PHM-ENSEMBLE：通过集成密度估计实现可扩展且稳定的概率多分辨率聚类

Ali Turfah
密歇根大学生物统计学系，
美国密歇根州安娜堡，邮编
48109 aturfah@umich.edu

Xiaoquan Wen
密歇根大学生物统计学系，
美国密歇根州安娜堡，邮编
48109 xwen@umich.edu

arXiv:2611.32945v1 [stat.ML] 24 Nov 2026

摘要

聚类是各个科学领域中探索性数据分析的基础工具。我们认为，聚类方法应从四个维度进行评估：对数据底层结构的保真度、所识别群组的可解释性、面对数据扰动和算法随机性时的稳定性，以及对现代科学数据集的可扩展性。然而，现有方法很少能同时满足这四项标准。为弥补这一不足，我们提出 PHM-ensemble，这是一种用于多分辨率聚类的通用概率框架：它通过汇集一组拟合结果中的混合模型来估计底层数据密度，随后通过对混合成分进行层次合并来构建聚类层次结构。在模拟数据集和 MNIST 基准上，PHM-ensemble 在与真实值的一致性方面达到或优于广泛使用的聚类算法，在不同运行中产生更稳定的结果，并揭示复杂的底层结构。将该方法应用于人类免疫健康图谱中的 182 万个外周血细胞时，PHM-ensemble 与专家整理的主要细胞类型注释的一致性高于原研究中使用的 Leiden聚类（调整兰德指数为 0.965，而非 0.789），且所需计算时间仅为其一半。除了识别主要免疫谱系外，该方法还揭示了细胞类型之间具有生物学一致性的关系，并解析了各个谱系内部更精细的结构，发现了与已知生物学相符的亚型和细胞状态。

关键词 无监督学习 · 聚类分析 · 多分辨率聚类 · 高斯混合模型 · 集成学习 · 单细胞RNA测序

1 引言

聚类是无监督数据分析中广泛使用的计算工具，使研究人员能够在不了解底层组标签的情况下，揭示数据中的结构。它对于探索性数据分析尤其有用，并已广泛应用于众多科学领域。在基因组学中，聚类通常用于根据单细胞基因表达数据识别细胞类型；在天文学中，用于根据
光度和光谱测量对星系和恒星进行分类；在经济学中，用于细分市场和消费者群体；在社会科学中，用于识别社交网络中的社区。然而，随着这些领域的数据集在规模和复杂性方面不断增长，现有聚类方法面临着日益严峻的挑战。要满足当代科学应用的需求，就需要开发能够同时提升保真度、可解释性、稳定性和可扩展性的新型聚类方法。

保真度和可解释性密切相关，分别关乎聚类结果的质量与表示方式。我们用保真度来指输出反映数据真实内在结构的程度，也就是说，聚类是否忠实地捕捉了数据中的模式。相比之下，可解释性关注的是，研究人员能否理解并有意义地诠释所恢复的聚类结构：各个聚类是否清晰区分，是否对应已知的协变量或领域知识，以及聚类结果是否易于解释。虽然

保真度本身至关重要，但当真实值不可用时，可解释性就变得尤为关键，而这在无监督学习中十分常见。许多现有的聚类方法通过对数据进行单一的平面划分来表示其内在结构。然而，在许多科学应用中，内在的群体结构源于连续的分化过程，且本质上具有层次性。将这种复杂结构压缩为单一的平面划分，可能会损害保真度和可解释性。多分辨率表示通过保留多个尺度上的结构，克服了这些局限，使研究人员能够针对特定科学问题选择合适的粒度，并揭示不同分辨率下群体与子群体之间的关系。

稳定性最初被定义为确定性算法的输出对输入数据扰动的稳健性 [1], ，后来这一概念扩展到随机化算法 [2] ，以考虑算法自身随机性所带来的变异。稳定性的这两个来源都与聚类直接相关。长期以来，人们一直认为，对数据扰动的稳定性是一项重要的验证标准 [3, 4, 5]：如果恢复出的结构是真实的，而非特定样本造成的假象，那么它就应当能在不同但同样有效的样本中重现。同时，许多聚类方法包含会引入额外变异性的随机分量。例如， k-均值算法和用于拟合高斯混合模型（GMMs）的 EM算法，在同一数据集上可能产生不同结果，因为不同的随机初始化可能导致不同的局部最优解。显著的不稳定性意味着，恢复出的结构可能反映的是抽样变异或算法随机性，而非数据中的真实模式，从而损害保真度和可解释性。因此，我们认为稳定性是可靠聚类的一项基本要求，尤其是在科学应用中。

随着样本量和复杂度迅速增长，现代科学数据集的规模已超出许多现有聚类方法的计算能力。单细胞图谱可能包含数千万个细胞 [6], ，而天文巡天则可能编录数亿个天体 [7]。许多具有坚实统计基础的聚类方法，包括经典层次聚类和贝叶斯非参数方法，扩展性较差，处理大规模数据集时可能面临难以承受的计算开销。更根本地说，如果一种聚类方法无法处理现代科学应用中常见的数据规模，那么无论其保真度、可解释性或稳定性如何，其实际价值都有限。因此，可扩展性是发挥保真度、可解释性和稳定性优势、将其应用于现代科学应用的实际前提。

鉴于对保真度、可解释性、稳定性和可扩展性的需求，我们提出了一种用于概率多分辨率聚类的新型通用计算框架。该框架融合了统计推断和表示学习的思想。它突破了将聚类视为直接在观测样本上定义划分问题的传统观点，转而着眼于学习潜在的数据生成分布，并通过构建聚类层次结构来解读其结构。这种以分布为核心的视角，使我们能够利用密度估计和集成学习等成熟的机器学习技术，在降低大规模数据集计算负担的同时，提升聚类结果的保真度和稳定性。在此前关于 P_{mc} 层次合并（PHM）算法 [8], 的工作的基础上，我们提出直接根据估计得到的分布构建聚类层次结构，从而提供一种多分辨率表示，揭示不同尺度上的结构，而非预先限定为单一分辨率，进而提升可解释性。

本文其余部分的组织结构如下。第2节介绍所提出聚类框架的理论依据与方法，并分析其计算复杂度。第3节使用模拟数据集和真实基准数据集评估所提出的方法，包括 MNIST 和图谱规模的单细胞 RNA 测序数据集。第4节通过讨论主要发现与未来方向作出总结。

2 方法

2.1 作为概率表示学习的聚类

我们将聚类视为一项表示学习任务 [9]。在这一设定下，我们关注的表示并不是学习嵌入或特征空间，而是作为观测数据基础的数据生成分布。具体而言，目标是估计该分布，并通过聚类结构对其进行解释。

令 $\boldsymbol{x} \in \mathbb{R}^d$ 表示由未知分布 $P(\boldsymbol{x})$ 生成的 d 维观测值。我们的目标是使用高斯混合模型（GMM）从观测数据中学习该分布。当分量数量趋于无穷时，GMM 是密度函数的通用逼近器，能够以任意精度逼近 $P(\boldsymbol{x})$ [10, 11]；然而在实际应用中，我们使用的是有限的、由数据驱动的近似。我们将学习到的分布记为 $\hat{}P(\boldsymbol{x})$，并假设其由 K 个混合成分组成：

$$\hat{P}(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

其中， α_k 是第 k 个分量的混合比例，而 $\mathcal{N}(\boldsymbol{x};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 表示一个多元高斯密度，其估计均值向量为 $\boldsymbol{\mu}_k$ ，协方差矩阵为 $\boldsymbol{\Sigma}_k$。

现有的基于高斯混合模型的聚类方法通常会将混合成分直接解释为聚类 [12]。然而，基于模型的聚类文献早已认识到，这种一一对应关系并非必然成立：单个高斯成分通常过于受限，难以表示非高斯子群体；而贝叶斯信息准则等模型选择准则会在拟合优度与模型复杂度之间进行权衡，因此可能选择多于有意义聚类数量的成分[13, 14]。我们采用这一更灵活的观点，将总体数据分布建模为高斯混合，其中每个潜在簇本身可能包含多个高斯成分。在此表述下，聚类就成为在估计得到的分布 $\hat{}P(\boldsymbol{x})$ 中识别并分组不可分离或高度重叠的混合成分的任务。

这一视角自然引出了聚类的多分辨率视角，该视角已经构成现有方法的基础：这些方法通过组合混合成分来表示有意义的聚类 [13, 14]。更具体地说，任何有限混合分布都可以通过递归聚合其原子混合成分而得到层次表示。这直接源于凸组合的结合律：它使得成分子集能够表示为“宏观密度”，随后在更高层次的抽象层面上进行分组，而无需在任何聚合阶段重新估计密度。由此形成的层次结构是对总体在多种分辨率下的嵌套表示：从根节点处的全局分布（即单个聚类），逐步延伸至叶节点处越来越局部化的子群体，同时保留底层概率分布。然而，这种层次结构通常并不唯一，因为不同的聚合策略可能产生显著不同的聚类结构。因此，该框架的核心挑战在于量化混合成分之间的重叠程度，并决定它们应按何种顺序进行聚合。尽管已有研究提出了基于熵和基于误分类的方法来合并混合成分 [13, 14], ，但这些方法计算成本高，且难以扩展到大型数据集。

