arXiv:2303.00318v1 [stat.ME] 1 Mar 2023

贝叶斯结果导向多视图混合模型及其在分子精准医学中的应用

Paul D. W. Kirk、Filippo Pagani 和 Sylvia Richardson，英国剑桥 MRC生物统计学组

2023年3月2日

摘要

聚类通常作为初步分析步骤，用于揭示组学数据集中的结构，例如发现疾病的分子亚型。这些数据集具有高通量、高维的特性，因此能够提供有关各种不同生物分子过程和通路的信息。不同的变量组（例如基因或蛋白质）会参与不同的生物分子过程，因此，仅对整个数据集进行分析以识别单个聚类划分，可能会将这些不同过程所产生的多个 聚类结构混为一谈。为解决这一问题，我们提出一种多视图贝叶斯混合模型，用以识别变量组（“视图”），每组变量都定义一种独特的聚类结构。我们考虑分层医学中的应用，其主要目标是识别患者簇，从而界定具有临床干预价值的不同疾病亚型。我们采用半监督、结果导向的贝叶斯轮廓回归混合建模方法，利用响应变量引导推断，使其聚焦于在分层医学背景下最相关的聚类。我们介绍该模型，并给出说明性模拟示例和泛癌蛋白质组学示例。我们展示如何运用该方法进行整合聚类，并探讨一个整合不同组学数据集以进行乳腺癌分型的示例。

1 引言

聚类被广泛用于组学数据分析，以揭示这些大型高维数据集中的结构和模式（例如 Eisen 等，1998；Heyer 等，1999；Alon 等，1999；Ben-Dor 等，1999；Son 等，2005）。我们尤其关注分子精准医学的应用，其分析目标是识别疾病的分子亚型（例如 Golub 等，1999；Perou 等，2000；Sørlie 等，2001；癌症基因组图谱网络，2012；Kuijjer 等，2018）。在这一背景下，目标是依据多种分子变量识别患者簇，而这些变量的测量值是通过高通量组学技术获得的。这些数据集的一个特点是通常具有高维性，因此往往需要采用变量筛选或选择策略（例如 Witten 和 Tibshirani，2010；Fop 和 Murphy，2018；Crook 等，2018），或其他降维技术（例如 Yeung 和 Ruzzo，2001；McLachlan 等，2002；Taschler 等，2019）。然而，一个常被忽视的挑战是，由于不同变量子集（例如对应于基因或蛋白质功能组的变量）参与多种不同的生物分子过程，组学数据集往往会在患者群体中定义多个 聚类结构。因此，根据所考虑的变量子集，我们可能识别出不同的患者簇。

已有不少论文提出了识别多重聚类结构的方法（例如 Cui 等人，2007年；Niu 等人，2010年；Guan 等人，2010；Li 和 Shafto，2011；Niu 等人，2014）。在相关文献中，定义同一聚类结构的一组变量被称为视图，而识别视图及其对应聚类结构的任务则被称为多视图聚类 （Cui 等人，2007年）或交叉聚类 （Li 和 Shafto，2011）。这些方法面临的一个潜在挑战是，如何判断已识别出的聚类结构中，哪一种对给定任务最有用或最相关。这在分层医学和疾病分型应用中尤为重要，因为我们希望找到能够界定患者群体的簇（分层）；例如，这些患者具有相似的预后，或对治疗有相似的反应（例如 Perou 等人，2000；Sørlie 等人，2001）。在实践中，要判断某个聚类结构是否“相关”，通常会利用一个留出的结局变量 y （例如生存数据），并评估不同簇是否与不同分布相关联 y （例如 Curtis 等人，2012）。另一种方法是采用半监督（结局引导）方法，在进行聚类分析时利用

y。

贝叶斯轮廓回归是这类半监督混合建模方法之一，它利用结局/响应变量引导推断，使其指向相关的聚类结构（Moli- tor 等，2010）。通俗地说，对于给定的响应变量，如果被分配到同一簇的个体往往具有相似的响应值，或者更一般地说，如果同一簇中个体的响应能够用一个共同模型准确描述，那么这种聚类就被称为相关聚类。更准确地说，如果个体响应的取值与其簇分配并非相互独立，那么对于给定的响应变量，该聚类就被定义为相关聚类。从这个定义可以清楚地看出，聚类是否相关只能相对于给定的响应变量来界定；尤其是对于不同的响应变量，相关的聚类可能各不相同。

例如，我们可以使用贝叶斯轮廓回归，根据遗传或基因组学数据对患者进行回顾性聚类，并以其生存时间作为响应变量，引导聚类识别具有预后意义的疾病亚型。如果改用不同的响应变量（例如身高），最终得到的聚类结构可能会截然不同。这两种聚类结构都未必是“错误”的——它们只是分别相对于不同的响应变量具有相关性。关键在于，如果采用无监督聚类方法（这通常是默认的分析选择），就无法保证它能识别出对任一响应变量都相关的聚类结构。

正如我们在第 5.1 节中所展示的，贝叶斯剖面回归模型的一个局限是，随着数据维度增加，响应变量（通常维度较低）对聚类结构推断的影响会减弱。在此，我们提出一种半监督多视图贝叶斯聚类模型，同时解决这一局限，以及现有多视图方法所面临的挑战：如何识别出（最）相关的聚类结构。

2 剖面回归

我们假设拥有由聚类变量向量 $\mathbf{x}$ 和响应变量 $\mathbf{y}$ 的观测值组成的数据。我们将聚类变量与响应变量的连接向量记为$\mathbf{v} = [\mathbf{x}; \mathbf{y}],$，并使用包含 K 个成分的混合模型（其中 K 可以是

有限的或无限的），如下所示：

$$p(\mathbf{v}|\boldsymbol{\rho},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{v}}(\mathbf{v}|\boldsymbol{\rho}_k) \tag{1}$$

$$= \sum_{k=1}^{K} \pi_k f_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\phi}_k), \tag{2}$$

其中，π_k 是与第 k个成分相关的混合权重，$\boldsymbol{\rho}_k$ 表示与第 k个成分相关的参数，我们用 $\boldsymbol{\rho}$ 和 $\boldsymbol{\pi}$ 分别表示 $\{\boldsymbol{\rho}_k\}_{k=1}^{K}$ 和 $\{\pi_k\}_{k=1}^{K}$。假设联合密度 $f_{\mathbf{v}}$ 可分解为 $f_{\mathbf{x}}$ 和 $f_{\mathbf{y}}$，如公式（2）所示，其中 $\boldsymbol{\phi}_k$ 表示**x**模型的参数，$\boldsymbol{\theta}_k$ 表示**y**模型的参数。我们用 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Theta}$ 分别表示 $\{\boldsymbol{\phi}_k\}_{k=1}^{K}$ 和 $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$。在剖面回归模型中，进一步假设 $f_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}_k,\mathbf{x}) = f_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}_k)$；即给定 $\boldsymbol{\theta}_k$ 后，**y** 与**x** 条件独立（Molitor 等，2010）。Shahbaba 和 Neal（2009）给出了一种相关模型，该模型不作此条件独立假设。

Molitor 等人（2010）允许模型 **y** 依赖于额外的调整协变量 **w**，这些协变量可能有助于预测响应变量，但我们不希望它们参与聚类（例如，我们希望加以控制的混杂因素，如年龄或性别），同时还包括相关的“全局”参数（即非特定于成分的参数）$\boldsymbol{\beta}$（详见附录）。一般的剖面回归模型如下：

$$p(\mathbf{x},\mathbf{y}|\boldsymbol{\Phi},\boldsymbol{\Theta},\boldsymbol{\pi},\boldsymbol{\beta},\mathbf{w}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}_k,\mathbf{w},\boldsymbol{\beta}) f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\phi}_k). \tag{3}$$

剖面回归模型的原始表述——我们在此也采用该表述——具体是使用狄利克雷过程先验的无限混合模型（Molitor 等，2010）；不过，我们注意到，该模型同样适用于有限 K的情形。

2.1 狄利克雷过程表述

在 Molitor 等（2010年）中，考虑了贝叶斯非参数情形的有限近似，采用狄利克雷过程的截断棒折断构造（Ishwaran 和 James，2001年）来定义公式（3）中 π_k 的先验分布。后续的贝叶斯轮廓回归论文（Hastie 等，2014）和实现方法（Liverani 等，2015）也考虑了棒折断

构造，并通过切片抽样（Walker，2007年；Kalli 等，2011）进行推断。遵循 Neal（2000）和 Rasmussen（2000）在无监督情形下的推导，本文则将狄利克雷过程混合模型视为（有限）K 成分混合模型的极限情形，其中，对混合权重 π_k 施加参数为 α/K 的对称狄利克雷先验（另见 Ishwaran 和 Zarepour，2002）。这与狄利克雷过程的 P´olya 罐（Blackwell 和 MacQueen，1973）及中国餐馆过程（Aldous 等，1985）构造密切相关，并可通过折叠吉布斯采样器进行推断，其中混合权重被边缘化。详细内容见附录，采样按 Neal（2000）所述进行——但需指出，在此背景下也可采用其他贝叶斯混合模型推断方法（例如 Richardson 和 Green，1997；Green 和 Richardson，2001年；Jain 和 Neal，2004、2007年；Walker，2007年；Kalli 等，2011；Miller 和 Harrison，2017）

简要概述如下，令 $D = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}_i)\}_{i=1}^n$ 表示一个数据集，其中包含 n 个体的 ($\mathbf{x}$、$\mathbf{y}$、$\mathbf{w}$) 三元组，使得 ($\mathbf{x}_i$，$\mathbf{y}_i$，$\mathbf{w}_i$) 对应第 i 个体。与混合模型中的常见做法一样，我们引入潜在成分分配变量 z_i；若第 i 个体与第 k 个成分相关联，则 $z_i = k$，而 $p(z_i = k|\boldsymbol{\pi}) = \pi_k$。我们将所有 n 成分分配构成的多重集记为 $\mathbf{z} = \{z_1, \ldots, z_n\}$。与第 i 个体相关联的成分条件似然为：

$$p(\mathbf{x}_i, \mathbf{y}_i|z_i = k, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{w}_i) = f_{\mathbf{y}}(\mathbf{y}_i|\boldsymbol{\theta}_k, \mathbf{w}_i, \boldsymbol{\beta}) f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_k). \tag{4}$$

对于成分特定参数和全局参数，分别采用独立先验 $p(\boldsymbol{\theta})$ $p(\boldsymbol{\phi})$ 和 $p(\boldsymbol{\beta})$。随后，折叠吉布斯采样器在以下步骤之间迭代：

- 步骤 (i)：在给定数据 D、最近一次采样得到的参数 $\boldsymbol{\Theta}$，$\boldsymbol{\Phi}$，$\boldsymbol{\beta}$，以及其他分配变量最近一次采样得到的值 $\mathbf{z}_{-i} = \mathbf{z}\backslash\{z_i\}$ 的条件下，更新成分分配 z_i；以及

- 步骤 (ii)：更新参数 $\boldsymbol{\Theta}$，$\boldsymbol{\Phi}$，$\boldsymbol{\beta}$，给定数据 D 以及最近采样得到的成分分配 $\mathbf{z}$。

在有限 K 情形下，若对混合权重取参数为 α/K 的对称狄利克雷先验，$\pi_1, \ldots, \pi_K \sim \text{Dir}(\alpha/K)$，则将个体分配到

indi i 分配给第 k个成分的条件后验概率——这是上述步骤 (i) 所需的——为：

$$p(z_i = k|D,\mathbf{z}_{-i}, \mathbf{\Phi}, \mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{w}_i, \alpha) = b\frac{n_{-i,k}}{n-1+\alpha} f_{\mathbf{y}}(\mathbf{y}_i|\boldsymbol{\theta}_k, \mathbf{w}_i, \boldsymbol{\beta}) f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_k), \tag{5}$$

其中， b是一个归一化常数，用于确保 $\sum_{k=1}^{K} p(z_i = k|D,\ \mathbf{z}_{-i},\ .....) = 1$；$n_{-i,\ k}$是当前分配给成分 k的个体数，不包括第 i个个体。也就是说，若令 $\mathbb{1}_a$ 表示指示函数（当 a 为真时等于 1，否则为 0），则$n_{-i,\ k} = \sum_{z_j \in \mathbf{z}_{-i}} \mathbb{1}_{z_j = k}$。

狄利克雷过程（DP）混合模型可通过考虑极限 $K \to \infty$推导得出；在此情形下，将个体 i 分配到现有成分（即当前已有其他个体分配至其中的成分）的条件后验概率为：

$$p(z_i = k,\ \text{where } k \in \mathbf{z}_{-i}|D,\mathbf{z}_{-i}, \mathbf{\Phi}, \mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{w}_i, \alpha) = b\frac{n_{-i,k}}{n-1+\alpha} f_{\mathbf{y}}(\mathbf{y}_i|\boldsymbol{\theta}_k, \mathbf{w}_i, \boldsymbol{\beta}) f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_k), \tag{6}$$

其中，$n_{-i,\ k}$ 与前文相同，且 b 同样是一个归一化常数（见下文公式 (8)）。对于 DP，我们还需要将个体 i 分配到新成分的条件后验概率，其表达式为：

$$p(z_i \notin \mathbf{z}_{-i}|D,\mathbf{z}_{-i}, \mathbf{\Phi}, \mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{w}_i, \alpha) = b\frac{\alpha}{n-1+\alpha} \int_{\boldsymbol{\phi},\boldsymbol{\theta}} f_{\mathbf{y}}(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{w}_i, \boldsymbol{\beta}) f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}) p(\boldsymbol{\phi}) p(\boldsymbol{\theta}) d\boldsymbol{\phi} d\boldsymbol{\theta}. \tag{7}$$

公式 (6) 和 (7) 中的归一化常数 b 用于确保：

$$p(z_i \notin \mathbf{z}_{-i}|D, \mathbf{z}_{-i}, \ldots) + \sum_{k \in \mathbf{z}_{-i}} p(z_i = k,\ \text{where } k \in \mathbf{z}_{-i}|D, \mathbf{z}_{-i}, \ldots) = 1. \tag{8}$$

给定成分分配变量 $\mathbf{z}$ 的值后，执行步骤 (ii) 相对简单；也就是说，对参数 $\mathbf{\Theta}, \mathbf{\Phi}, \boldsymbol{\beta}$ 进行抽样。完整细节见附录，其中还介绍了如何根据 Escobar 和 West（1995）所述的方法，对 DP 超参数 α 进行抽样。

