arXiv:2602.09936v1 [stat.ML] 10 Feb 2026

高维下的 k-均值算法的灾难性失效，以及 Hartigan 算法如何避免这一问题

Roy R. Lederman [1] **David Silva-S´anchez** [2] **Ziling Chen** [1] **Gilles Mordant** [2] **Amnon Balanov** [3] **Tamir Bendory** [3]

摘要

Lloyd 算法是一种应用最广泛的聚类方法之一。我们证明，在高维、高噪声的情形下，该算法会出现灾难性失效：以高概率，数据的几乎每一种划分都是一个不动点。因此，Lloyd 算法只会返回其初始划分——即使其他方法能够轻而易举地恢复潜在的簇。相比之下，我们证明，Hartigan 算法不会出现这种病态现象。我们的结果揭示了这两种算法之间的鲜明差异，并为高维下使用 k-均值算法时经常观察到的经验性困难提供了理论解释。

[1]耶鲁大学统计学与数据科学系，美国康涅狄格州纽黑文 [2]耶鲁大学应用与计算数学系，美国康涅狄格州纽黑文 [3]特拉维夫大学电气与计算机工程学院，以色列特拉维夫。通讯作者：Roy R. Lederman <>。

预印本。*2026*年*2*月*11*日。

1. 引言

聚类是统计学和机器学习中的核心问题。该问题最常见的形式之一是 k-均值（MacQueen，1967），其目标是最小化簇内方差；历史回顾可参见Bock (2008)。解决 k-均值问题通常采用 Lloyd 的迭代 k-均值算法进行近似求解（Lloyd，1982；Forgy，1965），该算法与此问题几乎成了同义词，以至于它被称为 k-均值算法。时至今日，它仍被视为数据分析领域最重要的算法之一（Wu 等人，2008）。

然而，随着高维数据分析应用的发展，例如基因表达、文本分析和成像，人们发现 Lloyd 的 k-均值算法在高维设置下会遇到困难（例如，Hartigan，1975；Steinley，2006；Zha 等人，2001；Ding 和 He，2004）。本文证明，这些困难反映了一个关键问题，它会导致该算法灾难性失效，即使是在非常简单的问题中也如此。此外，为了完整呈现问题全貌，我们还证明了 Hartigan 的 k-均值算法（Hartigan，1975），即 Lloyd 算法的一种贪心变体，能够

在 Lloyd 算法失效之处取得成功。

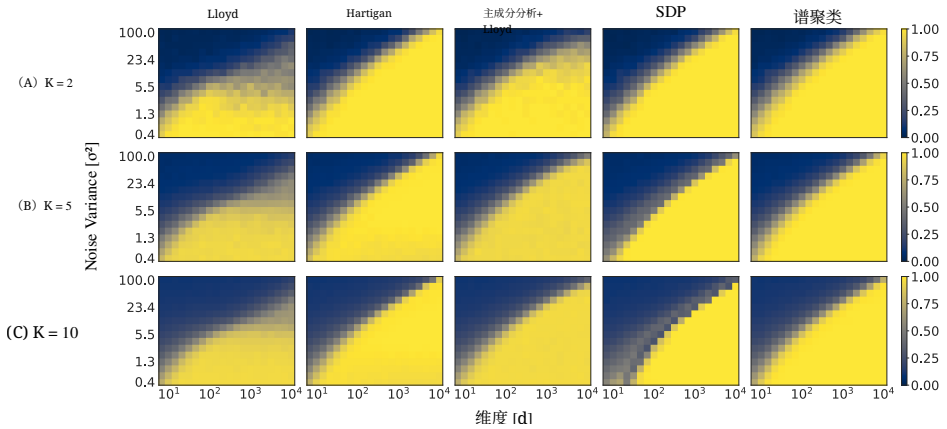


图*1.* 归一化互信息（NMI；见定义 A.13）介于真实划分与各聚类算法的输出之间。每个条目所报告的均值来自 100 次独立试验：在每次试验中，我们从高斯混合模型（高斯混合模型），即模型 2.1 （推广至 $K \geq 2$），其中 $\tau^2 = 1.0$，且 20 个每类样本，并将各算法运行至收敛。结果表明，在高噪声、高维情形下，Lloyd 的 k-均值相对于其他方法表现较差。相比之下，Hartigan 算法取得了与谱聚类以及基于半定规划（半定规划）的聚类相当的性能。详见第 4.1节。