总而言之，我们提出的计算框架并未将聚类视为直接针对观测样本进行算法式划分的问题，而是将其视为一项统计推断任务：估计底层的数据生成分布，并识别其中的自然分组。具体而言，我们将聚类分解为两个连续阶段：（i）根据观测数据估计底层概率分布；（ii）构建聚类层次结构，以揭示所学习分布在多种分辨率下的自然分组。这种分解使我们能够在每个阶段分别应对大规模数据聚类的挑战，并采用针对相应任务定制的计算技术。重要的是，所提出的框架并未摒弃现有的基于划分的聚类方法。相反，它可以将这些方法作为归纳偏置融入密度估计中，以提升估计效果，详见第2.2节。

2.2 集成密度估计

期望最大化（EM）算法已被广泛用于拟合高斯混合模型，以进行密度估计 [15], ，其中包括广受欢迎的 mclust 软件包 [16]。然而，将标准 EM 算法应用于大规模数据集时，会面临若干显著挑战。首先，通过 EM 进行高斯混合模型估计对初始化十分敏感，容易陷入局部最优解，并且依赖于分量数量的选择 K [17, 18, 12]。这些敏感性可能导致密度估计不稳定，并在我们提出的框架中进一步造成聚类结果退化。其次，使用 EM 拟合一个含 K 个分量的高斯混合模型，对 N 个样本在 $\mathbb{R}^d$ 中进行拟合时，其每次迭代的计算复杂度为 $O(NKd^2)$，这给大规模数据集的可扩展性带来了重大挑战。

为解决这些问题，我们提出采用集成学习策略：最终密度估计不再依赖于单个拟合的高斯混合模型，而是通过汇总一组高斯混合模型拟合结果中的估计值构建。具体而言，我们将集成密度估计记为

$$\hat{P}_e(\boldsymbol{x}) = \sum_{m=1}^{M} w_m \hat{P}_m(\boldsymbol{x}), \tag{2}$$

其中

$$\hat{P}_m(\boldsymbol{x}) = \sum_{k=1}^{K_m} \alpha_{mk}\, \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{mk}, \boldsymbol{\Sigma}_{mk}).$$

集成权重向量 $\boldsymbol{w} = (w_1、\ .....、\ w_M)$ 位于概率单纯形上，表示集合中相应高斯混合模型拟合结果的相对重要性。本文主要关注一种简单方案：计算最终密度估计时，为所有高斯混合模型拟合结果赋予相同权重，即令 $w_m = 1/M$ for $m = 1、.....、M$。我们的软件还支持其他更通用的加权方案（见《补充说明》第 1.3 节）。每个重复的密度估计 $\hat{}\ P_m(\boldsymbol{x})$，都是通过对独立构造的子样本和初始化应用标准高斯混合 EM算法得到的。具体而言，对于每次重复，我们首先对子样本数据进行子采样，然后在该子样本上构造分区引导初始化，最后通过 EM算法拟合一个高斯混合模型。因此，所得集成

捕捉了两种不同的变异性来源：子采样在降低对完整数据集拟合 EM算法的计算负担的同时，也会扰动观测数据；而初始化的多样性则能减轻 EM算法对非凸似然面局部最优解的敏感性 [17]。

子采样由用户指定的比例 $\rho \in (0, 1]$ 控制。对于每个重复样本 m，我们从大小为 N 的完整数据集中抽取 $\lfloor \rho N \rfloor$ 个观测值，不放回；各重复样本之间相互独立。这种方案称为子采样聚合（即子采样聚合），与在完整样本量 N 下进行有放回自助法重采样（即装袋法）相比，计算成本更低 [19]。尽管多次重复的子采样不一定会降低总体空间复杂度，但它可以并行执行，从而缩短密度估计的实际运行时间。
给定第 m 个子样本，接下来我们对其中的点应用基础划分算法。随后可以通过两种不同方式将所得划分纳入基于 EM 算法的密度估计。第一种方法称为全局密度估计，即直接对整个子样本拟合单个高斯混合模型，并利用划分信息初始化其关键参数。该方法直接沿用 mclust [16] 和 scikit-learn [20] 所采用的 EM 算法初始化策略。第二种方法称为局部密度估计，采用分治策略：使用加权期望最大化 [16, 21], , 分别对与各个划分相关联的子样本概率加权子集拟合高斯混合模型；随后将这些局部的、分区特异性密度估计合并为该重复样本的单一估计。概率加权方案会考虑划分边界附近点分配的不确定性，从而将这种不确定性纳入每个划分的局部密度估计中[8]。在这两种方法中，全局密度估计已广为采用；在这种情况下，基于质心的算法以及采用 Ward连接法的层次聚类预计尤其适合作为基础划分方法，因为它们产生的划分天然符合高斯混合假设 [22, 23, 24, 25]。相比之下，局部密度估计更加灵活，可与任何基础划分方法结合使用，包括基于图的方法和基于密度的方法。不过，目前尚未就单一的最优概率加权方案达成共识。

为了使不同重复样本中的 EM 估计具有多样性，我们可以使用同一种基础划分算法（例如 k-均值）的随机起始值，或者在不同重复样本中改变基础划分算法的选择。使用多种初始化方式可以促使不同重复样本探索似然面的不同区域。因此，整个集成对任何单个重复样本收敛到无代表性的局部最优解的敏感性都会降低。

关于集成密度估计过程实现细节的说明，见《补充说明》第 1 节。

2.3 聚类层次结构的构建

从估计得到的高斯成分构建聚类层次结构，为在多种分辨率下系统地识别、合并密度成分并将其转化为聚类结构提供了一种方法。此阶段的核心计算挑战在于确定应如何合并高斯成分以及高斯混合成分。文献 [14, 13] 中已经提出了各种解决方案；然而，就计算效率、可解释性以及对大规模数据集的适用性而言，没有一种方案完全令人满意。
从概念上讲，我们认为误分类概率是对由高斯成分构成的聚类之间的分离度进行形式化描述的自然方式。在现有文献中，误分类概率通常被定义为给定聚类配置下，采用 0-1损失的最优分类器所对应的贝叶斯风险，其中，最优分类器是这样一种确定性决策规则：将每个数据点分配给后验分类概率最高的聚类。它通过衡量从任意聚类生成的数据点被分配到其他聚类的总体概率，来量化聚类分离度。因此，较小的误分类概率表示聚类之间具有良好的分离性，而较大的误分类概率则反映出某些聚类之间存在 substantial overlap，且可区分性较差。然而，在实践中，这一量通常难以计算，因而成为我们当前场景实现可扩展性的主要障碍。巴塔查里亚距离可用于推导两个聚类之间误分类概率的上界。不过，它只有在两个高斯分布组成的聚类之间才具有闭式表达，无法直接推广到两个高斯混合之间 [14]。

最近，我们研究了一种新的误分类概率，记为 P_{mc}。该概率基于 0-1损失，由一种替代性的随机分类器诱导；该分类器通过从相应的后验分类概率 [8] 中进行抽样，将每个数据点分配到某个聚类中。这种随机分类器广泛用于序贯决策和强化学习中的玻尔兹曼探索 [26, 27]。尽管从这样的意义上说，该随机分类器并非最优：对于相同的聚类配置，它得到的误分类概率不小于最优确定性分类器的误分类概率，但我们发现，在衡量高斯成分以及高斯成分混合分布之间的分离程度时，这两个量在数值上非常接近，尤其是在部分重叠和充分分离的情形下。此外，在我们提出的聚类框架中，分类器是隐式且固定的，而误分类概率仅用于量化不同聚类配置之间的分离程度，因此没有必要与最优分类器进行比较。

与我们的问题最为相关的是， P_{mc} 具有一项特别有用的计算性质：它可以进行成对分解，即

$$P_{\mathrm{mc}} = \sum_{i<j} \Delta P_{\mathrm{mc}}^{(i,j)}. \tag{3}$$

这里， $\Delta P_{\mathrm{mc}}^{(i,j)}$表示合并现有聚类 i 和 j所带来的整体 P_{mc} 的减少。对于一般的聚类生成分布，这种分解均成立，包括高斯分布和高斯混合分布（详情参见 [8] 中的命题1）。此外，一旦将聚类 i 和 j 合并为新聚类 $i \oplus j$，新聚类与任一现有聚类 k 之间的成对贡献便可方便地更新为

$$\Delta P_{\mathrm{mc}}^{(i\oplus j,k)} = \Delta P_{\mathrm{mc}}^{(i,k)} + \Delta P_{\mathrm{mc}}^{(j,k)}. \tag{4}$$

为了根据集成密度估计构建聚类层次结构，一种自然的做法是定义一个基于 $\{\Delta P_{\mathrm{mc}}^{(i,j)}\}$的连接函数，用于在每次凝聚迭代中评估并排序当前所有聚类对之间的分离程度。

在最简单的情况下，连接函数

$$d(i,j) = -\log \Delta P_{\mathrm{mc}}^{(i,j)} \tag{5}$$

会生成一种聚类层次结构，在每一步合并时贪心地降低整体 P_{mc}。然而，在我们的设定中，层次结构是根据集成密度估计构建的，其中可能包含许多高度相似、权重较低的密度成分；当独立拟合的重复样本合并在一起时，这种情况自然会出现。由于

$$\Delta P_{\mathrm{mc}}^{(i,j)} \propto \alpha_i \alpha_j, \tag{6}$$

原始连接函数 $d(i,\ j)$ 系统性地不利于这些恰好冗余的配对：较小的 $\alpha_i\alpha_j$ 会抑制 $\Delta P_{\mathrm{mc}}^{(i,j)}$，无论分量 i 和 j 的重叠程度有多高，从而延迟它们的合并。为解决这一问题，我们对$\Delta P_{\mathrm{mc}}^{(i,j)}$进行重新缩放，以消除其对混合比例的依赖，使近重复分量能够因为其本身是近重复分量而较早合并，而不是因为其权重所致。更具体地说，我们定义

$$\Delta P_{\mathrm{mc}}^{*(i,j)} = \frac{\Delta P_{\mathrm{mc}}^{(i,j)}}{\alpha_i \alpha_j}, \tag{7}$$

and

$$d^*(i,j) = \log\left(1 + \frac{1}{\Delta P_{\mathrm{mc}}^{*(i,j)}}\right). \tag{8}$$