3 剖面回归中的变量选择

我们在第 4 节提出的多视图模型可视为聚类变量选择的一种扩展形式。因此，我们首先介绍一种剖面回归模型，该模型允许以与 Law 等（2003，2004）、Tadesse 等（2005）和 Papathomas 等（2012）提出的模型相同的方式进行变量选择。

我们再次考虑与第 i 个个体相关的成分条件似然，并额外假设聚类变量条件独立：

$$f(\mathbf{x}_i, \mathbf{y}_i | z_i = k, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{w}_i) = f_{\mathbf{y}}(\mathbf{y}_i | \boldsymbol{\theta}_k, \boldsymbol{\beta}, \mathbf{w}_i) f_{\mathbf{x}}(\mathbf{x}_i | \boldsymbol{\phi}_k)$$
$$= f_{\mathbf{y}}(\mathbf{y}_i | \boldsymbol{\theta}_k, \boldsymbol{\beta}, \mathbf{w}_i) \prod_{j=1}^{P} f_x(x_{ij} | \boldsymbol{\phi}_k), \quad (9)$$

其中， P是$\mathbf{x}$的维度，因此$\mathbf{x}_i = [x_{i1}, \ldots, x_{iP}]^\top$。下文中，我们始终假设所有聚类变量具有相同的参数形式 f_x；不过，我们也可以直接扩展模型，为不同的聚类变量设定不同的参数形式，从而例如能够对混合型（连续型和类别型）数据进行建模。

为进行变量选择，我们采用 Law 等（2003，2004）和 Tadesse 等（2005）中的相同方法，并引入潜在二元指示变量 γ_j，使得若第 j个聚类变量对聚类结构有贡献，则 $\gamma_j = 1$ ，否则为 0。我们将满足 $\gamma_j = 0$ 的聚类变量建模为具有密度 $q_x(x_j | \boldsymbol{\phi}_{0j})$，其中 $\boldsymbol{\phi}_{0j}$ 是一个“全局”（即非成分特异性）参数，而 q_x 不必与 f_x具有相同的参数形式。我们用 $\boldsymbol{\phi}_{0\cdot} = \{\boldsymbol{\phi}_{0j}\}_{j=1}^{P}$表示所有 $\boldsymbol{\phi}_{0j}$ 参数的集合。分配条件似然为：

$$f(\mathbf{x}_i, y_i | z_i = k, \boldsymbol{\phi}, \boldsymbol{\phi}_{0\cdot}, \boldsymbol{\theta}, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f_{\mathbf{y}}(\mathbf{y}_i | \boldsymbol{\theta}_k, \boldsymbol{\beta}, \mathbf{w}_i) f_{\mathbf{x}}(\mathbf{x}_i | \boldsymbol{\phi}_k)$$
$$= f_{\mathbf{y}}(\mathbf{y}_i | \boldsymbol{\theta}_k, \boldsymbol{\beta}, \mathbf{w}_i) \prod_{j=1}^{P} q_x(x_{ij} | \boldsymbol{\phi}_{0j})^{\mathbb{1}_{\gamma_j = 0}} f_x(x_{ij} | \boldsymbol{\phi}_k)^{\mathbb{1}_{\gamma_j = 1}}. \quad (10)$$

公式（10）中引入潜在指示变量后，聚类变量被划分为两个互不相交的集合。沿用 Cui 等人（2007年）的术语，我们将这些集合称为视图。我们将视图 1 定义为构成聚类结构的聚类变量集合（其中$\gamma_j = 1$），将视图 0 定义为不构成聚类结构的变量集合（其中 $\gamma_j = 0$）。我们将后一组聚类变量称为空视图。因此，变量选择模型不仅可以将个体归为不同的簇，还可以将聚类变量归为不同组（共 2 个视图）。视图分配变量 γ_j 的推断可通过吉布斯抽样进行，即根据给定第 j个聚类变量 $\{x_{ij}\}_{i=1}^{n}$ 的观测数据时各模型（即“无聚类结构”模型 q_x 和“有聚类结构”模型 f_x）对应的后验概率进行采样。详情参见附录。

4 多视图贝叶斯轮廓回归

变量选择模型的一个自然扩展是允许存在 L 个视图（此时 $L \in \mathbb{Z}^+$ 可以大于 2），方法是允许指示变量 γ_j 为具有 L 个类别的类别型变量：0，1，2，……，$L-1$。由此可见，γ_j 变量是视图分配变量，其作用类似于成分分配变量 z_i，但它们是将聚类变量而非个体归为一组。

与变量选择模型一样，引入多个视图也需要引入视图特定模型。在变量选择的情形下，视图 1 中的聚类变量共同构成一个混合模型（用于定义个体之间的簇），而空视图中的聚类变量则对应单个密度（表示所有个体都属于同一个簇）。一般而言，视图特定模型可以有不同的选择。本文考虑一个 L 视图的情形，其中视图 0 为空视图，而视图 1、……、$L-1$ 各自关联一个混合模型，使得每个视图都定义个体之间一种不同的聚类结构。我们仅将响应变量与视图 1 相关联，并将该视图称为相关视图。因此，视图 1 的模型是一个（半监督）贝叶斯轮廓回归模型，而视图 2、……、$L-1$ 的模型均为无监督混合模型。

4.1 分配条件似然

对于每个非空视图，我们引入成分分配变量 $z_{i\ell}$，其中 $z_{i\ell}$表示在第 ℓ 个视图的混合模型中，负责第 i 个个体的成分。我们用 K_ℓ 表示第 ℓ 个视图的混合模型中的混合成分数量，其中——与前文相同——K_ℓ 可以是有限的，也可以是无限的。此外，我们用 $\boldsymbol{\phi}_{\ell k}$ 表示第 ℓ 个视图中与第 k 个成分相关联的参数，并将 $\boldsymbol{\phi}_{\ell\cdot} = \{\boldsymbol{\phi}_{k\ell}\}_{k=1}^{K_\ell}$ 定义为与第 ℓ 个视图相关联的全部分量特定参数。模型中与 $\mathbf{y}$ 相关的 $\boldsymbol{\theta}$ 参数仅与视图 $\ell = 1$ 相关联。

似然相关　　在给定这些成分分配的条件下，与第 i 个个体相关联　　，

于是得到:

$$f(\mathbf{x}_i, y_i|\{z_{i\ell}=k_\ell, \boldsymbol{\phi}_{\ell\cdot}\}_{\ell=1}^{L-1}, \boldsymbol{\phi}_{0\cdot}, \boldsymbol{\theta}, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f_{\mathbf{y}}(\mathbf{y}_i|\boldsymbol{\theta}_{k_1}, \boldsymbol{\beta}, \mathbf{w}_i) \prod_{j=1}^{P} q_x(x_{ij}|\boldsymbol{\phi}_{0j})^{\mathbb{1}_{\gamma_j=0}} \prod_{\ell=1}^{L-1} f_x(x_{ij}|\boldsymbol{\phi}_{\ell k_\ell})^{\mathbb{1}_{\gamma_j=\ell}}. \quad (11)$$

定义 $\Gamma_\ell = \{j:$ 令 $\gamma_j = \ell\}$ 为分配到第 ℓ 个视图的聚类变量的索引集，则上述内容也可改写如下:

$$f(\mathbf{x}_i, y_i|\{z_{i\ell}=k_\ell, \boldsymbol{\phi}_{\ell\cdot}\}_{\ell=1}^{L-1}, \boldsymbol{\phi}_{0\cdot}, \boldsymbol{\theta}, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \underbrace{\left(\prod_{j\in\Gamma_0} q_x(x_{ij}|\boldsymbol{\phi}_{0j})\right)}_{\text{View 0}} \underbrace{\left(f_{\mathbf{y}}(\mathbf{y}_i|\boldsymbol{\theta}_{k_1}, \boldsymbol{\beta}, \mathbf{w}_i) \prod_{j\in\Gamma_1} f_x(x_{ij}|\boldsymbol{\phi}_{1k_1})\right)}_{\text{View 1}} \prod_{\ell=2}^{L-1} \underbrace{\left(\prod_{j\in\Gamma_\ell} f_x(x_{ij}|\boldsymbol{\phi}_{\ell k_\ell})\right)}_{\text{View } \ell}, \quad (12)$$

其中，每个方括号中的项分别对应一个视图，如图所示。该表达式清楚地表明，在给定聚类变量到各视图的分配后，各视图的建模彼此独立，这一点也如图1所示：视图 0 被建模为单个聚类，视图 1 使用剖面回归模型进行建模（参见公式(9)），其余视图则分别使用各自的（无监督）混合模型进行建模。

4.2 推断

给定视图分配后，描述各视图的模型彼此独立，因此我们可以使用现有方法，在每个视图内对成分分配和参数进行推断：对于视图 1，采用贝叶斯剖面回归模型的方法（Molitor 等，2010）；对于视图 2、...、 $L-1$，则采用无监督贝叶斯混合模型的方法（例如 Richardson 和 Green，1997；Neal，2000）。因此，我们采用吉布斯抽样方法，在视图分配变量的采样与各视图模型的推断之间迭代。下面将详细介绍视图分配指示变量的更新。视图内参数和潜变量的更新按照 Neal（2000）的方法进行；更多细节请参见附录。

4.2.1 更新视图分配指标

定义 ν_ℓ 为聚类变量被分配至第 ℓ 个视图的先验概率。给定第 ℓ 个视图的成分分配 $z_{\cdot\ell} = \{z_{i\ell}\}_{i=1}^n$（对于 $\ell = 1$、.....、 $L-1$），后验

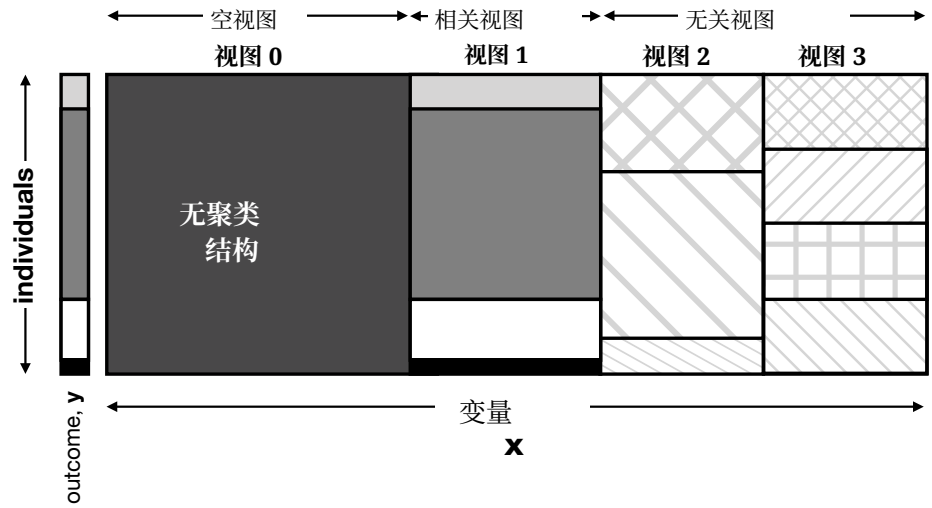


图1：半监督多视图模型的示意图。在给定聚类变量向各视图的分配后，每个视图均独立建模：视图 0（空视图）被建模为不具有聚类结构；视图 1 则采用剖面回归（半监督聚类）模型，将聚类变量所定义的聚类结构与响应变量 $\mathbf{y}$ 联系起来；其余视图各自采用无监督混合模型进行建模。上图中，数据矩阵左侧的一列表示响应变量 $\mathbf{y}$ 的水平。

概率 概率为第 j 个聚类变量被分配至第 ℓ 个视图的概率，即 :

$$p(\gamma_j = \ell | z_{\cdot\ell}, \boldsymbol{\phi}_{\ell\cdot}) = \frac{1}{Z}\nu_\ell \prod_{i=1}^{n} f_x(x_{ij}|\boldsymbol{\phi}_{\ell z_{i\ell}}) \text{ for } \ell = 1, \ldots, L-1, \tag{13}$$

其中， Z 是确保后验视图分配概率之和为 1 的归一化常数；与之前一样， $\boldsymbol{\phi}_{\ell z_{i\ell}}$ 表示与第 $z_{i\ell}$ 个视图中的第 ℓ 个成分相关联的参数，而 $\boldsymbol{\phi}_{\ell\cdot}$ 是与第 ℓ 个视图相关联的完整分量特定参数集合。

对于空视图，我们有：

$$p(\gamma_j = 0 | \boldsymbol{\phi}_{0j}) = \frac{1}{Z}\nu_0 \prod_{i=1}^{n} q_x(x_{ij}|\boldsymbol{\phi}_{0j}), \tag{14}$$

而现在已经很清楚，

$$Z = \nu_0 \prod_{i=1}^{n} q_x(x_{ij}|\boldsymbol{\phi}_{0j}) + \sum_{\ell=1}^{L-1} \left(\nu_\ell \prod_{i=1}^{n} f_x(x_{ij}|\boldsymbol{\phi}_{\ell z_{i\ell}}) \right).$$

如果为分量特定参数 $\phi_{\ell k}$ 和/或空视图模型的参数 $\phi_{0\cdot}$ 设定共轭先验，这些参数可以被积分消去，似然函数 f_x 和 q_x 可由边际似然替代；参见附录。

4.3 L的选择：视图数量

值得注意的是，对视图分配的推断与对每个（非空）视图内混合成分分配的推断有些类似。在后一种情况下，个体更有可能被分配到那些具有相似聚类变量特征的个体也被分配到的成分中；而在前一种情况下，聚类变量更有可能被分配到那些包含定义相似聚类结构的聚类变量的视图中。

类似地，多视图模型中视图数量 L 的选择，类似于传统混合模型中成分数量 K 的选择。此外，与混合模型中成分先验分配概率可以作为待推断的参数 π_k（例如，采用狄利克雷或狄利克雷过程先验分布）这一做法大致相同，也可以将多视图模型中的视图先验分配概率 ν_ℓ 视为参数。通过采用狄利克雷（或狄利克雷过程）先验分布，并令 L 取较大值（或无穷大），原则上可以自动推断视图数量。然而，这会带来相应的计算开销，因为每增加一个视图，就会引入一个混合模型，其参数和潜在成分分配都必须进行推断。针对这些模型的近似推断程序，例如变分贝叶斯（Guan 等人，2010 曾在多视图聚类背景下对其进行探讨），将是未来研究的重要方向；但在本文中，我们专注于较小的 $L \geq 2$ 值，因为此时通过马尔可夫链蒙特卡罗（MCMC）进行推断是可行的。