1.1. 主要结果

以下是结果的简要版本，省略了第2.1节、推论3.8以及推论3.12中的完整表述所包含的一些细节。

定理 1.1（非正式表述：高噪声、高维情形下 Lloyd 与 Hartigan 的有限样本行为）。考虑 $n \in \mathbb{N}$ 来自一个 $\mathbb{R}^d$ 中的双簇（$K = 2$）高斯混合模型的观测样本 $x_1, \ldots, x_n$，其均值 $\mu_1^\star, \mu_2^\star \in \mathbb{R}^d$ 服从标准正态分布，噪声协方差 $\sigma^2 I_d$为各向同性。令 $\mathcal{F}_{\text{Lloyd}}$ 表示这样一个事件：所有（极端不平衡的划分除外）划分 都是 *Lloyd* 算法的不动点；令 $\mathcal{F}_{\text{Hart}}$ 表示这样一个事件：存在一个错误划分 是 *Hartigan* 算法的不动点。那么，对于 $\sigma^2 > n$,

$$1 - \mathbb{P}(\mathcal{F}_{\text{Lloyd}}) \lesssim 2^n\, n \left(1 - \frac{1}{n^2}\right)^{d/4}, \tag{1}$$

$$\mathbb{P}(\mathcal{F}_{\text{Hart}}) \lesssim 2^n \left(1 - \frac{1}{4\sigma^4}\right)^{d/4}, \tag{2}$$

这导致了当 $d, n \to \infty$时截然不同的行为：

$$\mathbb{P}(\mathcal{F}_{\text{Lloyd}}) \to 1 \quad \textit{if} \quad d \gtrsim n^3, \tag{3}$$

$$\mathbb{P}(\mathcal{F}_{\text{Hart}}) \to 0 \quad \textit{if} \quad d \gtrsim n\sigma^4, \tag{4}$$

其中，$a \lesssim b$表示存在某个 $C > 0$ 使得$a \le Cb$。

也就是说，在定理1.1和推论3.8所述的高噪声、高维情形下，几乎每个划分都是 *Lloyd* 更新映射的不动点。由于 Lloyd 算法会在不动点处终止，这直接带来了算法层面的影响：以高概率，对于几乎任何初始化（极端不平衡的初始化除外），Lloyd 算法都会在第一次更新后停止并返回相同的划分，也就是说，除了初始化之外，它几乎毫无进展。

人们很容易将这一现象归因于高维中固有的几何退化（例如，“所有距离几乎相等”），从而认为在该情形下有意义的聚类是不可能的。然而，我们对 Hartigan 算法的结果以及相应实验表明，这一结论并不正确：在适当的情形下，即使 Lloyd 算法卡在其初始化处，贪心局部改进动力学仍能避开虚假不动点，并以高概率恢复正确的聚类。

特别地，由于 Hartigan 算法单调降低 k-均值目标函数，并且保证在某个不动点处终止，定理1.1 和推论 3.12 表明，在高噪声、高维情形下，以很高的概率，不存在错误的不动点。因此，从几乎任意初始化出发，Hartigan 算法都会在正确划分处终止。

1.2. 主要实证结果

为将这些现象置于更广泛的背景下，我们将 Lloyd 和 Hartigan 算法与几种替代性的聚类方法进行比较：(i) PCA+ k-均值方法，该方法首先应用主成分分析，然后在降维后的空间中运行 Lloyd 算法（Zha 等人，2001年；Ding 和 He，2004年）；(ii) 来自现代k-meansSDP 系列的半定规划（SDP）松弛方法（Peng 和 Wei，2007年）；以及 (iii) 谱聚类算法（Shi 和 Malik，2000年；Ng 等人，2001年）。需要指出的是，后者并非针对相同的目标函数设计，而后两者在计算可扩展性方面较差，且 n。

图1 给出了比较这些算法的数值实验摘要；详细描述见第4.1节。在所有 K中，Lloyd 算法都表现出明显的失败区域（较低的归一化互信息（NMI）），并且该区域一直延伸到其他算法已经成功的情形。有趣的是，在我们的定理适用范围之外，Hartigan 算法的表现似乎与最先进的替代算法相当。

1.3. 既有研究

多年来，k-均值聚类以及 Lloyd k-均值算法的行为一直受到广泛研究。大量研究给出了统计和计算条件，在这些条件下，Lloyd 算法或密切相关的过程能够恢复正确的聚类或簇均值（例如，Lu & Zhou，2016年；Gao & Zhang，2022年；Ndaoud，2022年；Chen& Yang，2021年）。此外，先前对 Lloyd 算法与 Hartigan 算法的比较（Telgarsky& Vattani，2010年；Slonim等人，2013 年）表明，Hartigan 算法的不动点是 Lloyd 算法不动点的子集。

