类中的变量选择。《统计学与计算》，27(4):1049–1063。

Marbac, M.、Sedki, M. 和 Patin, T.（2020）。混合数据聚类的变量选择：在人类群体基因组学中的应用。《分类学杂志》，37(1):124–142。Marbac, M. 和 Vandewalle, V.（2018）。一种易处理的多分区聚类。《计算统计与数据分析》。McLachlan, G. 和 Krishnan, T.（1997）。*EM*算法。Wiley概率与统计丛书：应用概率与统计，Wiley-Interscience，纽约。McLachlan, G. 和 Peel, D.（2000）。有限混合模型。Wiley概率与统计丛书：应用概率与统计，Wiley-Interscience，纽约。McNicholas, P. D.（2016）。基于混合模型的分类。CRC出版社。Poon, L. K.、Zhang, N. L.、Liu, T. 和 Liu, A. H.（2013）。高维数据的基于模型的聚类：变量选择与方面确定。国际近似推理杂志，54(1):196–215。Rodriguez-Sanchez, F.、Bielza, C. 和 Larra˜naga, P.（2022）。利用贝叶斯网络对混合数据进行多分区聚类。国际智能系统杂志，37(3):2188–2218。Schwarz, G.（1978年）。模型维数的估计。《统计年刊》，6(2):461–464。Toussile, W. 和 Gassiat, E.（2009）。使用多位点基因型数据进行基于模型的聚类中的变量选择。数据分析与分类进展，3(2):109–134。Vandewalle, V.（2020）。多分区子空间聚类。数学，8(4):597。