4.4 初始化

我们发现，一种有效的初始化策略是从相关视图中的所有变量开始，这样在后续迭代中就会“筛除”无关变量和空变量。我们在所有示例中都采用这种初始化策略

5 示例

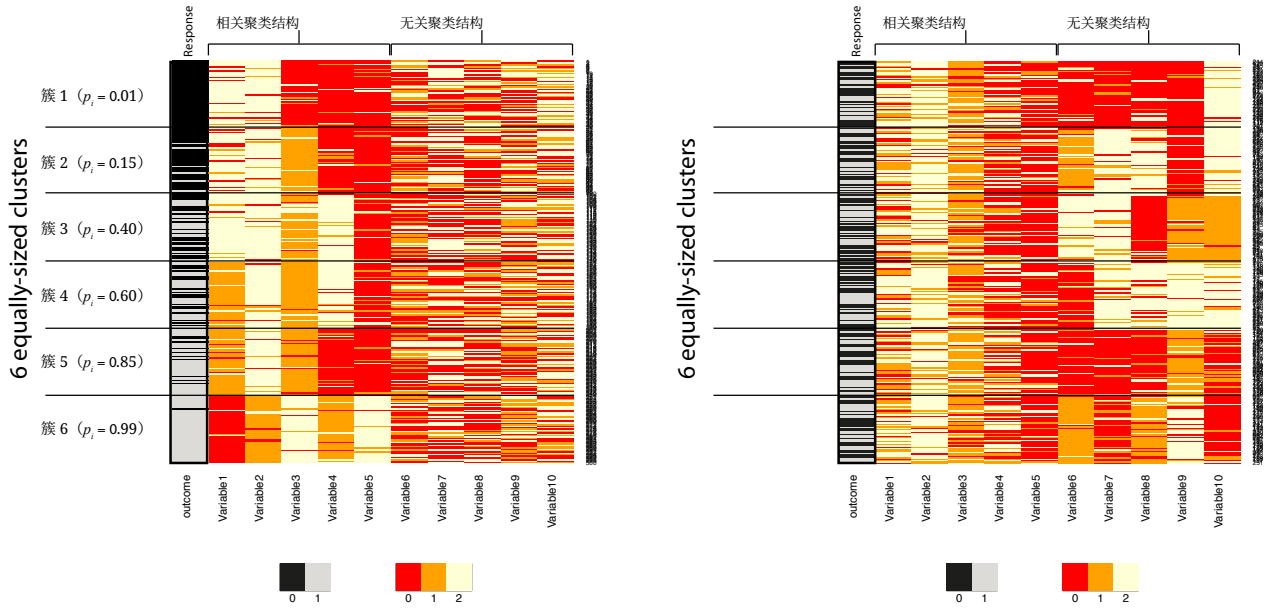
尽管到目前为止我们有意保持论述的普适性，但在接下来的示例中，我们将研究范围限定为非空视图采用狄利克雷过程混合模型的情形。第 5.1 节将介绍模拟示例，以说明即使利用了响应信息并采用了变量选择，不对多重聚类结构建模的聚类方法仍可能失效。第 5.2 节将介绍一个乳腺癌亚型表征背景下的整合聚类示例，其中我们拟合一个包含 3 个视图（包括一个空视图）的多视图模型。

5.1 用于说明忽略多个视图的方法局限性的模拟研究

我们构建了一个模拟示例，以展示现有未考虑多重聚类结构的半监督聚类方法的局限性。我们考虑包含 $n = 300$ 个体、 $p = 10$ 个类别型聚类变量（每个变量有 3 个类别）以及一个单变量二元响应 y 的模拟数据集。这些聚类变量定义了 2 个视图，其中前 q 个聚类变量（$q \in \{3, 5, 7, 9\}$）定义了一个相关的聚类结构（与响应变量相关），其余 $p - q$ 个变量定义了一个无关的聚类结构（与响应变量无关）。相关视图（视图 1）中的聚类变量定义了 6 个等规模聚类，这些聚类与响应变量之间的关系遵循 $P(y_i = 1|z_{i1} = k) = \theta_k$，其中 $\theta_k \in \{0.01, 0.15, 0.40, 0.60, 0.85, 0.90\}$。无关视图（视图 2）中的聚类变量同样定义了 6 个簇，但这些簇与响应变量之间不存在联系；即，无论视图 2 中的成分分配如何， $P(y_i = 1|z_{i2} = k) = P(y_i = 1) = 0.485$。图 2 展示了 $q = 5$ 数据集的示例，其中图 2a 显示了由视图 1 中聚类变量定义的聚类结构，图 2b 显示了由视图 2 中聚类变量定义的聚类结构（该结构与所示响应变量无关）。

在每个视图中，分类数据根据混合分布进行模拟。以 $w \in [0, 1]$ 的概率，我们根据以下方式模拟 x_{ij}：

$$
\begin{aligned}
x_{ij}|z_i = k, \gamma_j = \ell &\sim \text{Categorical}(\boldsymbol{\phi}_{\ell kj}), \\
\text{where } \boldsymbol{\phi}_{\ell kj} = [\phi_{\ell kj1}, \phi_{\ell kj2}, \phi_{\ell kj3}] &\sim \text{Dirichlet}(0.01, 0.01, 0.01),
\end{aligned} \tag{15}
$$



(a) 按行排序以突出相关聚类结构。 (b) 按行排序以突出无关聚类结构。

图 2：同一模拟数据集的两种不同的行排序。每个视图中的数据均根据方程（15）进行模拟，其中 $w = 0.8$。(a) 对行进行排序，以显示相关视图中由聚类变量（变量 1 至 5）定义的簇。个体的响应变量为 0/1 的概率取决于其所属的簇。(b) 与 (a) 相同，但对行进行排序，以突出显示无关视图中由聚类变量（变量 6 至 10）定义的簇。正如正文所述，这些（无关）簇与响应变量之间没有关联。

并且以概率 $1-w$，我们根据三个类别上的离散均匀分布模拟 x_{ij}。因此，参数 w 使我们能够控制簇可分性。当 w 接近 1 时，簇很容易分开；但随着 w 趋近于 0，簇中的噪声会越来越多，也越来越难以分开。

5.1.1 结果

我们将一种现有的半监督聚类方法应用于每个数据集，该方法由用于贝叶斯轮廓回归的 PReMiuM R 包实现（Liverani 等，2015）。对于每个数据集，

我们分别在启用和不启用变量选择的情况下运行 PReMiuM。若要启用变量选择，我们将 PReMiuM 中的“varSelectType”选项设为“continuous”，该设置会按照 Papathomas 等（2012）的描述，利用潜在选择权重进行变量选择。为了获得能够反映仅选择相关聚类变量时可达到性能的“金标准”，我们还在移除所有不相关聚类变量后，将 PReMiuM 应用于每个数据集。在实际示例中，不可能以这种方式事先识别相关聚类变量；但在此我们假设可以使用一个能够提供此信息的预言机。在所有情况下，我们均使用 PReMiuM 执行 10,000 次吉布斯抽样迭代，舍弃前 1,000 次作为预热期，并通过抽稀保留每 5 次迭代中的一次抽样。

为总结马尔可夫链蒙特卡罗（MCMC）的输出，我们计算了相关视图中的真实聚类结构与从后验分布中抽样得到的每个保留聚类结构（划分）之间的调整兰德指数（ARI）。因此，对于每次独立的 PReMiuM 运行，我们都得到一个 ARI 值分布，用于评估推断出的聚类结构与相关视图中的聚类结构的匹配程度。ARI 的最大值为 1（表示完全匹配），而 0 则表示匹配程度不优于随机情况下的预期。

图3展示了4个数据集的典型输出，每个数据集的 q （相关聚类变量的数量）取值各不相同。其他模拟数据集的对应图示见补充结果，其定性表现相似。正如预期，随着 q 减少，我们推断出正确聚类结构的能力也随之下降。或许不那么直观的是，我们还看到，由于未能对多个视图进行建模，使用变量选择运行 PReMiuM 并不一定有助于改善相关聚类结构的推断，实际上还可能造成损害。图3表明，当相关聚类变量相对于不相关聚类变量的数量较多时，变量选择可能是有用的（例如，10个变量中有7个相关）。在这些情况下，不相关聚类变量会被舍弃（即被大幅降低权重），从而取得与人为移除不相关聚类变量时相当的性能（例如，参见图3中的 $q = 7$ ）。然而，当相关聚类变量相对于不相关聚类变量的数量较少时（例如，10个变量中只有3个相关），变量选择可能会降低性能；因为在这些情况下，被舍弃的是相关聚类变量。由于 PReMiuM 将数据建模为仅具有一个（非空）聚类结构，因此它往往会

聚焦于单一的“主导”聚类结构（此处指由大多数聚类变量定义的结构）。变量选择会通过移除或降低定义不同聚类结构的聚类变量的权重来强化这一点；因此，如果主导聚类恰好也是相关聚类，性能会更好，否则会更差。

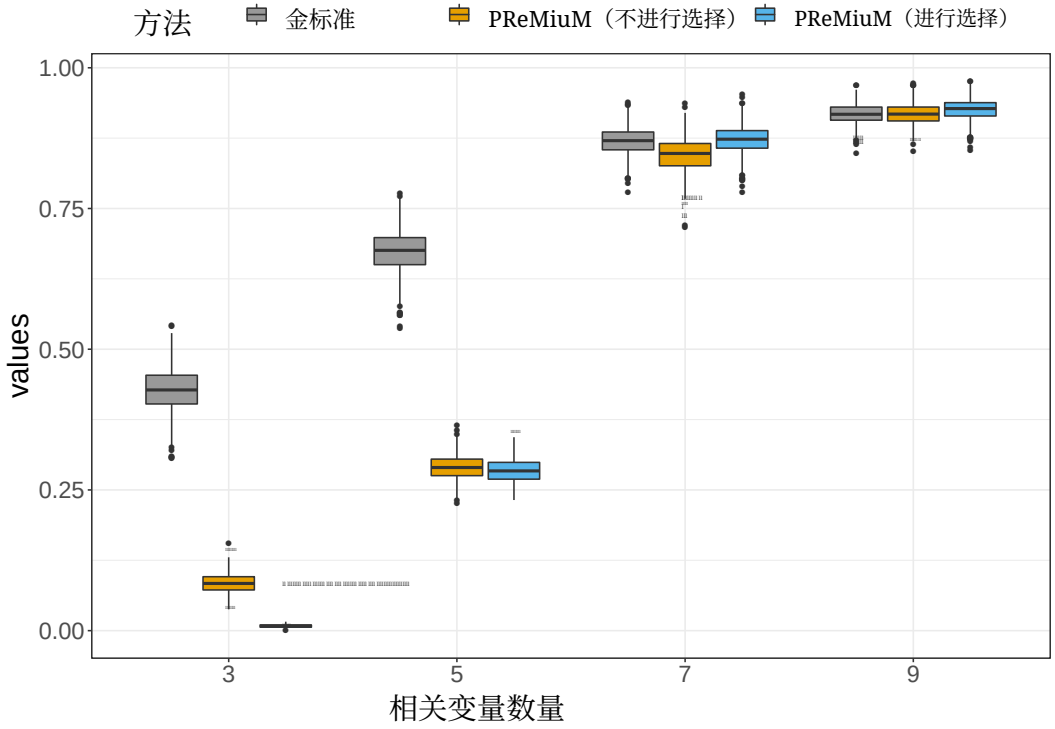


图3：将 PReMiuM 应用于包含相关和无关聚类变量的数据集时，所得到的调整兰德指数（ARI）值的后验分布。将 PReMiuM 应用于同时包含相关和无关变量的数据集时，可选择进行变量选择（蓝色）或不进行变量选择（金色）；也可将其应用于移除无关变量后的相同数据集（灰色；“金标准”表现）。实线：对每个数据集应用不进行变量选择的 PReMiuM。虚线：对每个数据集应用进行变量选择的 PReMiuM。点线：人为移除无关聚类变量后，将 PReMiuM 应用于每个数据集。

至关重要的是，我们强调，这里所见的结果并非 PReMiuM 中所实现的推断或变量选择程序所特有，而是在数据中存在多个（非零）视图时，未能对其进行建模所导致的。

5.2 乳腺癌亚型划分

近年来，许多作者（例如 Shen 等人，2009；Kirk 等人，2012；Lock 和 Dunson，2013；Savage 等人，2013）探讨了如何利用多个组学数据集进行整合聚类，以便在分子层面更好地表征癌症亚型（另见 Kristensen 等人，2014 年的综述）。尽管此前尚未将多视图方法用于这一目的，但它们可以直接用于整合聚类。在多视图模型中，如果聚类变量定义了相同的聚类结构，就会被分配到同一个视图。无论这些聚类变量是否来自同一个数据集，或是否属于同一种数据类型，都无关紧要；唯一重要的是它们所定义的聚类结构。

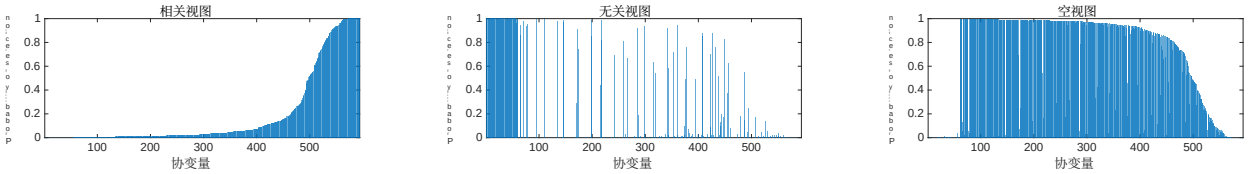
在此，我们基于来自癌症基因组图谱（TCGA；癌症基因组图谱，2012年）的反相蛋白质阵列（RPPA）和微小RNA（miRNA）数据，应用半监督多视图建模，对乳腺癌肿瘤样本进行聚类。现有大量研究探讨了如何利用mRNA表达数据识别癌症亚型，其中包括PAM50预测模型，该模型根据50个基因的表达情况对乳腺癌肿瘤样本进行分类（Parker 等，2009）。我们研究了 $n = 108$ 个乳腺癌肿瘤样本，其中66个基于mRNA表达数据被划分为基底样型，42个被划分为管腔*A*型。我们将该分类作为二元响应变量，用于指导基于TCGA的RPPA和miRNA数据进行聚类。RPPA数据包含171种蛋白质的测量值，miRNA数据则包含423种微小RNA的测量值。