缩放后的量 $\Delta P_{\mathrm{mc}}^{*(i,j)}$是聚类 i和 j [8]中密度成分之间巴塔查里亚系数的上界。我们将基于由 ΔP_{mc} 导出的连接函数构建的凝聚算法称为 P_{mc} 层次合并（PHM）算法。默认情况下，我们使用公式（8）作为连接函数，从集成密度估计中构建聚类层次结构。

构建的层次结构为簇分析提供了有价值的信息。与扁平表示相比，它能够自然地揭示嵌套聚类结构，并允许用户比较多种分辨率下的分组模式。包括热图和树状图在内的可视化工具，还能进一步帮助用户探索这些多分辨率模式。最后，尽管我们不主张将单一扁平聚类作为默认表示形式，但在需要时，可以通过对 P_{mc}应用阈值来获得这种表示。

2.4 计算复杂度

集成密度估计随子样本大小 ρN 线性扩展。由于 M 重复样本基于独立子样本和初始化构建，因此易于并行处理 [28]。通过 EM算法完成模型拟合后，只需保留 K 合并后原子分量的参数；无需保留子样本或其责任矩阵。因此，集成密度估计的持久内存需求与 N 无关。

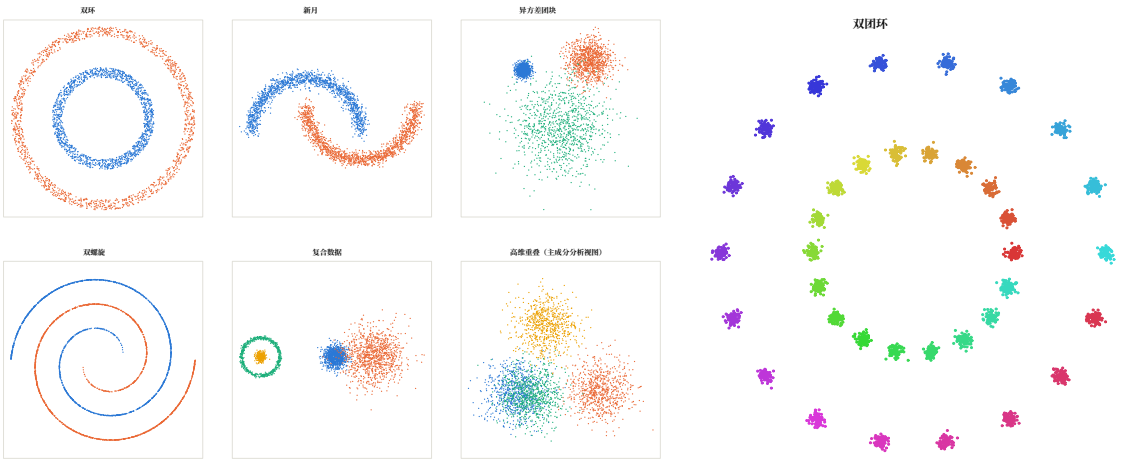
对于簇层次结构构建，评估成对的 ΔP_{mc} 值和执行 PHM 合并的时间复杂度分别为 $O(SKd^2 + SK^2)$ 和 $O(K^2)$，其中 S 是固定的蒙特卡洛采样预算。当 K 较大时，我们会应用预处理过程，识别并合并高度重叠的高斯成分，从而减少原子分量的数量 K，并进一步降低计算成本。PHM 算法的持久空间复杂度为 $O(K^2)$，与 N完全解耦。相比之下，先前的合并方法 [13, 14] 必须在整个合并过程中始终保留一个大小为 N的责任矩阵，导致 $O(NK^2)$的时间复杂度和 $O(NK)$的空间复杂度。由于 $N \gg K$在大规模数据集（例如图谱规模的单细胞 RNA 测序数据）中，所提出的 PHM 算法相比现有方法成为一种效率和可扩展性显著更高的替代方案。

3 结果

3.1 合成数据示例

我们在一系列合成数据集上展示 PHM-ensemble，以评估其保真度、可解释性、稳定性和可扩展性，并将其性能与现有聚类方法进行基准比较。

第一组合成数据集由六个具有挑战性的聚类问题组成（图1A），旨在突出各种现有聚类方法的局限性（补充说明第2.1节）。前五个数据集包含二维点，其聚类要么呈现非凸形状（即双环、新月和双螺旋），要么具有不同程度的簇内异质性（即异方差团块），或者同时具备这两种特征（即“复合数据”）。第六个数据集由60维空间中的点组成，这些点形成具有不同方差和协方差结构的重叠簇。每个数据集包含3,000个点，并具有相对较高的信噪比，因为人眼可以清晰地观察到其中的聚类模式。我们使用 $M = 5$和 $\rho = 1.0$ 对全部六个数据集应用PHM-ensemble程序。在每次集成重复中，我们执行全局密度估计，最多允许使用100个高斯成分，并使用mclust中实现的随机层次聚类程序初始化EM算法。作为比较，我们还应用了七种常用的聚类方法： k-均值、采用Ward方法和单链接的层次聚类、使用mclust的基于模型的聚类、使用HDBSCAN和OPTICS的基于密度的聚类，以及使用Leiden方法的社区检测。为评估聚类保真度，我们将每个数据集的聚类数固定为真实数量，并计算每种方法得到的聚类分配与已知真实标签之间的调整兰德指数（ARI）。结果汇总于表1。PHM-ensemble和Leiden是仅有的两种在全部六个数据集上始终保持较高调整兰德指数的方法，表明它们能够在一系列具有挑战性的场景中忠实地恢复潜在的聚类结构。此外，在保持 $M = 5$ 不变的情况下，当子采样比例 ρ 降低至0.75、0.50和0.25时，PHM-ensemble的性能几乎没有变化（补充表1）。相比之下，其余每种方法至少在一个数据集上表现不佳，凸显了它们对
特定类型的聚类结构。例如，mclust、 k-means 和采用 Ward连接法的层次聚类，都无法恢复双环、新月、双螺旋和复合数据集中的非凸聚类结构。另一方面，基于密度的方法 HDBSCAN 和 OPTICS 即使面对适度的维度增加也难以应对：尽管所有方法都仅使用 60 维数据中前四个主成分（PC）来处理“高维重叠”数据集（这些主成分通过碎石检验选出），但随着维度增加，局部密度对比会减弱，导致这些方法的性能显著下降。这是基于密度的聚类方法面临的一个众所周知的根本性挑战。



(A) (B)

图1：**用于评估聚类性能的合成数据集** (A) 六个旨在挑战现有聚类方法的合成数据集。(B) 一个具有清晰多分辨率结构的合成数据集：在较粗分辨率下，由两个间隔明显的环组成；在较细分辨率下，每个环又进一步划分为 18 个由点构成的独立团。

表 1：八种聚类方法在六个合成数据集上的调整兰德指数（ARI）。

方法	双环	新月	异方差团块	双螺旋	复合数据	高维重叠
k均值	0.001	0.257	0.844	0.029	0.662	0.955
层次聚类（Ward法）	0.000	0.191	0.942	0.037	0.693	0.909
层次聚类（单链接法）	1.000	1.000	0.000	1.000	0.474	0.000
mclust	0.000	0.503	0.961	0.046	0.776	0.956
OPTICS	0.000	0.995	0.928	0.977	0.500	0.006
HDBSCAN	1.000	0.965	0.807	0.964	0.785	0.230
Leiden	1.000	1.000	0.950	1.000	0.940	0.954
PHM-ensemble	1.000	1.000	0.961	1.000	0.938	0.957

接下来，我们使用另一个模拟数据集来评估可解释性，该数据集称为“双环团簇”，由3,000个具有清晰多分辨率结构的二维数据点组成（图1B）。在较粗分辨率下，该数据集由两个彼此分离的环组成；而在较细分辨率下，每个环又进一步划分为18个不同的点团。使用 $M = 5$ 和 $\rho = 1.0$得到的PHM-ensemble结果正确恢复了这两个层次的结构，并且该结构可通过所得的树状图和热图直接呈现且易于解释（图2）。将子采样比例 ρ 降至0.5或0.1后，结果几乎完全相同。相比之下，Leiden在分辨率参数的几乎整个范围内都恢复出36个独立团，但未能在任何分辨率下识别出双环结构。这可能是因为所构建的图中没有连接不同团的边，从而无论选择何种分辨率参数，基于模块度的社区检测方法都无法将它们合并。