聚类前，我们按照 Lock 和 Dunson (2013) 的方法处理微小RNA和反相蛋白质阵列数据集。为了增强对连续数据模型设定错误的稳健性，我们额外进行了预处理：在每个肿瘤样本内采用三分位离散化，并将所得数据视为类别型。预处理后，我们将数据集合并，因此最终使用的数据集包含 $p = 594$ 个聚类变量（对应 171 种蛋白质和 423 种微小RNA）。

5.2.1 结果

我们将半监督多视图模型拟合到拼接后的癌症基因组图谱 miRNA 和反相蛋白质阵列数据上，并假设 $L = 3$ 个视图（1 个相关视图、1 个无关视图和 1 个空视图）。我们进行了 10,000 次吉布斯抽样迭代，去除前 5,000 次作为预热期，然后进行抽稀，每 5 次抽样保留 1 次。对于每个聚类变量，我们计算了其被选入各视图的概率的蒙特卡罗估计。

变量被选入各视图的概率。图4展示了这些概率的条形图。大多数聚类变量被选入空视图的概率较高，而被选入其他两个视图的概率较高的变量相对较少。

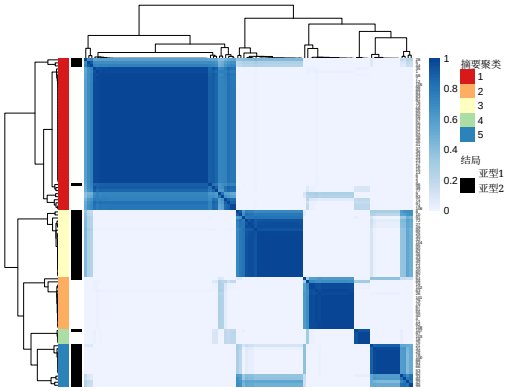


(a) (b) (c)

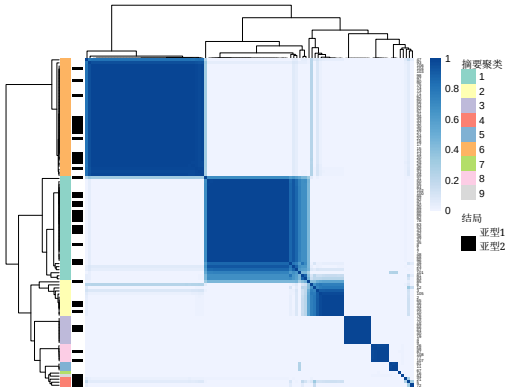
图4：每个聚类变量被选择进入（a）相关视图；（b）无关视图；以及（c）空视图的概率。在每个图中，聚类变量沿 x-轴排列，其顺序依据被选择进入相关视图的概率递增。

为总结每个非空视图中的聚类结构，我们首先计算了每个视图的后验相似性矩阵（PSM）。PSM 是一个 $n \times n$ 矩阵，其 i，j项表示从后验分布中抽取的聚类结果中，肿瘤样本 i 和 j 具有相同簇标签的比例（Fritsch 和 Ickstadt，2009）。由于 PSM 总结了成对共聚类概率，因此它们提供了对马尔可夫链蒙特卡罗输出的概括，并避免了与标签切换相关的难题。非空视图的 PSM 见图5。从图5a可以清楚看出，正如预期，相关视图中的聚类结构与亚型标签具有很强的关联。如图5b所示，无关视图同样具有明显的聚类结构，但该结构与亚型标签无关。

为了将相关视图中的簇可视化，我们对选择概率设定阈值，仅保留那些被选入相关视图的概率至少为 0.90 的聚类变量。这样仅剩 53 个聚类变量，其中 32 个是微小RNA，21 个是蛋白质。图6a 展示了这 53 个聚类变量的数据热图表示，从中可以清楚地看出聚类结构及其与响应变量（亚型）的关联。我们以类似方式将无关视图中的簇可视化，仅保留那些被选入无关视图的概率至少为 0.90 的聚类变量。这样剩下 76 个聚类变量，且全部是蛋白质。由此得到的热图表示见图6b。在这种情况下，虽然存在明显的聚类结构，但它与肿瘤亚型无关。



(a) 后验相似度矩阵：相关视图



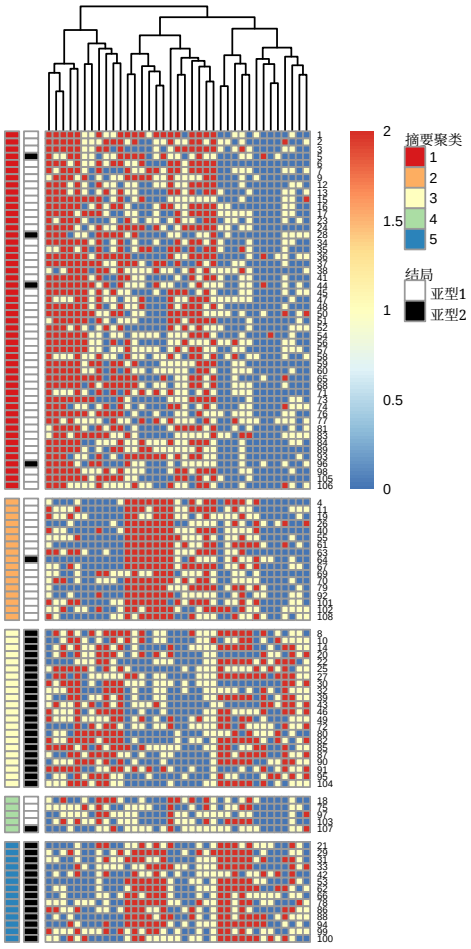
(b) 后验相似度矩阵：无关视图

图5：后验相似度矩阵总结了在（a）相关视图和（b）无关视图中个体被分配到各簇的情况。为便于可视化，对后验相似性矩阵的行和列均进行了层次聚类（如树状图所示）。

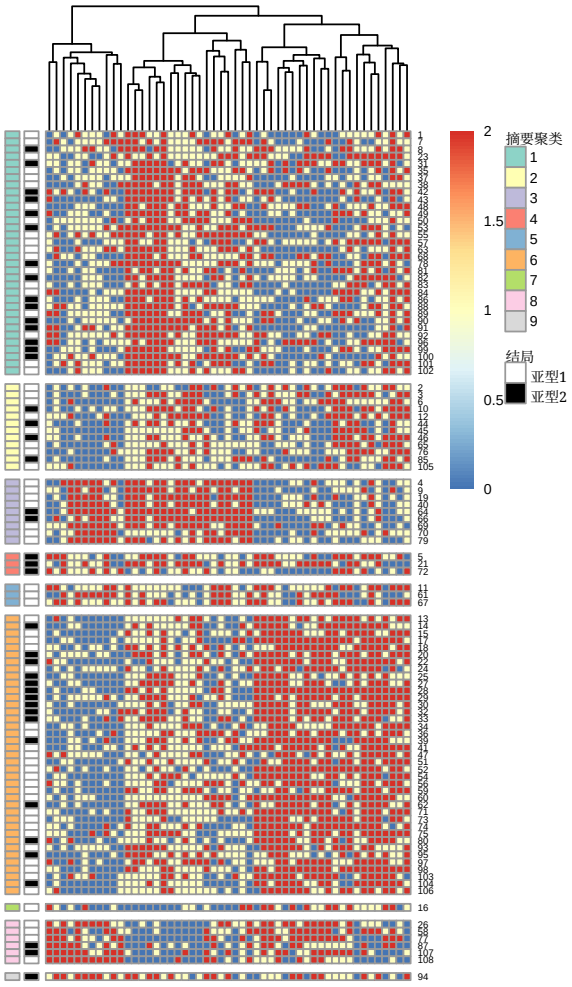
6 讨论

我们已经证明，当数据中存在多重聚类结构时，现有的（单视图）聚类方法可能无法找出最相关的聚类结构，即便有适当的响应变量作为引导（第 5.1.1 节）。此外，传统的聚类变量选择方法未必能改善这一情况，因为无论聚类结构是否与关注的响应变量相关，这些方法往往都会选择定义主导聚类结构的变量。在第 5.2.1 节中，我们证明了真实的分子数据集确实可能具有多重聚类结构，而且我们的半监督多视图模型能够识别相关结构和无关结构。

尽管多视图方法相较于现有的（单视图）替代方案显然具有优势，但计算方面的考量仍是一项显著挑战。狄利克雷过程混合模型本身计算成本就很高，而本文提出的多视图方法会针对每个新增视图引入一个额外的狄利克雷过程混合模型。尽管在本文考察的示例中，限制视图数量也能取得足够好的结果，但如果数据中“真实”的视图数量大于为 L 选定的值，情况就未必如此。两方面的计算



（a）聚类结构：相关视图



（b）聚类结构：无关视图

图6：（a）相关视图和（b）无关视图中聚类结构的可视化。行对应于肿瘤样本，每个图中的列对应于被选入（a）相关视图和（b）无关视图的概率至少为0.90的聚类变量。对每个图的行和列均应用了层次聚类，树状图对此进行了展示。每个肿瘤样本的PAM50亚型通过显示在每个图左侧的附加列标示。

方法（例如 Suchard 等，2010 中的并行化）和快速近似推断程序（例如 Guan 等人，2010 中的变分贝叶斯）将是未来工作中的重要考量。

资助

PDWK 和 SR 分别获得了 MRC（MC UU 00002/13 和 MC UU 00002/10）的支持。FP 和 SR 获得了 EPSRC 项目 EP/R018561/1 的支持。

7 附录

7.1 无监督贝叶斯混合模型

我们首先考虑以下一般形式的无监督混合模型：

$$p(\mathbf{x}|\boldsymbol{\phi}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\phi}_k), \tag{16}$$

其中，$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$是混合权重向量，$\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$是所有分量特定参数的集合。我们先考虑有限 K的情形，稍后再回到无限混合模型。

与混合模型中的常见做法一样，我们引入潜在成分分配变量 z_i，其中$z_i = k$ 表示第 i个观测$\mathbf{x}_i$ 与第 k个成分相关联，而 $p(z_i = k|\pi) = \pi_k$。于是，

$$p(\mathbf{x}_i|z_i, \boldsymbol{\phi}) = f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_{z_i}), \tag{17}$$

因此

$$p(\mathbf{x}_i, z_i = k|\boldsymbol{\phi}, \pi) = f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_k)p(z_i = k|\boldsymbol{\pi}) \tag{18}$$

$$= f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_k)\pi_k. \tag{19}$$

对所有 K 个可能取值求和以消去 z_i 后，我们得到（正如我们所希望的）：

$$p(\mathbf{x}_i|\boldsymbol{\phi}, \pi) = \sum_{k=1}^{K} \pi_k f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_k). \tag{20}$$

采用通常的 c　　在条件独立性假设下，$\mathbf{x}_i$、 z_i、 $\boldsymbol{\phi}$、 $\boldsymbol{\pi}$ 的完整联合模型为　　：

$$p(\mathbf{x}_i, z_i, \boldsymbol{\phi}, \pi) = f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_{z_i})p(z_i|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\phi}) \tag{21}$$

$$= f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_{z_i})p(z_i|\boldsymbol{\pi})p(\boldsymbol{\pi})\prod_{k=1}^{K} p(\boldsymbol{\phi}_k), \tag{22}$$

其中，我们假设分量特定参数具有独立先验 ϕ_k，混合权重具有对称狄利克雷先验 π_1，**.....**，$\pi_K \sim$ **Dir(**α/K**)**。

对于完整数据集，我们有：

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, z_1, \ldots, z_n, \boldsymbol{\phi}, \boldsymbol{\pi}) = \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_{z_i})p(z_i|\boldsymbol{\pi})\right) p(\boldsymbol{\pi})p(\boldsymbol{\phi}) \tag{23}$$

$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_{z_i})p(z_i|\boldsymbol{\pi})\right) p(\boldsymbol{\pi})\prod_{k=1}^{K} p(\boldsymbol{\phi}_k). \tag{24}$$

7.1.1 通过吉布斯抽样进行推断（有限 K 情形）

根据公式（24），可以直接写出用于吉布斯抽样的条件分布。暂时假设 K 有限（稍后将由此推导出无限极限）。

条件分布 $\boldsymbol{\phi}_k$ 通过考察方程（24）的右侧，我们得到：

$$p(\boldsymbol{\phi}_k|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_1, \ldots, z_n, \boldsymbol{\phi}_{-k}, \pi_1, \ldots, \pi_K) \propto p(\boldsymbol{\phi}_k) \prod_{i:z_i=k} f_{\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\phi}_{z_i}), \tag{25}$$

其中， $\boldsymbol{\phi}_{-k}$ 表示由所有满足 $j \neq k$的 $\boldsymbol{\phi}_j$ 组成的集合。因此， $\boldsymbol{\phi}_k$ 的条件分布是给定所有满足 $z_i = k$的$\mathbf{x}_i$ 时， $\boldsymbol{\phi}_k$ 的后验密度。若采用共轭先验，则该后验分布可以通过解析方法求得；否则，可以通过例如 Metropolis-Hastings 算法抽取样本。请注意，如果不存在满足 $z_i = k$ 的 $\mathbf{x}_i$ （即第 k个成分没有关联的观测值），那么 $\boldsymbol{\phi}_k$ 只需从先验分布 p**(**$\boldsymbol{\phi}_k$**)**中抽样。

的条件分布通过考察方程（24）右侧可得，　　：

$$p(\pi_1, \ldots, \pi_K|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_1, \ldots, z_n, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K) \propto \left(\prod_{i=1}^{n} p(z_i|\pi_1, \ldots, \pi_K)\right) p(\pi_1, \ldots, \pi_K). \tag{26}$$

因此， π 的条件分布是给定分类潜在分配变量 z_i、 $i = 1$、**.....**、 n取值后， π 的后验分布。若采用共轭狄利克雷先验，该后验分布可写成闭式解。

的条件分布通过考察方程（24）右侧可得 :

$$p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K, \pi_1, \ldots, \pi_K, z_{-i}) \propto p(z_i = k|\pi) f_{\mathbf{x}}(\mathbf{x}_i|\phi_k), \tag{27}$$

$$= \pi_k f_{\mathbf{x}}(\mathbf{x}_i|\phi_k), \tag{28}$$

其中 z_{-i} 表示由所有满足 $j \neq i$ 的 z_j 构成的集合。由于 $\sum_{k=1}^{K} p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\phi}, \pi, z_{-i}) = 1$，因此条件分布为:

$$p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\phi}, \pi, z_{-i}) = \frac{\pi_k f_{\mathbf{x}}(\mathbf{x}_i|\phi_k)}{\sum_{k=1}^{K} \pi_k f_{\mathbf{x}}(\mathbf{x}_i|\phi_k)}, \tag{29}$$