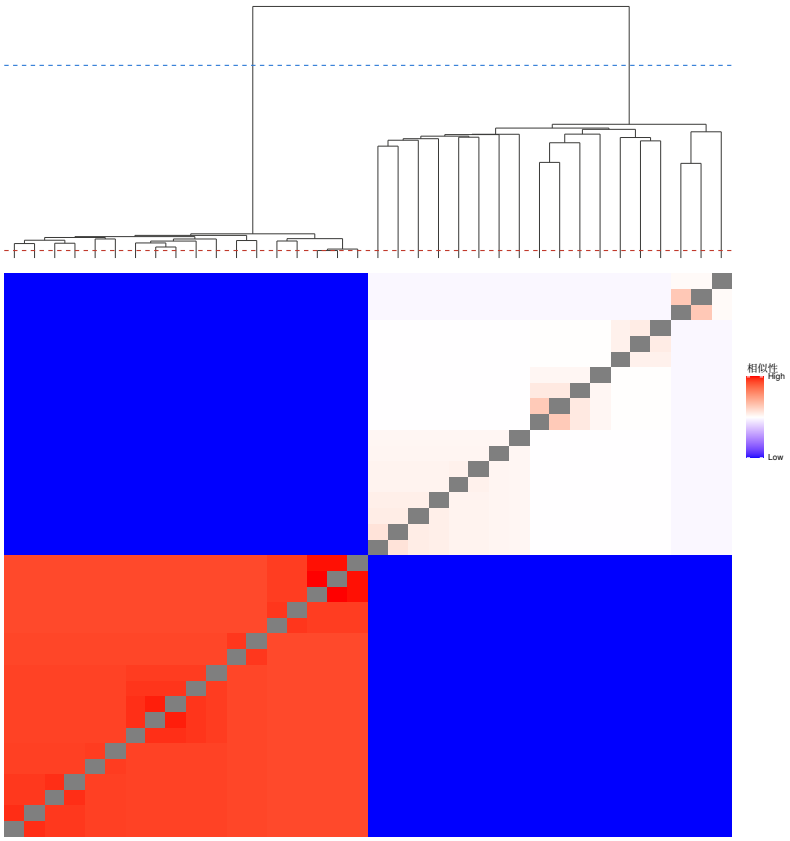
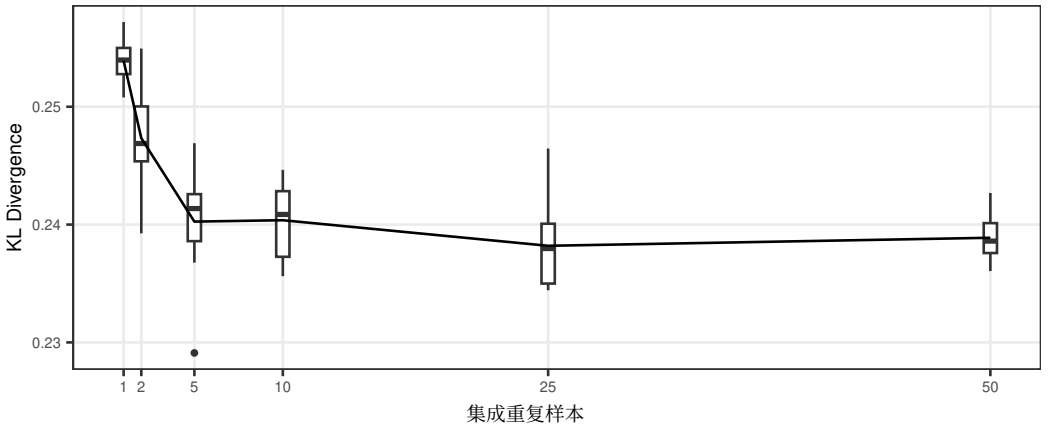
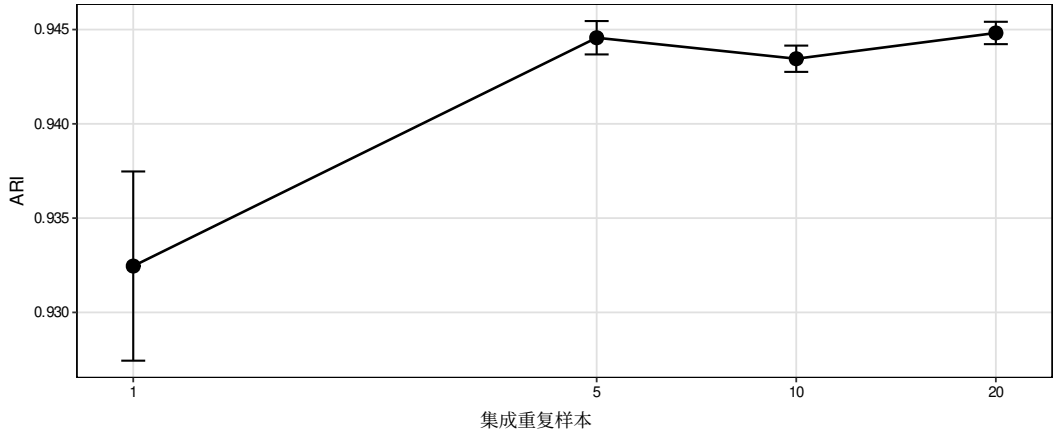


图2：**PHM-ensemble 在双环团数据集中恢复多分辨率结构**使用 $M = 5$ 和 $\rho = 1.0$得到的 PHM-ensemble 结果，正确恢复了较粗分辨率下的双环结构（蓝色点线），以及较细分辨率下的 36个独立团（红色点线）。生成的树状图和热图直观且易于解读地呈现了多分辨率结构。内环中的团彼此比外环中的团更接近，这一点在树状图中体现为内环的分支长度更短。

为了评估 PHM-ensemble 的稳定性及其集成密度估计程序的效果，我们首先关注第一组合成数据集中的双环数据，其中模拟数据点均匀分布在两个环带上。我们通过改变重复样本数量（$M = 1$，2、5、10、25、50），比较均匀分布与密度估计之间的 KL散度，并采用全局和局部密度估计方法。对于全局密度估计，每个重复样本使用 24 个高斯成分。对于局部密度估计，我们使用 k-均值和 $K = 6$ 个划分初始化每个重复样本，并允许使用最多四个高斯成分来估计每个分区特异性密度。在每种估计设置下，我们重复实验 10 次，以量化所得 KL散度的变异性。对于这两种密度估计程序，我们观察到，随着 M 增加，估计准确性提高，而估计结果的变异性降低（图 3A，补充图 1）。这一现象符合我们对集成密度估计的设计预期：即使每个重复样本中拟合的模型存在错设，组合多个重复样本也能使模型更准确、可靠地捕捉真实数据密度。为进一步说明集成密度估计的改进能够提升聚类性能，我们转而考察“复合数据”数据集；对于该数据集，所评估的方法均未能完美恢复真实聚类。我们比较 PHM-ensemble 在 $M = 1$、5、10 和 20 时得到的 ARI，同时固定$\rho = 1$ （图 3B）。随着 M 增加，聚类准确率和稳定性均有所提高，表现为 ARI 更高，且重复运行之间的变异性更低。



(A) 双环



(B) 复合数据

图 3：**集成提升了密度估计的准确性和聚类稳定性**(A) 双环数据的全局密度估计过程中，随着集成规模（M）增加，KL散度的箱线图。实线表示平均 KL散度。随着 M 增加，KL散度持续下降，这表明集成密度估计能够更准确地捕捉潜在密度。(B) 随着集成规模（M）增加，PHM-ensemble 在复合数据上的 ARI。实线表示平均 ARI，竖线表示各次运行的 $\pm$1标准差。

最后，为评估可扩展性，我们报告了使用 PHM 及另外两种方法 clustCombi [13], 和 DEMP/DEMP + [14, 29] 在双环数据上构建聚类层次结构所需的实际运行时间（图 4）。这三种方法都能合并集成密度估计中的高斯成分，以构建聚类层次结构，并在通过不同数量的集成重复样本获得的密度估计中，以 $K = 2$ 的分辨率恢复双环结构（补充图 2 和图 3）。在合并过程中，PHM 和 DEMP 均将 $K = 2$ 识别为最显著的结构变化；而 clustCombi 的内置选择程序倾向于选择聚类数量较多的方案（补充说明 2.2）。在这三种方法中，随着集成重复样本数量增加，PHM 的可扩展性最佳；当 $M = 10$时，所需时间不到 100 秒，对应 240 个混合成分。相比之下，即使集成重复样本数量适中，clustCombi 的计算成本也很高；当 $M = 10$时，完成计算需要一个多小时。DEMP 的可扩展性显著优于 clustCombi，但随着混合成分数量增加，其可扩展性仍不及 PHM：在 $M = 10$时，DEMP 耗时 8.3 分钟，而 PHM 耗时 1.3 分钟。

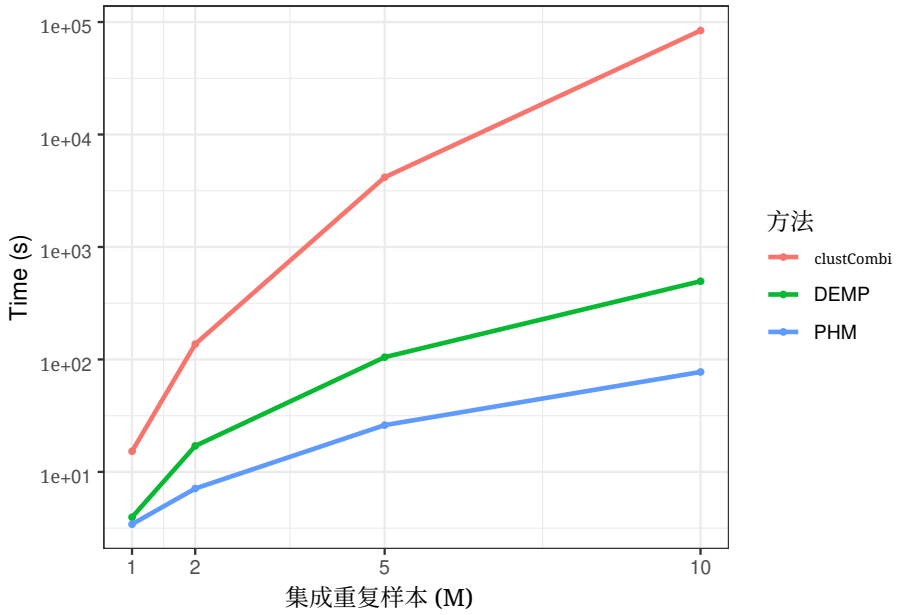


图 4：**使用 PHM、DEMP 和 clustCombi 构建聚类层次结构的运行时间** 图中展示了将 PHM、DEMP 和 clustCombi 程序应用于全局集成密度估计时的实际运行时间。每个绘制点表示在相同设置下重复运行 10 次所得的平均实际运行时间。集成规模$M = 1$、2、5和10 分别对应包含 24、48、120和240 个分量的混合密度。在这三种方法中，随着集成规模的增大，PHM 展现出最佳的可扩展性。

3.2 MNIST 手写数字

在本示例中，我们将 PHM-ensemble 应用于 MNIST 手写数字数据集 [30]。该数据集包含 70,000 张从 0 到 9 的手写数字灰度图像，每张图像由一个 28×28 像素网格表示，广泛用于聚类方法的基准测试。我们将每张图像展平为一个向量，保留前 100 个主成分（PC），并按照在 t-SNE 之前进行 PCA 预处理的常见做法，将所得数据嵌入到二维空间中 [31, 32]。随后，将所得的 t-SNE 嵌入作为聚类的输入（补充说明的第3节）。

我们分别采用全局密度估计和局部密度估计，对 PHM-ensemble 进行评估；每个集成包含 $M = 10$ 个重复样本，每个重复样本的子采样比例为 $\rho = 0.1$。为基准评估其性能，我们还评估 k-均值、层次聚类（Ward连接法）、Leiden聚类、mclust、HDBSCAN 和 SC3s（一种基于 k-均值的集成共识聚类方法）[33]。我们使用聚类准确率（ACC）、归一化互信息（NMI）和 ARI，根据真实数字标签评估聚类结果的质量。对于每种方法，我们均针对 10 个聚类的解计算这些指标；HDBSCAN 除外，对于该方法，我们使用 11 个聚类的解。具体而言，我们为 k-均值、mclust 和 SC3s 设置 $K = 10$；调整 Leiden 的分辨率参数，使其恰好得到 10 个社区；并将层次聚类和 PHM-ensemble 生成的树状图在能够产生 10 个聚类的高度处进行截切。对于 HDBSCAN，没有任何最小簇大小参数的取值能够恰好产生 10 个聚类，因此我们使用能够产生 11 个聚类的最小整数值。除层次聚类和 HDBSCAN 外，所有方法都包含随机成分，包括 k-均值、mclust 和 SC3s 中的随机初始化、PHM-ensemble 中的子采样，以及 Leiden聚类中的随机局部移动启发式算法。为评估这些方法的稳定性，我们使用不同的随机种子对每种随机算法运行 100 次，并考察所得性能指标的变异性。有关聚类流程的更多细节，请参见补充说明的第3节。

相对于真实标签，在所评估的聚类方法中，PHM-ensemble 在全部三项指标上均表现最佳（图 5 和补充表 2）。其准确率接近理想分类器的水平；该分类器使用基于真实标签拟合的高斯混合模型，分别估计各数字类别的密度，并将每幅图像分配给后验概率最高的数字类别。全局密度估计与局部密度估计的结果相近，这表明两者的集成密度估计都能准确捕捉底层数据分布。
PHM-ensemble 在随机化聚类算法中也得到了最稳定的结果（图 5 和补充表 2）。为了评估稳定性的提升是否归因于集成策略，我们还

评估了原始 PHM 程序；该程序使用针对整个数据集拟合的单个密度估计，等同于 $M = 1$和 $\rho = 1$ 条件下的 PHM-ensemble。与原始 PHM 程序相比，PHM-ensemble 不仅提高了准确率（平均 ARI 从 0.891 提高到 0.910），也提升了稳定性（ARI 的标准差从 0.030 降至 0.013）。此外，值得注意的是，PHM-ensemble 在减少算法变异性方面的效果远胜于 SC3s；后者同样采用集成策略，但聚合的是划分而非密度估计。

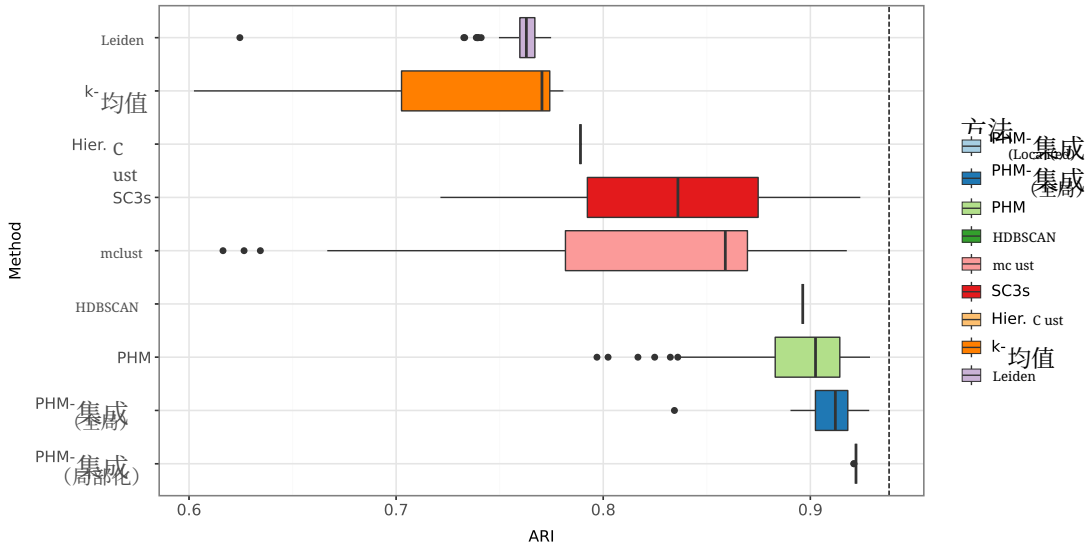


图 5：**PHM 集成在 MNIST 数据上实现了高聚类保真度和稳定性** 应用于 MNIST 数据集的各方法 ARI 值箱线图。理想分类器的性能以竖直虚线表示。PHM 方法优于其他基线方法，其中 PHM-ensemble 在独立运行中表现出非常一致的性能。

最后，PHM-ensemble 能够揭示数据中超越单一平面划分的复杂多分辨率结构，与现有方法相比，可提供更强的可解释性。图6展示了一次 PHM-ensemble 运行得到的 PHM树状图和热图，其中显示了层次结构中的前 20 个聚类。可视化结果表明，对应于各个数字的 10 个聚类通常彼此分离良好，同时也揭示了若干数字聚类内部存在显著异质性，其中的子簇对应不同的书写风格。例如，代表数字 2 的聚类包含两个不同的子簇：一个对应常规字形 2，另一个对应底部带圈的写法 2。其他数字中的子簇，包括 1、4、7 和 9，也同样体现了书写风格上的差异，例如直立与倾斜的写法，如 9 与 *9*（补充图 4）。在较粗分辨率下，层次结构进一步揭示，10 个数字聚类的分离程度并不相同，某些手写数字组表现出更高的结构相似性。在层次结构的最高层，树状图将数字分为两大组：一组主要由直线型数字 1、4、7 和 9 组成，另一组则由曲线型数字 0、2、3、5、6 和 8 组成。热图还表明，9、4 和 7 之间，以及 8、5 和 3 之间具有较强的相似性。此前已有研究指出 9 和 4 具有共同的结构特征，见 [34]。在此，PHM-ensemble 以全面且系统的方式揭示了手写数字之间的这些及其他关系。

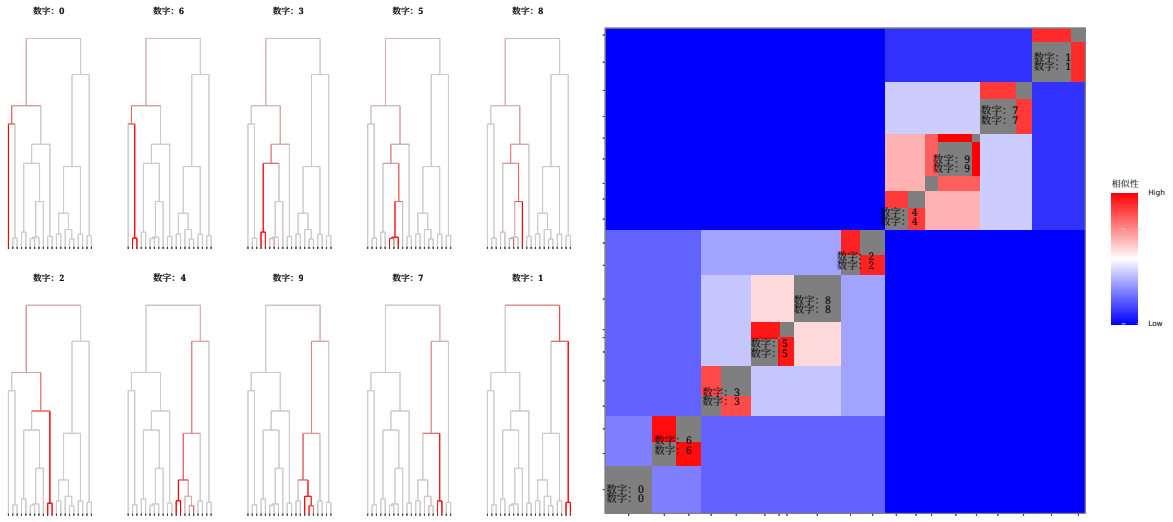


图6：**PHM揭示MNIST中的多分辨率结构**对MNIST数据进行分析后得到的PHM合并流程树状图，展示前20个聚类。树状图的颜色对应于某一分支中的观测值属于某个数字类别的比例。树状图高度和热图颜色均恢复出与独特数字相对应的分量，并展现了数字之间的关系。