对于有限的 K，可以直接计算。

边缘化 π 对于 π 采用共轭狄利克雷先验，另一种策略是对 $\boldsymbol{\pi}$ 进行边缘化，而不是对其进行抽样。我们假设其先验分布为对称狄利克雷分布，浓度参数为 α/K。

注意， z_i 仅在给定 $\boldsymbol{\pi}$ 的条件下彼此条件独立，因此如果我们对 $\boldsymbol{\pi}$ 进行边缘化，就必须在 z_i 的条件分布中谨慎地建模 z_i 对 z_{-i} 的依赖关系。

我们有

$$p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\phi}, \boldsymbol{\pi}, z_{-i}, \alpha) \propto p(z_i = k|z_{-i}, \pi, \alpha) f_{\mathbf{x}}(\mathbf{x}_i|\phi_k) \qquad \text{[cf. Equation (27)]}. \tag{30}$$

因此，要对 $\boldsymbol{\pi}$ 进行边缘化，我们必须计算 $\int_{\pi} p(z_i = k|z_{-i}, \boldsymbol{\pi}, \alpha) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi} = p(z_i = k|z_{-i}, \alpha)$，这是给定其他潜在分配变量 z_{-i} 的情况下，z_i 的条件先验分布。我们有，

$$p(z_i = k|z_{-i}, \alpha) = \frac{p(z_i = k, z_{-i}|\alpha)}{p(z_{-i}|\alpha)} \tag{31}$$

$$= \frac{\int_{\boldsymbol{\pi}} p(z_i = k, z_{-i}|\boldsymbol{\pi}, \alpha) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi}}{\int_{\boldsymbol{\pi}} p(z_{-i}|\boldsymbol{\pi}, \alpha) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi}} \tag{32}$$

$$= \frac{\int_{\boldsymbol{\pi}} p(z_i = k, z_{-i}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi}}{\int_{\boldsymbol{\pi}} p(z_{-i}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi}}, \tag{33}$$

其中，在最后一行中，我们利用了这样一个事实：给定 π 后，z_i 与 α 条件独立。

为了继续推导，我们必须计算这个分数。为此，我们需要用到一个关于狄利克雷分布的标准结果：根据该分布抽取的随机变量，其矩可以表示为

趋于对称　参数为 α/K 的狄利克雷分布如下　　:

$$E\left[\prod_{k=1}^{K}\pi_k^{m_k}\right]=\frac{\Gamma(\sum_{k=1}^{K}(\alpha/K))}{\Gamma(\sum_{k=1}^{K}((\alpha/K)+m_k))}\times\prod_{k=1}^{K}\frac{\Gamma((\alpha/K)+m_k)}{\Gamma(\alpha/K)},\tag{34}$$

其中，m_k是任意自然数。

此外，我们注意到以下两个等式:

$$p(z_i=k,z_{-i}|\pi)=\pi_k^{n_{-i,k}+1}\prod_{\substack{c=1,\ldots,K\\c\neq k}}\pi_c^{n_{-i,c}},$$

and

$$p(z_{-i}|\pi)=\pi_k^{n_{-i,k}}\prod_{\substack{c=1,\ldots,K\\c\neq k}}\pi_c^{n_{-i,c}},$$

其中，$n_{-i,\ c}$ 是满足 $z_j=c$的 z_j的数量，其中 $j\neq i$ 。因此，我们可以使用公式（34）中的结果来计算公式（33）右侧的分子和分母。经过一些代数运算，并利用伽马函数的性质Γ($t+1$) $=t$Γ(t)，可得:

$$p(z_i=k|z_{-i},\alpha)=\frac{n_{-i,k}+\alpha/K}{n-1+\alpha}.\tag{35}$$

因此,

$$p(z_i=k|\mathbf{x}_1,\ldots,\mathbf{x}_n,\boldsymbol{\phi},z_{-i},\alpha)\propto\frac{n_{-i,k}+\alpha/K}{n-1+\alpha}\times f_{\mathbf{x}}(\mathbf{x}_i|\phi_k).\tag{36}$$

此外，由于 K 是有限的，我们可以直接计算出以下等式:

$$p(z_i=k|\mathbf{x}_1,\ldots,\mathbf{x}_n,\boldsymbol{\phi},z_{-i},\alpha)=\frac{1}{Z}\frac{n_{-i,k}+\alpha/K}{n-1+\alpha}\times f_{\mathbf{x}}(\mathbf{x}_i|\phi_k),\tag{37}$$

其中

$$Z=\sum_{c=1}^{K}\left(\frac{n_{-i,c}+\alpha/K}{n-1+\alpha}\times f_{\mathbf{x}}(\mathbf{x}_i|\phi_c)\right).\tag{38}$$

边缘化 ϕ 同样，如果存在适用于 ϕ_k的共轭先验分布，那么也可以对其进行边缘化。**请注意**（与边缘化 π时处理 z_i的情况类似），只有在给定 ϕ_k和 z_i的条件下，$\mathbf{x}_i$之间才条件独立。因此，如果我们对ϕ 进行边缘化，就必须在 z_i的条件分布中谨慎地对 v_i 与 v_{-i} 之间的依赖关系建模。

经过一些代数运算，可以很容易地证明，对 ϕ 进行边缘化后， z_i的条件分布如下：

$$p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}, \alpha) = \frac{1}{Z}\frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha} \times \int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_i|\phi_k)p(\phi_k|\mathbf{x}_{-i,k})d\phi_k, \tag{39}$$

其中，$\mathbf{x}_{-i,k}$ 表示所有满足 $j \neq i$ 和 $z_j = k$的观测值$\mathbf{x}^j$ ，因此 $p(\phi_k|\mathbf{x}_{-i,k})$是 ϕ_k 的后验分布，条件是当前与成分 k 相关联的所有观测值（不包括$\mathbf{x}_i$）。如果不存在满足 $j \neq i$ 和 $z_j = k$ 的$\mathbf{x}^j$（即 k 是一个没有分配其他观测值的成分），则称第 k个成分为空$empty$ ，并将$p(\phi_k|\mathbf{x}_{-i,k})$定义为 ϕ_k的先验分布$= p(\phi_k)$。

实现采样器时，注意到以下这一点很有用：

$$p(\phi_k|\mathbf{x}_{-i,k}) = \frac{f_{\mathbf{x}}(\mathbf{x}_{-i,k}|\phi_k)p(\phi_k)}{\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_{-i,k}|\phi_k)p(\phi_k)d\phi_k}, \text{ if the } k\text{-th component is not empty.}$$

因此，仍假设第 k个成分非空，公式（39）中的积分为

$$\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_i|\phi_c)p(\phi_k|\mathbf{x}_{-i,k})d\phi_k = \frac{\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_{-i,k}|\phi_k)p(\phi_k)d\phi_k}{\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_{-i,k}|\phi_k)p(\phi_k)d\phi_k}, \tag{40}$$

这是边际似然之比：一个边际似然将$\mathbf{x}_i$ 纳入与成分 k相关联的观测值中，另一个则将$\mathbf{x}_i$ 排除在与成分 k相关联的观测值之外。

这个表达式有助于理解采样器：在每次迭代中，对于每个成分，我们都会权衡$\mathbf{x}_i$ 与成分 k 相关联的证据，以及$\mathbf{x}_i not$ 与成分 k 不相关联的证据（给定当前与该成分相关联的其他观测值 $\mathbf{x}_{-i,k}$）。直观来说，这个表达式确保我们更有可能将$\mathbf{x}_i$ 分配给一个此前已分配过相似观测值的成分。

还要注意，公式（39）中的项 $\frac{n_{-i,k}+\alpha/K}{n-1+\alpha}$ 表示将$\mathbf{x}_i$ 分配给成分 k 的条件先验概率，反映了在给定其他所有观测值的分配情况后，我们倾向于将$\mathbf{x}_i$ 分配给成分 k的先验信念。由于 $n_{-i,k}$ 位于分子中，这表达了一种“富者愈富”的先验信念；也就是说，先验地，我们更倾向于将$\mathbf{x}_i$ 分配给已经分配了许多观测值的成分，而非观测值较少的成分。

最后说明 请注意，在对 π_k 和 ϕ_k 积分消去后，我们可以使用公式（39）仅对 z_i 进行采样，而无需对其他任何参数进行采样。唯一的例外是 α

超参数，我们可以将其固定，也可以进行采样（例如，采用 Escobar 和 West（1995）中介绍的方法）。

7.1.2 吉布斯抽样的条件分布（K 无限）

通过考虑极限 $K \to \infty$，可以从有限情形推导出无限情形。关键变化在于成分分配变量的条件先验分布， $p(z_i = k|z_{-i}, \alpha)$。需要考虑两种情况：

1. k **是已有（非空）成分的标签：**这种情况是指，对于某个 $j \neq i$， $k = z_j$。考虑公式（35）中 $K \to \infty$ 的极限，可得：

$$\text{If } k = z_j \text{ for some } j \neq i, \text{ then} \qquad p(z_i = k|z_{-i}, \alpha) = \frac{n_{-i,k}}{n - 1 + \alpha}. \tag{41}$$

2. k **是新标签：**这种情况是指，对于所有 $j \neq i$， $k \neq z_j$。因此，我们要计算的概率为 p（对于所有 $j \neq i|z_{-i}, \alpha, z_i \neq z_j$）。借助以下恒等式，可以计算该概率：

$$1 = p(z_i \neq z_j \text{ for all } j \neq i|z_{-i}, \alpha) + \sum_{k \in \mathcal{C}_{-i}} p(z_i = k|z_{-i}, \alpha),$$

其中 $\mathcal{C}_{-i} = \{k：\text{对于某个 } j \neq i\}$， $z_j = k$ 是所有当前成分标签（不包括 z_i）构成的集合。利用该恒等式，经过一些代数运算后，可得：

$$p(z_i \neq z_j \text{ for all } j \neq i|z_{-i}, \alpha) = \frac{\alpha}{n - 1 + \alpha}. \tag{42}$$

的条件分布利用公式（41）和（42），可以直接证明（参见公式（39），并利用公式（40）），我们得到以下关于 z_i 的条件分布：

$$\text{If } k = z_j \text{ for some } j \neq i, \text{ then} \qquad p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}, \alpha) = b\frac{n_{-i,k}}{n - 1 + \alpha}\frac{\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_{-i,k}|\phi_k)p(\phi_k)d\phi_k}{\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_{-i,k}|\phi_k)p(\phi_k)d\phi_k},$$

$$\text{且 } p(z_i \neq z_j \text{对于所有 } j \neq i|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}, \alpha) = b\frac{\alpha}{n - 1 + \alpha}\int_{\phi_k} f_{\mathbf{x}}(\mathbf{x}_i|\phi_k)p(\phi_k)d\phi_k, \tag{43}$$

(44)

其中 b 是一个归一化常数，用于确保

$$p(z_i \neq z_j \text{ for all } j \neq i|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}, \alpha) + \sum_{k \in \mathcal{C}_{-i}} p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}, \alpha) = 1.$$

除符号记法的变更及部分项的展开外，这些表达式与 Neal（2000）中的公式（3.7）完全相同。

7.2 针对类别型数据的剖面回归

回顾一下，我们的数据由输入—输出（即聚类变量向量—响应变量）对组成， $D = \{\mathbf{v}_i\}_{i=1}^n$，其中$\mathbf{v}_i = [\mathbf{x}_i, \ y_i]^\top$，其中$\mathbf{x}_i$ 和 y_i 分别表示第 i个聚类变量向量和响应变量。现在我们考虑聚类变量和响应变量均为类别型的情况。

7.2.1 对聚类变量建模

现在假设每个聚类变量（即向量$\mathbf{x}_i$中的每个元素）均为类别型，其中第 j个聚类变量有 R_j 个类别，记为 1、2、.....、 R_j。我们使用分类分布对数据建模。定义$\phi_{k,\ j,\ r}$ 为第 j个聚类变量在第 k个成分中取值为 r 的概率，并将与第 k个成分中第 j个聚类变量相关的概率集合记为$\Phi_{k,\ j} = [\phi_{k,j,1}, \phi_{k,j,2}, \ldots, \phi_{k,j,R_j}]$。我们进一步将与第 k个成分相关的所有概率（涵盖全部 J个聚类变量）记为 $\Phi_k = \{\Phi_{k,1}, \Phi_{k,2}, \ldots, \Phi_{k,J}\}$，并将与非空成分相关的所有 Φ_k记为 $\Phi = \{\Phi_k\}_{k \in \mathcal{C}}$（此处， $\mathcal{C} = \{k:$ 对于某个 $i \in \{1, \ldots, n\}\}$， $z_i = k$）。

我们假设，给定聚类变量的成分分配后，这些变量条件独立，因此

$$f_{\mathbf{x}}(\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iJ}]|\mathbf{\Phi}, z_i = k) = \phi_{k,j,x_{i1}} \phi_{k,j,x_{i2}} \ldots \phi_{k,j,x_{iJ}} \tag{45}$$

$$= \prod_{j=1}^{J} \phi_{k,j,x_j} \tag{46}$$

$\Phi_{k\text{的条件分布},\ j}$ 根据公式（25），吉布斯抽样所需的条件分布是在给定与第 k 个成分相关的观测值时，$\Phi_{k,j}$的后验分布。对于每个 j,

我们为 $\Phi_{k,j}$采用共轭狄利克雷先验，

$$\Phi_{k,j} \sim \text{Dirichlet}(\mathbf{a}_j),$$

其中 $\mathbf{a}_j = [a_{j,1}, \ldots, a_{j,R_j}]$ 是浓度参数向量。于是后验分布为：

$$\Phi_{k,j}|x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k},j}, \mathbf{a}_j \sim \text{Dirichlet}(\mathbf{a}_j + [s_{k,j,1}, s_{k,j,2}, \ldots, s_{k,j,R_j}]), \tag{47}$$

其中 $\mathbf{x}^{i_1}$，$\mathbf{x}^{i_2}$，……，$\mathbf{x}^{i_{n_k}}$是与成分 k相关联的观测值，而 $s_{k,\ j,\ r}$定义为与成分 k 相关联且第 j 个聚类变量属于类别 r的观测值数量。