3.3 人类免疫健康图谱单细胞RNA测序数据

在此示例中，我们将 PHM-ensemble 应用于人类免疫健康图谱 [6] 的图谱规模单细胞RNA测序（scRNA-seq）数据集，以进行簇分析。该图谱收录了 209 万个外周血单个核细胞，每个细胞测量了约 33,500 个基因。我们直接从原始论文下载经过质量控制和降维的数据。用于簇分析的最终数据集包含 182 万个细胞（经质量控制并去除双细胞和污染细胞后），每个细胞由 30 个主成分（PC）表示。我们使用这些数据，完整复现了 [6]中报告的Leiden聚类分析。

我们采用结合全局密度估计的PHM-ensemble流程，对该数据集进行簇分析。我们使用 $M = 50$ 个集成重复样本，每个重复样本包含100,000个子采样观测（即完整数据集的5.5%）。在对每个重复样本进行密度估计时，允许最多使用100个混合成分。最终的集成混合模型总共包含2,706个分量。

PHM-ensemble的结果（图7）成功识别出对应于主要免疫细胞类型的聚类。作为参考，我们使用艾伦免疫学研究所（AIFI）的细胞类型注释；这些注释经过人工整理，通过反复进行Leiden子聚类并检查经典RNA和表面蛋白（CITE-seq）标志物的表达，构建成三级层级 [6]。PHM-ensemble在 $K = 20$ 的结果相对于定义9种广泛细胞类型的第1级细胞类型注释，取得了0.965的ARI和0.908的归一化互信息（补充表3和补充图5）。相比之下，在所考察的0.005至1.0分辨率范围内（对应9至36个聚类），Leiden聚类在分辨率为0.01时取得最高ARI 0.789（归一化互信息 $= 0.841$），生成10个聚类。更重要的是，PHM结果揭示了免疫细胞类型在基因表达空间中的复杂关系，包括其内部异质性和层级结构。这些关系可以直接通过生成的树状图和热图（图7）进行解读。

在 $K = 20$处，我们观察到，代表主要谱系的聚类（T/NK细胞、B细胞和髓系细胞）彼此之间具有明显区分：任何淋巴系聚类与髓系聚类之间，以及T/NK细胞聚类与B细胞聚类之间的重叠都可以忽略不计。

几乎所有T细胞（超过99.99%）都归入单个聚类（簇1），该聚类与NK细胞簇形成一个分支。具体而言，T细胞聚类与CD56^{dim} NK聚类（簇6）的重叠最多，这与NK细胞和效应记忆性CD8 T细胞共享细胞毒性程序相一致；而CD56^{bright} （簇12）和增殖性NK（簇16）聚类则在之后加入该分支。

三个聚类（簇2、7和9）包含了所有已注释B细胞的98.5%（不含其他细胞类型），形成一个独特的分支，与所有其他细胞类型明显分离。两个最大的聚类分别对应初始B细胞（簇2，包括过渡性B细胞）和记忆B细胞（簇7，包括效应B细胞），且呈现出

该分支内唯一明显的重叠。这种模式与初始B细胞在抗原暴露后分化为记忆B细胞的过程一致 [35]。浆细胞形成一个独立的聚类（簇9），最后并入该分支，与初始B细胞和记忆B细胞聚类的重叠可以忽略不计。第四个较小的聚类（簇14，占B细胞的1.4%）由已注释的初始B细胞和记忆B细胞混合组成。该聚类与其他B细胞聚类相隔离，并在主B细胞分支之外并入层次结构，与一个较小的血小板聚类（簇18）并列。
单核细胞和2型常规树突状细胞 (cDC2) 占据三个相邻的簇（簇3、5和8），共同构成一个与淋巴系簇明显分离的髓系区块。这三个簇合计包含所有已注释单核细胞的99.97%。经典型CD14、非经典型CD16和中间型单核细胞构成最大的簇（簇3），在这一分辨率下无法彼此区分。该簇与cDC2簇（簇8）重叠，表明单核细胞与cDC2的表达谱呈连续变化。这种重叠与单核细胞样$CD14^+$ cDC2（DC3）的中间表达谱相符，也印证了已知的难以将DC3与$CD14^+$单核细胞区分开来的问题 [36, 37]。一个由$IL1B^+$ CD14单核细胞构成的小簇（簇5）可能反映了炎症状态，是最后并入髓系区块的簇。在该区块之外，浆细胞样树突状细胞 (pDC) 构成一个独立的簇（簇4），而1型常规树突状细胞 (cDC1) 则与祖细胞聚为一簇（簇13）。

其他注释为1级的细胞类型在聚类数据集中较为罕见。具体而言，血小板、祖细胞、红细胞和先天性淋巴样细胞（ILC）分别占所有细胞的0.43%、0.08%、0.08%和0.05%。血小板和红细胞各自形成一个主聚类（分别为簇11和簇10），且其中相应细胞类型占绝大多数（$\geq 96\%$）。祖细胞占一个聚类（簇13）的多数（53%），该聚类还包含树突状细胞（cDC1和ASDC）。相比之下，在这一分辨率下，ILC不会形成独立聚类，而是分布在T细胞和NK细胞簇中（主要为簇1和簇12）。
此外，PHM-ensemble的结果让我们能够通过直接放大层次结构中的各个分支，在较细分辨率下更详细地考察细胞类型之间的关系，而无需反复进行子聚类。
图8A展示了对 $K = 41$处跨越T细胞簇（簇1）的子树进行的更细致考察。在T细胞簇内部，最显著的区分是将一小群增殖性细胞（包括一些增殖性NK细胞）与其余T细胞区分开来。这种分离似乎反映的是共同的细胞周期程序，而非T细胞谱系之间的差异。其余按谱系定义的T细胞子簇清晰地分为三大组：常规初始T细胞和CD4 T细胞（初始CD4、记忆CD4、初始CD8和调节性T细胞，即Treg）、先天样T细胞（MAIT和 $\gamma\delta$T，即gdT），以及记忆CD8 T细胞，其中包括细胞毒性效应细胞。
图8B展示了对 $K = 27$处单核细胞区块（簇3、5和8）的细化视图。在该区块内，CD14单核细胞形成两个主要簇：一个以核心CD14单核细胞为主，另一个几乎完全由ISG+CD14单核细胞组成，代表具有最强干扰素反应特征的细胞。有趣的是，PHM-集成层次结构显示，核心CD14单核细胞簇先与CD16单核细胞簇合并，随后才与ISG+ CD14单核细胞簇合并，这表明在基因表达空间中，核心CD14单核细胞与CD16单核细胞的重叠程度高于其与ISG+ CD14单核细胞的重叠程度。这一观察结果与CD16单核细胞由CD14单核细胞分化而来的情况一致 [38, 39] ，并提示干扰素刺激状态在CD14单核细胞之间引入的转录差异，大于CD14到CD16转变所伴随的差异。

利用这一图谱规模数据集，我们进一步将PHM-ensemble与 [6]中使用的Leiden聚类方法进行计算效率基准比较。在配备3.1 GHz英特尔至强处理器的计算节点上，集成密度估计每个重复样本需要1至2小时的实际运行时间。通过在由SLURM管理的计算集群上并行运行全部50个重复样本，密度估计可在4小时内完成。层次结构构建在单个计算节点上约需21分钟的实际运行时间，其中包括计算 ΔP_{mc} 矩阵所需的15.5分钟，以及层次合并所需的5.7分钟。在整个PHM-ensemble流程中，所有处理单元的内存使用量始终低于2 GB。相比之下，我们对原始Leiden聚类的复现在同一系统上约需8.5小时，峰值内存使用量超过10 GB。

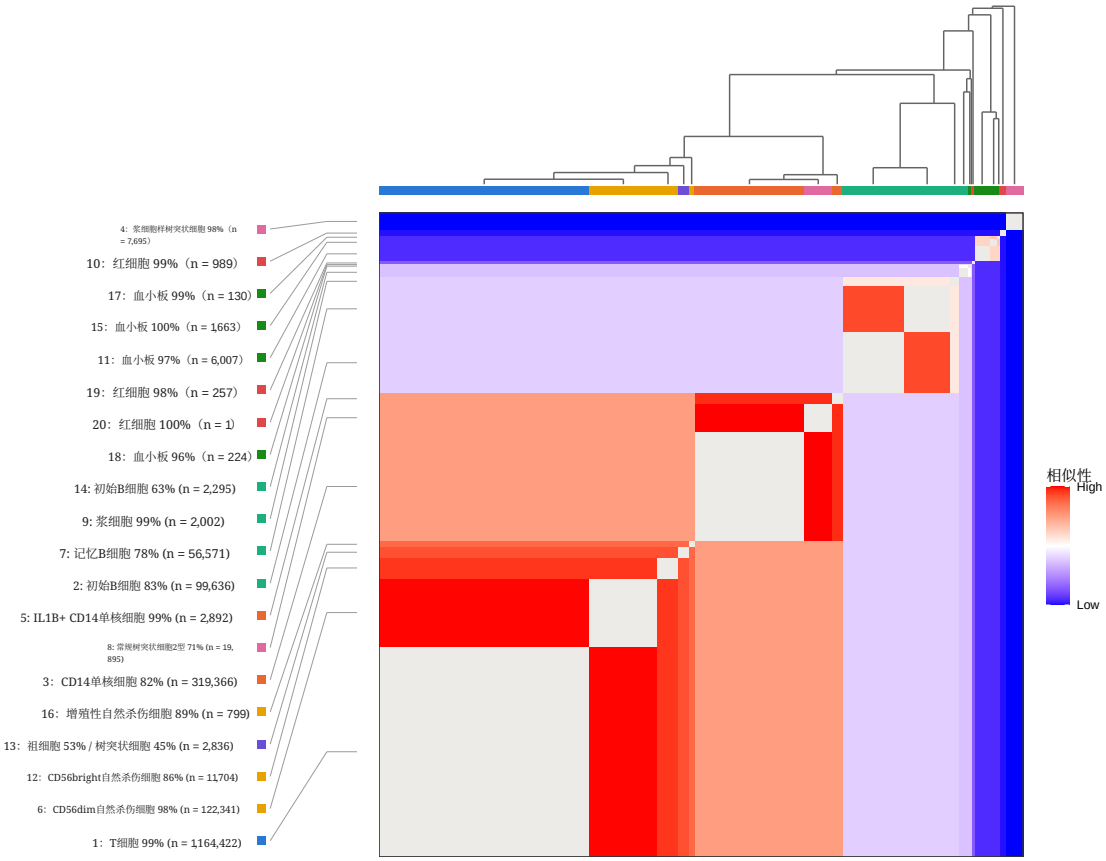
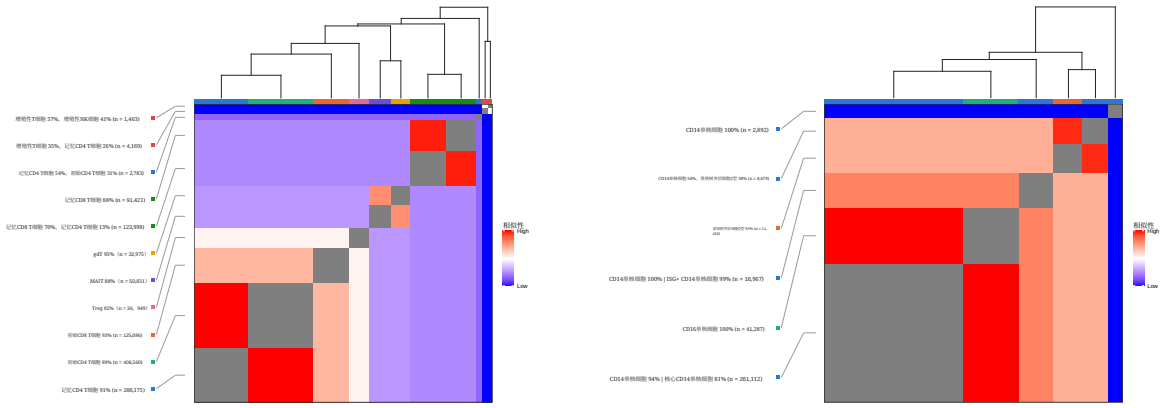


图7：**PHM 在表达空间中还原了宽泛的细胞类型组及其关系** 人类免疫健康图谱的 PHM 集成分析在分辨率 $K = 20$下生成的 PHM 树状图与热图叠加图。使用最大后验概率（MAP）规则将单个细胞分配到聚类中。每个聚类均标注其索引、最常见的 AIFI一级注释及携带该注释的细胞所占百分比，以及分配到该聚类的细胞总数（n）。聚类索引仅用作标识符。每个对角块的面积与分配到相应聚类的细胞数量成正比。树状图和热图揭示了主要细胞类型组及其关系。



(A) T细胞　　(B) 单核细胞

图 8：**PHM-ensemble 在较细分辨率下还原了 T细胞和单核细胞中的亚簇结构**PHM 集成热图和树状图：(A) T 细胞区块（$K = 20$ 处的簇1），显示于 $K = 41$；(B) 髓系区块（$K = 20$ 处的簇3、5 和 8），显示于 $K = 27$。每个聚类均标注其最常见的 AIFI 二级注释、具有该注释的细胞占聚类细胞总数的百分比，以及分配至该聚类的细胞总数（n）。当第二种注释占该聚类细胞至少 10% 时，也会列出该注释。在 (B) 中，两个主要的 CD14 单核细胞簇还在竖线后标注了其最常见的 AIFI 三级注释。每个对角区块的面积与分配至该聚类的细胞数成正比。PHM 集成分析揭示的 T 细胞区块和单核细胞区块内的亚组结构，与已知生物学特征一致。

4 讨论

本文提出了一种新颖的计算框架，用于对大规模数据集进行聚类，并解决保真度、可解释性、稳定性和可扩展性方面的挑战。该框架采用概率表示学习的视角，首先从观测样本中学习底层的数据生成分布，然后应用 PHM 算法，根据学习到的分布构建聚类层次结构。为了使该框架适用于大规模数据，我们开发了集成密度估计和层次结构构建的计算策略，尤其注重准确率和稳定性。通过合成数据和真实数据示例，我们展示了该方法能够有效地恢复大规模数据集中的有意义聚类结构，同时为数据提供灵活且可解释的多分辨率表示。

嵌入和聚类是针对高维复杂数据进行无监督学习的两个重要组成部分。尽管文献中已投入大量精力开发有效的嵌入方法，但对如何在生成的嵌入空间中进行聚类，关注相对较少。在本研究中，尤其通过 MNIST 示例，我们表明，聚类方法的选择及其性能会显著影响无监督学习流程的最终输出质量。更广泛地说，这一观察适用于许多实际场景（如空间转录组学）：如果没有准确且稳定的聚类，即便经过精心学习的嵌入，其优势也未必能得到充分发挥。

尽管 PHM-ensemble 已经可以应用于由深度学习模型生成的嵌入，但所提出的框架还可以进一步与现代深度学习架构相结合。这种整合的一个潜在益处是联合优化嵌入和密度估计目标，最终可能提升聚类性能。拓展现有的融合高斯混合模型 [40, 41] 的深度学习方法，是未来研究中一个充满前景且自然的方向。此外，当前框架中的集成实现也可以进行调整，以更好地适配深度学习架构，例如利用随机失活等机制来生成集成多样性。

从混合密度估计中有效构建层次结构，是在我们提出的框架中识别和解释聚类结构的关键步骤。PHM 算法为此问题提供了一种计算高效的解决方案。构建聚类层次结构的一项关键挑战，是缺乏一种有原则且计算上可行的簇间分离度度量。标准统计距离和散度，包括 KL散度、詹森-香农散度、巴塔查里亚距离及其系数、海林格距离、瓦瑟斯坦距离和全变差距离，通常无法为高斯混合 [14]导出闭式表达式。尽管我们在先前工作 [8], 中提出的 P_{mc}对于高维数据仍需要蒙特卡洛近似，但它具备一种尤为

有用的成对分解性质，能够在凝聚合并过程中实现高效计算和更新。在本研究中，我们进一步证明，P_{mc} 可用于构建合适的连接函数，并能自然地适用于集成密度估计。更一般地说，P_{mc} 的分解性质适用于广泛的一类分布，并不限于高斯混合。这种普适性为未来将所提出的框架扩展至其他密度估计方法奠定了基础。

我们提出的方法也与文献中的共识聚类方法相关 [42, 43], ，这类方法会为同一数据集生成一组划分，并构建共关联矩阵，以得出最终的共识划分。虽然这两种方法都旨在提高聚类稳定性，但它们构建的集成有所不同。共识聚类方法会汇总离散的划分标签，因此需要解决集成中各成员之间的簇对应问题，并生成一个 $N \times N$ 共关联矩阵。相比之下，我们提出的框架直接汇集集成成员的密度估计，从而避免这一对应问题，并生成一种组合表示；其规模取决于拟合分量的数量，而非数据点的数量。更重要的是，我们的方法符合多分辨率聚类这一更广泛的目标：不同于传统共识聚类止步于单个平面划分，我们的方法提供统一的密度表示，支持后续构建完整的聚类层次结构。

最后，我们并不认为所提出的聚类框架可以取代现有的基于划分的方法或软聚类方法。这些方法是该框架不可或缺的组成部分，在密度估计过程中提供关键的归纳偏置，有助于捕捉高维、多模态空间中的结构。我们希望，所提出的框架能够推动开发多样且新颖的聚类算法，以满足对大规模、高维、复杂科学数据进行有效簇分析的迫切需求。

软件与数据可用性

PHM-ensemble流程的软件实现可在 https://github.com/aturfah/PHM 获取。用于生成本文结果的代码和数据可在 https://github.com/aturfah/PHM_ensemble_repo 获取。