边缘化 $\Phi_{k,j}$ 我们还可以对 $\Phi_{k,j}$ 积分，以写出与 $x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k},j}$相关联的边际似然。请注意，边际似然（根据定义）是乘积 $\phi_k^{s_{k,\ j\ 1,\ j,1}} \cdots \phi_k^{s_{k,\ j\ R_j,\ j,R_j}}$ 的先验期望。因此，我们可以使用推导公式（34）时所用的同一标准结果，直接写出边际似然。仍假设 $\mathbf{x}^{i_1}, \mathbf{x}^{i_2}, \ldots, \mathbf{x}^{i_{n_k}}$ 是与成分 k 相关联的观测值，则有：

$$p(x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k},j}|\mathbf{a}_j) = \frac{\Gamma(\sum_{r=1}^{R_j} a_{j,r})}{\Gamma(\sum_{r=1}^{R_j}(a_{j,r} + s_{k,j,r}))} \times \prod_{r=1}^{R_j} \frac{\Gamma(a_{j,r} + s_{k,j,r})}{\Gamma(a_{j,r})}. \tag{48}$$

为简化记号，定义 $X_k = \{\mathbf{x}^{i_1}, \mathbf{x}^{i_2}, \ldots\ldots, \mathbf{x}_{i_{n_k}}\}$ 为与成分 k 相关联的观测值集合，并定义 $X_{k,\ j} = \{x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k},j}\}$ 为包含 j 个向量中第 j 个元素的集合。由于我们假设聚类变量在给定其成分分配的条件下相互独立，因此有：

$$p(X_k|\mathbf{a}_1, \ldots, \mathbf{a}_J) = \prod_{j=1}^{J} p(X_{k,j}|\mathbf{a}_j), \tag{49}$$

其中 $p(X_{k,\ j}|\mathbf{a}j) = p(x_{i_1,\ j},\ x_{i_2,\ j},\ \ldots\ldots,\ x_{i_{n_k},\ j}|\mathbf{a}j)$ 如公式（48）所示。

7.2.2 响应变量建模

我们假设响应变量 y_i 也为类别型（这次有 R_y 个类别），并使用分类分布进行建模。与之前类似，我们定义 $\theta_{k,\ r}$ 为 y在第 k 个成分中取值 r 的概率，并用 $\Theta_k = [\theta_{k,1}, \theta_{k,1}, \ldots, \theta_{k,R_y}]$ 表示与第 k 个成分相关的所有概率的集合。此外，定义 $Y_k = \{y_{i_1}, y_{i_2}, \ldots, y_{i_{n_k}}\}$ 为分配给成分 k 的 y 集合。

的条件分布我们对 θ_k 采用共轭狄利克雷先验，

$$\theta_k \sim \text{Dirichlet}(\mathbf{a}_y),$$

其中$\mathbf{a}_y = [a_{y,1}, \ldots, a_{j,R_y}]$是浓度参数向量。与之前类似， θ_k 的后验分布为

$$\theta_k | Y_k, \mathbf{a}_y \sim \text{Dirichlet}(\mathbf{a}_y + [s_{k,y,1}, s_{k,y,2}, \ldots, s_{k,y,R_y}]), \tag{50}$$

其中，$s_{k,y,r}$ 是 Y_k 中属于第 r类的观测值数量。

对 θ_k 进行边缘化。如前所述，我们也可以计算边际似然， $p(Y_k|\mathbf{a}_j)$。我们有

$$p(Y_k|\mathbf{a}_y) = \frac{\Gamma(\sum_{r=1}^{R_y} a_{y,r})}{\Gamma(\sum_{r=1}^{R_y}(a_{y,r} + s_{k,y,r}))} \times \prod_{r=1}^{R_y} \frac{\Gamma(a_{y,r} + s_{k,y,r})}{\Gamma(a_{y,r})}. \tag{51}$$

7.2.3 联合边际似然

为了继续推导，我们需要一个与 V_k $=\{X_k, Y_k\}$相关的边际似然表达式。我们假设，给定各自的成分分配后， y 与$\mathbf{x}$ 条件独立。因此

$$p(V_k|\mathbf{a}_1, \ldots, \mathbf{a}_J, \mathbf{a}_y) = p(Y_k|\mathbf{a}_y) \prod_{j=1}^{J} p(X_{k,j}|\mathbf{a}_j), \tag{52}$$

其中 $p(Y_k|\mathbf{a}_y)$ 和 $p(X_{k,\ j}|\mathbf{a}_j)$ 的表达式如前所述。需要注意的是 ：

1. 令 $V_k = \{\mathbf{x}_i,\ \mathbf{x}_{-i,\ k}\}$，我们可以计算公式（43）分子中的边际似然。

2. 令 $V_k = \{\mathbf{x}_{-i,\ k}\}$，我们可以计算公式（43）分母中的边际似然。

3. 令 $V_k = \{\mathbf{x}_i\}$，我们可以计算公式（44）中的边际似然。

因此，我们可以评估 z_i 的条件分布所需的所有项，从而（暂且不考虑对 α 进行采样的问题）获得执行模型推断所需的一切。

7.3 变量选择

7.3.1 聚类变量数据的零模型

我们在假设不存在聚类结构的前提下，为$\mathbf{x}$ 数据引入一个零模型。在零模型下，我们再次使用类别分布对聚类变量数据建模，并且——与前面类似——定义 $\Phi_{0,j} = [\phi_{0,j,1}, \phi_{0,j,2}, \ldots, \phi_{0,j,R_j}]$ 和 $\Phi_0 = \{\Phi_{0,1}, \Phi_{0,2}, \ldots, \Phi_{0,J}\}$。然而，在零模型下，我们假设不存在聚类结构（或者等价地，假设数据构成一个大的单一簇）。因此，与$\mathbf{x}_i$ 相关的似然不涉及成分分配变量 z_i，而是：

$$f_{\mathbf{x}}(\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iJ}]^\top | \mathbf{\Phi}_0) = \prod_{j=1}^{J} \phi_{0,j,x_{ij}} \tag{53}$$

在我们的零模型中，假设观测值 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 在给定$\mathbf{\Phi}_0$的条件下相互独立。因此，在零模型下，与完整聚类变量数据集相关的似然为：

$$f_{\mathbf{x}}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathbf{\Phi}_0) = \prod_{i=1}^{n} \prod_{j=1}^{J} \phi_{0,j,x_{ij}}. \tag{54}$$

7.3.2 变量选择

我们考虑这样一种可能性：只有部分聚类变量与确定聚类结构相关，而其余变量则无关。我们引入二元指示变量 γ_j，其中 $\gamma_j = 1$ 表示第 j 个聚类变量与聚类结构相关，而$\gamma_j = 0$ 表示该变量无关。

简单来说，我们希望对相关变量使用完整的混合模型，对无关变量使用零模型。更准确地说，我们针对 $\mathbf{x}$ 考虑以下模型：

$$f_{\mathbf{x}}(\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iJ}]^\top | \mathbf{\Phi}, \mathbf{\Phi}_0, z_i = k) = \prod_{j=1}^{J} \phi_{k,j,x_{ij}}^{\mathbb{I}(\gamma_j=1)} \phi_{0,j,x_{ij}}^{\mathbb{I}(\gamma_j=0)}. \tag{55}$$

7.3.3 推断

我们来考察引入 γ_j's如何影响我们推断模型参数。和之前一样，我们首先写出完整的联合模型：

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, z_1, \ldots, z_n, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Phi}_0, \pi, \alpha, \gamma) =$$
$$p(\pi|\alpha)p(\Phi_0)\left(\prod_{j=1}^{J} p(\gamma_j)\right)\left(\prod_{k=1}^{K} p(\Phi_k)p(\theta_k)\right)\left(\prod_{i=1}^{n} p(z_i|\pi)\theta_{z_i,y_i}\left(\prod_{j=1}^{J} \phi_{z_i,j,x_{ij}}^{\mathbb{I}(\gamma_j=1)}\phi_{0,j,x_{ij}}^{\mathbb{I}(\gamma_j=0)}\right)\right). \tag{56}$$

I推断 $\Phi_{k,\ j}$ 与前文相同，我们采用相互独立的狄利克雷先验分布作为$\Phi_{k,\ j}$ ，

$$\Phi_{k,j} \sim \text{Dirichlet}(\mathbf{a}_j), \tag{57}$$

其中，$\mathbf{a}_j = [a_{j,1}, \ldots, a_{j,R_j}]$ 是浓度参数向量。

由公式（56）可知，$\Phi_{k,\ j}$ 的条件分布为

$$p(\Phi_{k,j}|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_1, \ldots, z_n, \boldsymbol{\theta}, \boldsymbol{\Phi}_{-k,j}, \boldsymbol{\Phi}_0, \pi, \alpha, \gamma) \propto p(\Phi_{k,j}) \prod_{i=1}^{n}\prod_{j=1}^{J} \phi_{z_i,j,x_{ij}}^{\mathbb{I}(\gamma_j=1)}. \tag{58}$$

若 $\gamma_j = 1$，那么——与前文相同——$\Phi_{k,\ j}$ 的条件分布是$\Phi_{k,\ j}$ 的后验分布，该后验分布基于当前与成分 k相关联的所有观测值，即

$$\Phi_{k,j}|\gamma_j = 1, x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k},j}, \mathbf{a}_j \sim \text{Dirichlet}(\mathbf{a}_j + [s_{k,j,1}, s_{k,j,2}, \ldots, s_{k,j,R_j}]), \tag{59}$$

其中，$s_{k,\ j,\ r}$均如前所定义

若 $\gamma_j = 0$，那么$\Phi_{k,\ j}$ 的条件分布就是先验分布 $p(\Phi_{k,\ j})$，如公式（57）所示。[但请注意，在实际操作中，我们无需在这种情况下从条件分布中抽样]。

推断$\Phi_{0,\ j}$ 正如我们对$\Phi_{k,\ j}$所做的那样，我们为 $\Phi_{0,\ j}$采用狄利克雷先验分布，

$$\Phi_{0,j} \sim \text{Dirichlet}(\mathbf{a}_j^{(0)}), \tag{60}$$

其中 $\mathbf{a}_j^{(0)} = [a_{j,1}^{(0)}, \ldots, a_{j,R_j}^{(0)}]$ 是浓度参数向量。由公式（56）可清楚看出，$\Phi_{0,\ j}$ 的条件分布为

$$p(\Phi_{0,j}|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_1, \ldots, z_n, \boldsymbol{\theta}, \boldsymbol{\Phi}_{-k,j}, \boldsymbol{\Phi}_0, \pi, \alpha, \gamma) \propto p(\Phi_{0,j}) \prod_{i=1}^{n}\prod_{j=1}^{J} \phi_{0,j,x_{ij}}^{\mathbb{I}(\gamma_j=1)}. \tag{61}$$

如果 $\gamma_j = 1$，那么 $\Phi_{0,\ j}$ 的条件分布就是先验分布 $p(\Phi_{0,\ j})$，如公式（60）所示。[但请注意，在实践中，我们在这种情况下根本不需要从条件分布中采样]。

如果 $\gamma_j = 0$，那么 $\Phi_{0,\ j}$ 的条件分布是给定所有观测值时 $\Phi_{0,\ j}$ 的后验分布，即 .

$$\Phi_{0,j}|\gamma_j = 1, x_{1,j}, x_{2,j}, \ldots, x_{n,j}, \mathbf{a}_j^{(0)} \sim \text{Dirichlet}(\mathbf{a}_j^{(0)} + [s_{j,1}, s_{j,2}, \ldots, s_{j,R_j}]), \tag{62}$$

其中，$s_{j,\ r}$ 定义为第 j个变量属于类别 r 的观测值数量。

7.3.4 z_i 的条件分布

让我们暂时回到 K 有限的情形，并回顾公式（37），该公式给出了 z_i 的条件分布（假设对 θ_k 进行抽样，而非将其积分消去）。引入 γ_j（以实现变量选择），并利用给定 z_i 时变量之间的条件独立性，公式（37）变为:

$$p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\theta}, z_{-i}, \alpha, \gamma) = \frac{1}{Z}\frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha} f_y(y_i|\theta_k) \left(\prod_{j:\gamma_j=1} f_{\mathbf{x}}(x_{ij}|\Phi_{k,j})\right) \left(\prod_{j:\gamma_j=0} f_{\mathbf{x}}(x_{ij}|\Phi_{0,j})\right) \tag{63}$$

$$= \frac{1}{Z}\frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha} \theta_{k,y_i} \left(\prod_{j:\gamma_j=1} \phi_{k,j,x_{ij}}\right) \left(\prod_{j:\gamma_j=0} \phi_{0,j,x_{ij}}\right). \tag{64}$$

边缘化 θ_k 经过一些代数运算后，很容易证明，将分量特定参数 θ_k 边缘化后，z_i 的条件分布如下:

$$\begin{aligned} p(z_i = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}, \alpha) = \frac{1}{Z}\frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha} &\times \int_{\theta_k} f_y(y_i|\theta_k) p(\theta_k|\{y_m\}_{m\in\mathcal{M}_{-i,k}}) d\theta_k \\ &\times \prod_{j:\gamma_j=1} \left(\int_{\Phi_{k,j}} f_{\mathbf{x}}(x_{ij}|\Phi_{k,j}) p(\Phi_{k,j}|\{x_{mj}\}_{m\in\mathcal{M}_{-i,k}}) d\Phi_{k,j}\right) \\ &\times \prod_{j:\gamma_j=0} \left(\int_{\Phi_{0,j}} f_{\mathbf{x}}(x_{ij}|\Phi_{0,j}) p(\Phi_{0,j}|\{x_m\}_{m\in\mathcal{M}_{-i}}) d\Phi_{0,j}\right), \end{aligned} \tag{65}$$

其中

- $\mathcal{M}_{-i} = \{m:\ \ m \in \{1, \ldots, n\}$ 和 $m \neq i\}$；以及
- $\mathcal{M}_{-i,\ k} = \{m:\ \ m \in \{1, \ldots, n\}$ 和 $m \neq i$ 和 $z_m = k\}$。

因此，与之前一样，$p(\theta_k|\{y_m\}_{m\in\mathcal{M}_{-i,k}})$ 是 k 个成分中当前关联的观测值（不包括 y_i）对应的 θ_k 后验分布（$\Phi_{k,j}$ 也同理）。同样与之前一样，如果第 k 个

成分为空，则我们将 $p(\theta_k|\{y_m\}_{m\in\mathcal{M}_{-i,k}}) := p(\theta_k)$ 定义为 θ_k 的先验分布（$\Phi_{k,j}$ 也同理）。

请注意，涉及 $\Phi_{0,j}$ 的表达式略有不同：此处，我们得到的是在给定除第 i 个观测值之外的所有 观测值时，$\Phi_{0,j}$ 的后验分布。由于此表达式不涉及 k，我们可以将该项并入归一化常数 Z，于是：

$$p(z_i = k|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_{-i},\alpha) = \frac{1}{Z'}\frac{n_{-i,k}+\alpha/K}{n-1+\alpha} \times \int_{\theta_k} f_y(y_i|\theta_k)p(\theta_k|\{y_m\}_{m\in\mathcal{M}_{-i,k}})d\theta_k \\ \times \prod_{j:\gamma_j=1}\left(\int_{\Phi_{k,j}} f_{\mathbf{x}}(x_{ij}|\Phi_{k,j})p(\Phi_{k,j}|\{x_{mj}\}_{m\in\mathcal{M}_{-i,k}})d\Phi_{k,j}\right). \tag{66}$$

的条件分布 z_i (K **（无限）**与之前一样，我们考虑 $K\to\infty$ 时的极限。在这种情况下，z_i 的条件分布如下：

如果对于某个 $j\neq i$，$k=z_j$，则

$$p(z_i = k|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_{-i},\alpha) = b\frac{n_{-i,k}}{n-1+\alpha} \times \frac{\int_{\theta_k} f_y(y_i,\{y_m\}_{m\in\mathcal{M}_{-i,k}}|\theta_k)p(\theta_k)d\theta_k}{\int_{\theta_k} f_y(\{y_m\}_{m\in\mathcal{M}_{-i,k}}|\theta_k)p(\theta_k)d\theta_k} \\ \times \frac{\prod_{j:\gamma_j=1}\left(\int_{\Phi_{k,j}} f_{\mathbf{x}}(x_{ij},\{x_{mj}\}_{m\in\mathcal{M}_{-i,k}}|\Phi_{k,j})p(\Phi_{k,j})d\Phi_{k,j}\right)}{\prod_{j:\gamma_j=1}\left(\int_{\Phi_{k,j}} f_{\mathbf{x}}(\{x_{mj}\}_{m\in\mathcal{M}_{-i,k}}|\Phi_{k,j})p(\Phi_{k,j})d\Phi_{k,j}\right)}. \tag{67}$$

此外，

$$p(z_i\neq z_j \text{ for all } j\neq i|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_{-i},\alpha) = b\frac{\alpha}{n-1+\alpha} \times \int_{\theta_k} f_y(y_i|\theta_k)p(\theta_k)d\theta_k \\ \times \prod_{j:\gamma_j=1}\left(\int_{\Phi_{k,j}} f_{\mathbf{x}}(x_{ij}|\Phi_{k,j})p(\Phi_{k,j})d\Phi_{k,j}\right), \tag{68}$$

其中 b 是一个归一化常数，用以确保

$$p(z_i\neq z_j \text{ for all } j\neq i|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_{-i},\alpha) + \sum_{k\in\mathcal{C}_{-i}} p(z_i=k|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_{-i},\alpha) = 1.$$

请注意，这些表达式本质上与公式（43）和（44）中给出的相同，只是无关变量不会对条件分布产生影响（这正是我们所期望的）。边际似然可按前文所述进行计算（见第 7.2.1 节和第 7.2.2 节）。

7.3.5 采样 γ_j

通过考察公式（56），γ_j 的条件分布可立即得出

$$p(\gamma_j|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n,\boldsymbol{\theta},\boldsymbol{\Phi},\boldsymbol{\Phi}_0,\pi,\alpha,\gamma_{-j}) \propto p(\gamma_j)\left(\prod_{i=1}^{n}\phi_{z_i,j,x_{ij}}^{\mathbb{I}(\gamma_j=1)}\phi_{0,j,x_{ij}}^{\mathbb{I}(\gamma_j=0)}\right). \tag{69}$$

由于 $\gamma_j \in \{0,1\}$，我们有：

$$p(\gamma_j=1|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n,\boldsymbol{\theta},\boldsymbol{\Phi},\boldsymbol{\Phi}_0,\pi,\alpha,\gamma_{-j}) = \frac{1}{z}p(\gamma_j=1)\prod_{i=1}^{n}\phi_{z_i,j,x_{ij}}, \tag{70}$$

and

$$p(\gamma_j=0|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n,\boldsymbol{\theta},\boldsymbol{\Phi},\boldsymbol{\Phi}_0,\pi,\alpha,\gamma_{-j}) = \frac{1}{z}p(\gamma_j=0)\prod_{i=1}^{n}\phi_{0,j,x_{ij}}, \tag{71}$$

其中

$$z = p(\gamma_j=1)\prod_{i=1}^{n}\phi_{z_i,j,x_{ij}} + p(\gamma_j=0)\prod_{i=1}^{n}\phi_{0,j,x_{ij}}.$$

边缘化 $\Phi_{k,\ j}$ 和 $\Phi_{0,\ j}$ 与之前一样，我们还可以对$\Phi_{k,\ j}$ 和$\Phi_{0,\ j}$ 参数进行边缘化。定义 $X_k = \{\mathbf{x}^{i_1}$、$\mathbf{x}^{i_2}$、.....、$\mathbf{x}_{i_{n_k}}\}$ 为与成分 k相关联的观测值集合，并定义 $X_{k,\ j} = \{x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k},j}\}$ 为包含 j向量中第 j个元素的集合 X_k（同前）。另外，定义 $\mathcal{C} = \{k : z_i = k$ 对于某个 $i\}$ 为所有当前成分标签的集合，并定义 $X_{\cdot,\ j} = \cup_{k\in\mathcal{C}}\{X_{k,j}\}$ 为包含所有观测值中第 j 个元素的集合。

于是得到

$$p(\gamma_j=1|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n,\boldsymbol{\theta},\pi,\alpha,\gamma_{-j}) = \frac{1}{z'}p(\gamma_j=1)\prod_{k\in\mathcal{C}}p(X_{k,j}|\mathbf{a}_j), \tag{72}$$

and

$$p(\gamma_j=0|\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n,\boldsymbol{\theta},\pi,\alpha,\gamma_{-j}) = \frac{1}{z'}p(\gamma_j=0)p(X_{\cdot,j}|\mathbf{a}_j^{(0)}), \tag{73}$$

其中 z' 是一个归一化常数，用于确保概率之和为 1。边际似然可按照之前所述的方法进行计算（参见第 7.2.1 节）。

7.4 多视图聚类

上一节所述的变量选择方法可以视为多视图聚类的一个特例，其中我们允许数据具有多个聚类

结构（具体取决于我们认为哪一组变量是“相关的”）。在变量选择的情况下，我们将数据建模为具有 2 个聚类结构：一个非平凡的聚类结构（“相关”变量对此作出贡献），以及一个平凡的聚类结构（仅包含单个聚类，“不相关”变量对此作出贡献）。通过令 γ_j 成为一个类别型变量，使其取值表示第 j 个变量所贡献的聚类结构，可以将这一方法直接扩展到 L 个聚类结构。

7.4.1 符号约定

我们假设 $\gamma_j \in \{0, 1, 2, \ldots, V-1\}$ 是一个具有 L 个类别的分类变量。与前文一样，我们假设若 $\gamma_j = 0$，则第 j 个变量完全无关，不会对任何聚类结构产生贡献（除了仅包含一个簇的平凡结构）。

$$f(\mathbf{x}_i, y_i | \cdots) = \left(\prod_{j=1}^{J} \left(\prod_{v=1}^{V-1} f_{\mathbf{x}}(x_{ij} | \theta^{(v)}_{z_i^{(v)}})^{\mathbb{I}(\gamma_j = v)} \right) f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{z_i^{(1)}}, \mathbf{w}_i, \beta)$$

$$\begin{aligned} p(\gamma_j = 0 | \cdots) &\propto p(\gamma_j = 0) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_0) \\ p(\gamma_j = v | \cdots) &\propto p(\gamma_j = v) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta^{(v)}_{z_i^{(v)}}) \qquad \text{for } v = 1, \ldots, V-1 \end{aligned}$$

对于 $\ell = 1$，.....，$L-1$，我们假设存在互不重叠的变量集 $\mathcal{S}_\ell$，使得每个 $\mathcal{S}_\ell$ 都具有自己的聚类结构。在每个 $\mathcal{S}_\ell$ 内，我们使用混合模型对数据进行建模。为此，我们引入潜在成分分配变量 $z_i^{(\ell)}$，它们指示模型中负责生成 $\mathcal{S}_\ell$ 的第 i 个观测的成分。类似地，我们将 $\Phi_k^{(\ell)}$ 定义为与第 ℓ 个混合模型中的第 k 个成分相关联的参数。

此外，我们定义 $\gamma_j = \ell$，表示第 j 个变量是否属于 $\mathcal{S}_\ell$。请注意，先验地，我们并不知道哪些变量属于哪些 $\mathcal{S}_\ell$，因此必须对 γ_j 进行推断。

最后，我们引入另一个类别型变量 $\nu \in \{1, 2, \ldots, L-1\}$，使得 $\mathcal{S}_\nu$ 中存在的聚类结构与剖面回归相关，即 y_i 的似然，

$$f_{\mathbf{y}}(y_i|\boldsymbol{\theta}, z_i^{(1)}, \ldots, z_i^{(L-1)}, \nu) = f_{\mathbf{y}}(y|\boldsymbol{\theta}_{z_i^{(\ell)}}),$$

仅取决于 $z_i^{(\ell)}$ 成分分配变量。

O我们的条件模型为$\mathbf{x}_i = [\mathbf{x}_i,\ y_i]^\top$（给定 $z_i^{(\ell)}$、γ_j 和 ν）则为

$$\begin{aligned} &p(\mathbf{x}_i|z_i^{(1)}, \ldots, z_i^{(L-1)}, \boldsymbol{\theta}, \Phi_0, \Phi^{(1)}, \ldots, \Phi^{(L-1)}, \gamma, \nu) \\ &\qquad = f_{\mathbf{y}}(y|\boldsymbol{\theta}_{z_i^{(\ell)}}) \prod_{j=1}^{J} \left(f(x_{ij}|\phi_{0,j,x_{ij}})^{\mathbb{I}(\gamma_j=0)} \prod_{\ell=1}^{L-1} \left(f(x_{ij}|\phi^{(\ell)}_{z_i^{(\ell)},j,x_{ij}})^{\mathbb{I}(\gamma_j=\ell)} \right) \right). \end{aligned} \tag{74}$$

7.4.2 吉布斯抽样的条件分布

我们可以沿用第 3 节中提出的论证方法，推导所需的条件分布。为简洁起见，我们给出对分量特定参数积分消去后得到的条件分布。

条件分布 $z_i^{(\ell)}$：　$\nu = \ell$ 情形 在此情形下，我们处理的是与剖面回归相关的聚类结构，因此必须纳入响应变量 y 的贡献。我们有（参见公式 (67) 和 (68)）：若 $k = z_j^{(\ell)}$ 对某个 $j \neq i$，则

$$\begin{aligned} p(z_i^{(\ell)} = k|\mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}^{(\ell)}, \alpha^{(\ell)}, \nu = \ell) = b^{(\ell)} \frac{n_{-i,k}^{(\ell)}}{n - 1 + \alpha^{(\ell)}} \times \frac{\int_{\theta_k} f_y(y_i, \{y_m\}_{m \in \mathcal{M}_{-i,k}^{(\ell)}}|\theta_k) p(\theta_k) d\theta_k}{\int_{\theta_k} f_y(\{y_m\}_{m \in \mathcal{M}_{-i,k}^{(\ell)}}|\theta_k) p(\theta_k) d\theta_k} \\ \times \frac{\prod_{j:\gamma_j=1} \left(\int_{\Phi_{k,j}^{(\ell)}} f_{\mathbf{x}}(x_{ij}, \{x_{mj}\}_{m \in \mathcal{M}_{-i,k}^{(\ell)}}|\Phi_{k,j}^{(\ell)}) p(\Phi_{k,j}^{(\ell)}) d\Phi_{k,j}^{(\ell)} \right)}{\prod_{j:\gamma_j=1} \left(\int_{\Phi_{k,j}^{(\ell)}} f_{\mathbf{x}}(\{x_{mj}\}_{m \in \mathcal{M}_{-i,k}^{(\ell)}}|\Phi_{k,j}^{(\ell)}) p(\Phi_{k,j}^{(\ell)}) d\Phi_{k,j}^{(\ell)} \right)}, \end{aligned} \tag{75}$$

其中 $n_{-i,\,k}^{(\ell)}$ 是观测值的数量，$\mathbf{x}_m$，对于这些观测值，$z_m^{(\ell)} = k$（不包括第 i 个观测值），且 $\mathcal{M}_{-i,\,k}^{(\ell)} = \{m:\ m \in \{1, \ldots, n\}$ 和 $m \neq i$ 和 $z_m^{(\ell)} = k\}$。

此外　，

$$\begin{aligned} p(z_i^{(\ell)} \neq z_j^{(\ell)} \text{ for all } j \neq i|\mathbf{x}_1, \ldots, \mathbf{x}_n, c_{-i}^{(\ell)}, \alpha^{(\ell)}, \nu = \ell) = b^{(\ell)} \frac{\alpha^{(\ell)}}{n - 1 + \alpha^{(\ell)}} \times \int_{\theta_k} f_y(y_i|\theta_k) p(\theta_k) d\theta_k \\ \times \prod_{j:\gamma_j=1} \left(\int_{\Phi_{k,j}^{(\ell)}} f_{\mathbf{x}}(x_{ij}|\Phi_{k,j}^{(\ell)}) p(\Phi_{k,j}^{(\ell)}) d\Phi_{k,j}^{(\ell)} \right), \end{aligned} \tag{76}$$

其中 $b^{(\ell)}$ 是一个确保概率之和为 1 的常数，而 $\alpha^{(\ell)}$ 是第 ν 个混合模型中混合权重的狄利克雷过程先验分布的浓度参数。