致谢

本研究由美国国立卫生研究院（NIH）资助，资助项目编号为 R35GM138121 和 R01ES033634。

参考文献

[1] Olivier Bousquet 和 André Elisseeff。稳定性与泛化。《机器学习研究杂志》，2:499–526，2002年。[2] André Elisseeff、Theodoros Evgeniou 和 Massimiliano Pontil。随机学习算法的稳定性。《机器学习研究杂志》，6:55–79，2005年。[3] Asa Ben-Hur、André Elisseeff 和 Isabelle Guyon。一种用于发现聚类数据结构的基于稳定性的方法。载于太平洋生物计算研讨会，第6–17页，2002年。[4] Tilman Lange、Volker Roth、Mikio L Braun 和 Joachim M Buhmann。基于稳定性的聚类解验证。神经计算，16(6):1299–1323，2004年。[5] Ulrike von Luxburg。聚类稳定性：综述。机器学习基础与趋势，2(3):235–274，2010年。arXiv:1007.1075。[6] Qiuyu Gong、Mehul Sharma、Marla C Glass、E mma L Kuan、Aishwarya Chander、Mansi Singh、Lucas T Graybuck、Zachary J Thomson、Christian M LaFrance、Samir Rachid Zaim 等。多组学分析揭示健康成年人中年龄相关的免疫动态。《自然》，648(8094):696–706，2025年。[7] Kevork N Abazajian、Jennifer K Adelman-McCarthy、Marcel A Agüeros、Sahar S Allam、Carlos Allende Prieto、Deokkeun An、Kurt SJ Anderson、Scott F Anderson、James Annis、Neta A Bahcall 等。斯隆数字巡天第七次数据发布。《天体物理学杂志增刊》，182(2):543–558，2009年。[8] Xiaoquan Wen、Ali Turfah 和 Yuntian Wu。概率多分辨率聚类：基因组数据分析示例。*SSRN*电子期刊，2026年。SSRN工作论文7087586，发布于2026年7月9日。[9] Antonio Torralba、Phillip Isola 和 William T. Freeman。《计算机视觉基础》。MIT出版社，2024年。[10] Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。《深度学习》。MIT出版社，2016年。

[11] Christopher R. Genovese 和 Larry Wasserman。高斯混合筛的收敛速率。《统计年鉴》，28(4):1105–1127，2000年。[12] Chris Fraley 和 Adrian E. Raftery。基于模型的聚类、判别分析与密度估计。《美国统计学会会刊》，97(458):611–631，2002年。[13] Jean-Patrick Baudry、Adrian E. Raftery、Gilles Celeux、Kenneth Lo 和 Raphael Gottardo。合并混合成分以进行聚类。《计算与图形统计学杂志》，19(2):332–353，2010年。[14] Christian Hennig。合并高斯混合成分的方法。《数据分析与分类进展》，4(1):3–34，2010年。[15] A. P. Dempster、N. M. Laird 和 D. B. Rubin。通过EM算法从不完整数据中进行最大似然估计。《皇家统计学会杂志：B辑》，39(1):1–38，1977年。[16] Luca Scrucca、Michael Fop、Thomas Brendan Murphy 和 Adrian E. Raftery。mclust 5：使用高斯有限混合模型进行聚类、分类与密度估计。《R期刊》，8(1):289–317，2016年。[17] Chi Jin、Yuchen Zhang、Sivaraman Balakrishnan、Martin J. Wainwright 和 Michael I. Jordan。高斯混合模型似然中的局部极大值：结构性结果与算法后果。载于神经信息处理系统进展（NeurIPS），2016年。arXiv:1609.00978。[18] Sivaraman Balakrishnan、Martin J. Wainwright 和 Bin Yu。EM算法的统计保证：从总体分析到基于样本的分析。《统计年鉴》，45(1):77–120，2017年。[19] Peter Bühlmann 和 Bin Yu。分析装袋法。《统计年鉴》，30(4):927–961，2002年。[20] Fabian Pedregosa、Gaël Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vincent Dubourg、Jake Vanderplas、Alexandre Passos、David Cournapeau、Matthieu Brucher、Matthieu Perrot 和 Édouard Duchesnay。Scikit-learn：使用Python进行机器学习。《机器学习研究杂志》，12:2825–2830，2011年。arXiv:1201.0490。[21] Adrian O'Hagan、Thomas Brendan Murphy、Luca Scrucca 和 Isobel Claire Gormley。使用高斯混合模型，通过刀切法、自助法和加权似然自助法研究聚类中的参数不确定性。《计算统计学》，34(4):1779–1813，2019年。arXiv:1510.00551。[22] Christopher M. Bishop。《模式识别与机器学习》。Springer，2006年。[23] Gilles Celeux 和 Gérard Govaert。用于聚类的分类EM算法及其两个随机版本。《计算统计学与数据分析》，14(3):315–332，1992年。[24] Joe H. Ward。通过层次分组优化目标函数。《美国统计学会会刊》，58(301):236–244，1963年。[25] Jeffrey D. Banfield 和 Adrian E. Raftery。基于模型的高斯与非高斯聚类。《生物统计学》，49(3):803–821，1993年。[26] Cameron B. Browne、Edward Powley、Daniel Whitehouse、Simon M. Lucas、Peter I. Cowling、Philipp Rohlfshagen、Stephen Tavener、Diego Perez、Spyridon Samothrakis 和 Simon Colton。蒙特卡罗树搜索方法综述。《IEEE计算智能与游戏人工智能汇刊》，4(1):1–43，2012年。[27] Richard S. Sutton 和 Andrew G. Barto。《强化学习：导论》。MIT出版社，第2版，2018年。[28] Cleve Moler。分布式内存多处理器上的矩阵计算。载于 Michael T. Heath（编），《超立方体多处理器1986》，第181–195页。工业与应用数学学会，费城，PA，1986年。[29] Volodymyr Melnykov。通过成对重叠合并混合成分以进行聚类。《计算与图形统计学杂志》，25(1):66–90，2016年。[30] Yann LeCun、Léon Bottou、Yoshua Bengio 和 Patrick Haffner。将基于梯度的学习应用于文档识别。《IEEE会刊》，86(11):2278–2324，1998年。[31] Laurens van der Maaten 和 Geoffrey Hinton。使用t-SNE进行数据可视化。《机器学习研究杂志》，9:2579–2605，2008年。[32] Laurens van der Maaten。使用基于树的算法加速t-SNE。《机器学习研究杂志》，15:3221–3245，2014年。[33] Vladimir Yu Kiselev、Kristina Kirschner、Michael T Schaub、Tallulah Andrews、Andrew Yiu、Tamir Chandra、Kedar N Natarajan、Wolf Reik、Mauricio Barahona、Anthony R Green 等。Sc3：单细胞RNA测序数据的共识聚类。《自然·方法》，14(5):483–486，2017年。[34] Sarah M Goggin 和 Eli R Zunder。Eschr：一种超参数随机化集成方法，可在多样化数据集上实现稳健聚类。《基因组生物学》，25(1):242，2024年。

[35] Jason G. Cyster 和 Christopher D. C. Allen。B细胞应答：细胞相互作用动力学与决策。*Cell*，177(3):524–540，2019年。[36] Alexandra-Chloé Villani、Rahul Satija、Gary Reynolds、Siranush Sarkizova、Karthik Shekhar、James Fletcher、Morgane Griesbeck、Andrew Butler、Shiwei Zheng、Suzan Lazo、Laura Jardine、David Dixon、Emily Stephenson、Emil Nilsson、Ida Grundberg、David McDonald、Andrew Filby、Weibo Li、Philip L. De Jager、Orit Rozenblatt- Rosen、Andrew A. Lane、Muzlifah Haniffa、Aviv Regev 和 Nir Hacohen。单细胞RNA测序揭示人血树突状细胞、单核细胞和祖细胞的新类型。*Science*，356(6335):eaah4573，2017年。[37] Pierre Bourdely、Giorgio Anselmi、Kristine Vaivode 等。CD1c$^+$人树突状细胞的转录与功能分析鉴定出一个可引发CD163$^+$ CD8$^+$T细胞应答的CD103$^+$ 亚群。*Immunity*，2020年。PMID: 32610077。[38] Tamar Tak、Julia Drylewicz、Lennart Conemans、Rob J. de Boer、Leo Koenderman、José A. M. Borghans 和 Kiki Tesselaar。人体单核细胞亚群在体内的循环与成熟动力学。*Blood*，130(12):1474–1477，2017年。[39] Amit A. Patel、Yan Zhang、James N. Fullerton、Lies Boelen、Anthony Rongvaux、Alexander A. Maini、Venetia Bigley、Richard A. Flavell、Derek W. Gilroy、Becca Asquith、Derek Macallan 和 Simon Yona。稳态及全身性炎症状态下人单核细胞亚群的命运与寿命。*Journalof ExperimentalMedicine*，214(7):1913–1923，2017年。[40] Zhuxi Jiang、Yin Zheng、Huachun Tan、Bangsheng Tang 和 Hanning Zhou。变分深度嵌入：一种用于聚类的无监督生成式方法。载于第二十六届国际人工智能联合会议论文集（*IJCAI*），第1965–1972页，2017年。arXiv:1611.05148。[41] Nat Dilokthanakul、Pedro A. M. Mediano、Marta Garnelo、Matthew C. H. Lee、Hugh Salimbeni、Kai Arulkumaran 和 Murray Shanahan。使用高斯混合变分自编码器进行深度无监督聚类，2016年。arXiv:1611.02648。[42] Stefano Monti、Pablo Tamayo、Jill Mesirov 和 Todd Golub。共识聚类：一种用于基因表达微阵列数据类别发现与可视化的基于重采样的方法。*Machine Learning*，52(1–2):91–118，2003年。[43] Ana L. N. Fred 和 Anil K. Jain。利用证据累积组合多个聚类结果。*IEEE Transactionson Pattern Analysis and Machine Intelligence*，27(6):835–850，2005年。