关于 $z_i^{(\ell)}$ 的条件分布： $\nu \neq \ell$ **情形** 在此情形下，响应变量 y 不影响条件分布。我们有（参见公式 (67) 和 (68)）：如果对于某个 $j \neq i$，$k = z_j^{(\ell)}$，那么

$$p(z_i^{(\ell)} = k | \mathbf{x}_1, \ldots, \mathbf{x}_n, z_{-i}^{(\ell)}, \alpha^{(\ell)}, \nu \neq \ell) = b^{(\ell)} \frac{n_{-i,k}^{(\ell)}}{n - 1 + \alpha^{(\ell)}} \frac{\prod_{j:\gamma_j=\ell} \left(\int_{\Phi_{k,j}^{(\ell)}} f_{\mathbf{x}}(x_{ij}, \{x_{mj}\}_{m \in \mathcal{M}_{-i,k}} | \Phi_{k,j}^{(\ell)}) p(\Phi_{k,j}^{(\ell)}) d\Phi_{k,j}^{(\ell)} \right)}{\prod_{j:\gamma_j=\ell} \left(\int_{\Phi_{k,j}^{(\ell)}} f_{\mathbf{x}}(\{x_{mj}\}_{m \in \mathcal{M}_{-i,k}} | \Phi_{k,j}^{(\ell)}) p(\Phi_{k,j}^{(\ell)}) d\Phi_{k,j}^{(\ell)} \right)}. \tag{77}$$

其中，$n_{-i,\,k}^{(\ell)}$ 是满足 $z_m^{(\ell)} = k$ 的观测值 $\mathbf{x}_m$ 的数量（不包括第 i 个观测值）。此外，

$$p(z_i^{(\ell)} \neq z_j^{(\ell)} \text{ for all } j \neq i | \mathbf{x}_1, \ldots, \mathbf{x}_n, c_{-i}^{(\ell)}, \alpha^{(\ell)}, \nu \neq \ell) = b^{(\ell)} \frac{\alpha^{(\ell)}}{n - 1 + \alpha^{(\ell)}} \prod_{j:\gamma_j=\ell} \left(\int_{\Phi_{k,j}^{(\ell)}} f_{\mathbf{x}}(x_{ij} | \Phi_{k,j}^{(\ell)}) p(\Phi_{k,j}^{(\ell)}) d\Phi_{k,j}^{(\ell)} \right), \tag{78}$$

其中，$b^{(\ell)}$ 是确保概率之和为 1 的常数，而 $\alpha^{(\ell)}$ 是第 ℓ 个混合模型中混合权重的狄利克雷过程先验的浓度参数。

7.4.3 γ_j 的条件分布

定义 $X_k^{(\ell)} = \{\mathbf{x}_{i_1}、\mathbf{x}_{i_2}，\ldots\ldots，\mathbf{x}_{i_{n_k^{(\ell)}}}\}$ 为与第 ℓ 个混合模型的成分 k 相关联的观测值集合，并定义 $X_{k,\,j}^{(\ell)} = \{x_{i_1,j}, x_{i_2,j}, \ldots, x_{i_{n_k^{(\ell)}},j}\}$ 为包含向量 $X_k^{(\ell)}$ 中第 j 个元素的集合。还定义 $\mathcal{C}^{(\ell)} = \{k：$ 对于某个 $i\}$，$z_i^{(\ell)} = k$ 为第 ℓ 个混合模型的所有当前成分标签的集合。

我们有

$$p(\gamma_j = 0 | \ldots) = \frac{1}{z'} p(\gamma_j = 0) p(X_{\cdot,j} | \mathbf{a}_j^{(0)}), \tag{79}$$

并且，对于所有其他 $\ell \neq 0$,

$$p(\gamma_j = \ell | \ldots) = \frac{1}{z'} p(\gamma_j = \ell) \prod_{k \in \mathcal{C}^{(\ell)}} p(X_{k,j}^{(\ell)} | \mathbf{a}_j^{(\ell)}), \tag{80}$$

其中，z' 是确保概率总和为 1 的归一化常数。边际似然可按照前文所述的方法进行评估（参见第 7.2.1 节）。

7.4.4 条件分布 ν

ν的条件分布与 γ_j的条件分布非常相似。定义 $Y^{(\ell)}{}_k = \{y_{i_1}, y_{i_2}, \ldots, y_{i_{n_k^{(\ell)}}}\}$。那么，对于 $\ell \in \{1, \ldots, L-1\}$，有：

$$p(\nu = \ell | \ldots) = \frac{1}{z''} p(\nu = \ell) \prod_{k \in \mathcal{C}^{(\ell)}} p(Y_k^{(\ell)} | \mathbf{a}_y), \tag{81}$$

其中， z'' 是一个归一化常数，用于确保概率之和为 1 .

请注意，在对 ν 进行采样时，我们实际上是在比较对 y 进行划分的不同模型，并根据上述公式 (81) 给出的概率从中进行选择。

参考文献

Aldous、D. J.、I. A. Ibragimov、J. Jacod 和 P. L. Hennequin（编）。1985。《可交换性及相关主题》，柏林、海德堡。Springer Berlin Heidelberg。

Alon, U.、N. Barkai、D. A. Notterman、K. Gish、S.Ybarra、D. Mack 和 A.J. Levine, 1999年。通过对寡核苷酸芯片检测的肿瘤和正常结肠组织进行聚类分析，揭示基因表达的广泛模式。《美国国家科学院院刊》，96(12)：6745–6750。

Ben-Dor, A.、R. Shamir 和 Z. Yakhini

1999年。基因表达模式聚类。计算生物学杂志：计算与分子细胞生物学杂志，6(3-4)：281–297。

Blackwell, D. 和 J. B. MacQueen 1973年。通过P´olya urn方案构造Ferguson分布。统计年鉴，1(2)：353–355。癌症基因组图谱网络 2012年。人类乳腺肿瘤的全面分子图谱。《自然》，490(7418)：61–70。Crook, O. M.、L. Gatto 和 P. D. W. Kirk

2018. 狄利克雷过程混合模型中变量选择的快速近似推断及其在泛癌蛋白质组学中的应用。*arXiv.org*。

Cui, Y., X. Z. Fern, and J. G. Dy 2007. 通过正交化实现非冗余多视图聚类。载于第七届 *IEEE* 国际数据挖掘会议（*ICDM 2007*），第 133–142 页。IEEE。Curtis, C., S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gr¨af, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McK-inney, METABRIC 团队, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavare, C. Caldas, and S. Aparicio 2012. 对 2,000 个乳腺肿瘤的基因组和转录组架构进行分析，揭示了新的亚群。《自然》，486(7403):346–352。Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein 1998. 全基因组表达模式的聚类分析与展示。《美国国家科学院院刊》，95(25):14863–14868。Escobar, M. and M. West 1995. 使用混合模型进行贝叶斯密度估计与推断。*Journal of the AmericanStatistical Association*，90(430):577–588。Fop, M. and T. B. Murphy 2018. 基于模型的聚类中的变量选择方法。《统计学调查》，12:18–65。Fritsch, A. and K. Ickstadt 2009. 基于后验相似度矩阵的改进聚类准则。*BayesianAnalysis*，4(2):367–391。Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander 1999. 癌症的分子分类：通过基因表达监测进行类别发现和类别预测。《科学》（纽约州纽约市），286(5439):531–537。Green, P. J. and S. Richardson 2001. 使用和不使用狄利克雷过程对异质性建模。《斯堪的纳维亚统计学杂志》，28(2):355–375。

Guan, Y.、J. G. Dy、D. Niu 和 Z. Ghahramani，2010年。《非参数多重聚类的变分推断》。*MultiClust*研讨会。Hastie, D. I.、S. Liverani 和 S. Richardson，2014年。《从未知浓度参数的狄利克雷过程混合模型中抽样：大型数据实现中的混合问题》。*Statistics And Computing*，第1–15页。Heyer, L. J.、S. Kruglyak 和 S. Yooseph，1999年。《探索表达数据：共表达基因的识别与分析》。*Genomeresearch*，9(11)：1106–1115。Ishwaran, H. 和 L. F. James，2001年。《断棍先验的吉布斯抽样方法》。*Journal of the American Statistical Association*，96(453)：161–173。Ishwaran, H. 和 M. Zarepour，2002。《和狄利克雷过程的精确与近似表示》。*Can J Stat*，30(2)：269–283。Jain, S. 和 R. M. Neal，2004。《狄利克雷过程混合模型的分裂-合并马尔可夫链蒙特卡罗程序》。*Journal of Computational and Graphical Statistics*，13(1)：158–182。Jain, S. 和 R. M. Neal，2007年。《非共轭狄利克雷过程混合模型成分的拆分与合并》。*Bayesian Analysis*，2(3)：445–472。Kalli, M.、J. E. Griffin 和 S. G. Walker，2011。《混合模型的切片抽样》。*Statistics And Computing*，21(1)：93–105。Kirk, P.、J. E. Griffin、R. S. Savage、Z. Ghahramani 和 D. L. Wild，2012年。《用于整合多个数据集的贝叶斯相关聚类》。*Bioinformatics*，28(24)：3290–3297。

Kristensen, V. N.、O. C. Lingjærde、H. G. Russnes、H. K. M. Vollan、A. Frigessi 和 A.-L. Børresen-Dale，2014。《癌症整合基因组分析的原理与方法》。*Nature reviews. Cancer*，14(5)：299–313。Kuijjer, M. L.、J. N. Paulson、P. Salzman、W. Ding 和 J. Quackenbush，2018。《利用体细胞突变数据识别癌症亚型》。《***British Journal of Cancer***》，118(11)：1492–1501。Law, M. H.、A. K. Jain 和 M. Figueiredo，2003。《基于混合模型的聚类中的特征选择》。载于*Advances in Neural Information Pro-cessing Systems 15*，S. Becker、S. Thrun 和 K. Obermayer 编，第641–648页。MIT出版社。Law, M. H. C.、M. A. T. Figueiredo 和 A. K. Jain，2004。《使用混合模型同时进行特征选择与聚类》。*Pattern Analysisand Machine Intelligence, IEEE Transactions on*，26(9)：1154–1166。Li, D. 和 P. Shafto，2011。《贝叶斯层次交叉聚类》。载于第十四届人工智能与统计国际会议论文集，第443–451页，劳德代尔堡，佛罗里达州，美国。PMLR。Liverani, S.、D. I. Hastie、L. Azizi、M. Papathomas 和 S. Richardson，2015。《PReMiuM：使用狄利克雷过程进行剖面回归混合模型的 R 包》。《统计软件杂志》，64(7)。Lock, E. F. 和 D. B. Dunson，2013。《贝叶斯共识聚类》。*Bioinformatics*，第 btt425 页。McLachlan, G. J.、R. W. Bean 和 D. Peel，2002。《一种基于混合模型的微阵列表达数据聚类方法》。*Bioin-formatics (Oxford, England)*，18(3)：413–422。Miller, J. W. 和 M. T. Harrison

2017年。《具有成分数量先验的混合模型》。*Journal of the American Statistical Association*，113(521):340–356。Molitor, J., M. Papathomas, M. Jerrett, and S. Richardson 2010年。贝叶斯剖面回归及其在全国儿童健康调查中的应用。*Biostatistics (Oxford, England)*，11(3):484–498。Neal, R. M. 2000年。狄利克雷过程混合模型的马尔可夫链采样方法。*Journal ofComputational and Graphical Statistics*，9(2):249–265。Niu, D., J. G. Dy, and M. I. Jordan 2010年。多个非冗余谱聚类视图。载于第*27*届国际机器学习会议论文集，第831–838页，美国。Omnipress。Niu, D., J. G. Dy, and M. I. Jordan 2014年。迭代发现多个备选聚类视图。*Pattern Analysis and Machine Intelligence, IEEE Transactions on*，36(7):1340–1353。Papathomas, M., J. Molitor, C. Hoggart, D. Hastie, and S. Richardson 2012年。使用贝叶斯变量选择和狄利克雷过程探索遗传关联研究数据：在寻找基因—基因模式中的应用。*Genetic epidemiology*，36(6):663–674。Parker, J. S., M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard 2009年。基于内在亚型的乳腺癌监督式风险预测因子。*Journal of clinical oncology : official journal of the American Society of Clinical Oncology*，27(8):1160–1167。Perou, C. M., T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein 2000年。人类乳腺肿瘤的分子图谱。《自然》，406(6797):747–752。

Rasmussen, C. E. 2000年。无限高斯混合模型。载于神经信息处理系统进展，第*12*卷，第554–560页。MIT出版社。Richardson, S. and P. J. Green 1997年。关于成分数量未知的混合模型贝叶斯分析（含讨论）。*Journal Of The Royal Statistical Society Series B-Statistical Methodology*，59(4):731–792。Savage, R. S., Z. Ghahramani, J. E. Griffin, P. Kirk, and D. L. Wild 2013。通过结合基因组、转录组和表观基因组数据识别胶质母细胞瘤的癌症亚型。*arXiv.org*。Shahbaba, B. and R. Neal 2009年。使用狄利克雷过程混合模型的非线性模型。机器学习研究期刊（*JMLR*），10:1829–1850。Shen, R., A. B. Olshen, and M. Ladanyi 2009年。使用联合潜变量模型整合聚类多种基因组数据类型，并应用于乳腺癌和肺癌亚型分析。*Bioinformatics*，25(22):2906–2912。Son, C. G., S. Bilke, S. Davis, B. T. Greer, J. S. Wei, C. C. Whiteford, Q.-R. Chen, N. Cenacchi, and J. Khan 2005年。多种人体器官的mRNA基因表达谱数据库。*Genome research*，15(3):443–450。Sørlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale 2001年。乳腺癌的基因表达模式可区分具有临床意义的肿瘤亚类。《美国国家科学院院刊》，98(19):10869–10874。Suchard, M. A., Q. Wang, C. Chan, J. Frelinger, A. Cron, and M. West 2010年。理解用于统计计算的GPU编程：大规模并行海量混合模型研究。《计算与图形统计学杂志》，19(2):419–438。

Tadesse, M. G., N. Sha 和 M. Vannucci，2005年。《高维数据聚类中的贝叶斯变量选择》。*Journal of the American Statistical Association*，100(470):602–617。Taschler, B., F. Dondelinger 和 S. Mukherjee，2019年。《通过自适应投影实现超高维数据的基于模型的聚类》。*arXiv.org*。癌症基因组图谱，2012年。《人类乳腺肿瘤的全面分子图谱》。《自然》，487(7407):61–70。Walker, S. G.，2007年。《利用切片对狄利克雷混合模型进行采样》。统计学与计算模拟通讯，36(1):45–54。Witten, D. M. 和 R. Tibshirani，2010年。《聚类中特征选择的框架》。*Journal of the American StatisticalAssociation*，105(490):713–726。PMID：20811510。Yeung, K. Y. 和 W. L. Ruzzo，2001年。《用于基因表达数据聚类的主成分分析》。*Bioinformatics*（英国牛津），17(9):763–774。