我们的结果通过确定一个广泛的高噪声、高维有限样本区域，补充了上述研究。在该区域内，Lloyd 算法表现出不动点丰盛现象，因此以高概率无法改进其初始化；而 Hartigan 算法则通过以高概率不存在错误不动点，避免了这一病态现象。

2. 预备知识

本节汇集了全文中使用的定义和符号。我们用 $[n]$ 表示指标集 $\{1, 2, \ldots, n\}$。我们用 $\|\cdot\|$ 表示向量欧几里得范数。

2.1. 观测模型

我们分析双成分高斯混合模型（$K = 2$）中的 k-均值问题。本分析中所考虑的随机变量将在以下模型中形式化。

模型 2.1 (双成分各向同性高斯混合模型)。 观测值

tions $X = \{x_i\}_{i=1}^n$ 根据以下方式抽取 $x_i := \mu^\star_{z^\star_i} + \xi_i$，其中底层随机变量如下：

(a) 真实类别中心是随机且独立同分布的：$\mu^\star_1$，$\mu^\star_2$ 独立同分布 $\sim \mathcal{N}(0, \tau^2 I_d)$，且 $\tau \in \mathbb{R}^+$。
(b) 样本噪声是随机的、独立同分布的，且与真实中心相互独立：ξ_i 独立同分布 $\sim \mathcal{N}(0, \sigma^2 I_d)$。
(c) 真实类别分配由潜在标签 $z^\star_1$，……，$z^\star_n \in \{1,2\}$ 给出。我们将真实类别记为 $S^\star_\ell := \{i \in [n] : z^\star_i = \ell\}$。此处，我们不对类别分配的具体分布作任何假设，只要求两个类别均非空。

我们假设(a)、(b) 和(c) 相互独立。

2.2. k-均值问题

k-均值问题有几种等价的表述方式；我们采用标准的基于划分的表述。给定样本 $x_1, \ldots, x_n \in \mathbb{R}^d$ 和整数 $K \geq 2$，k- 均值问题 要求将指标集 $[n]$划分为K 个非空簇 $S = \{S_1, \ldots, S_K\}$，使簇内距离平方和最小，

$$\arg\min_{S} \sum_{k=1}^{K} \sum_{i \in S_k} \|x_i - \mu_k\|^2, \tag{5}$$

其中 $\mu_k = |S_k|^{-1} \sum_{i \in S_k} x_i$。这种损失通常称为惯性、簇内平方和（WCSS 或 WSS）、误差平方和（SSE）或失真。-均值问题已知是 NP-困难的（Aloise 等人，2009）。

以下定义有助于介绍和分析 Lloyd 和 Hartigan 针对 $K = 2$的算法。

定义 2.2（当前分配、簇与划分）。在第 t 次迭代时，当前聚类分配 是一个标记向量 $z^{(t)} = (z_1^{(t)}, \ldots, z_n^{(t)}) \in \{1,2\}^n$。它导出当前簇 $C_j^{(t)} := \{ i \in [n] : z_i^{(t)} = j \}$ 以及相应的（当前）划分 $\mathcal{P}^{(t)} := \{C_1^{(t)}, C_2^{(t)}\}$。除非另有说明，我们只考虑两个簇均非空的迭代状态，即 $|C_1^{(t)}|, |C_2^{(t)}| > 0$。在不需要时，我们省略迭代下标 t。

定义 2.3 (质心)。 考虑一个二划分$\mathcal{P}^{(t)} = \{C_1^{(t)}, C_2^{(t)}\}$；与之对应的经验质心为

$$\widehat{\mu}_j^{(t)} := \frac{1}{|C_j^{(t)}|} \sum_{i \in C_j^{(t)}} x_i, \qquad j \in \{1,2\}. \tag{6}$$

2.2.1. LLOYD 算法

用于 k-均值（Lloyd，1982）的 Lloyd 算法是一种交替最小化启发式，用于近似最小化方程（5）中的目标函数。算法从一个初始非空划分开始，并反复执行以下两个步骤。

分配步骤。 给定当前质心 $\widehat{\mu}_1^{(t)}, \widehat{\mu}_2^{(t)}$，将每个样本重新分配给最近的质心：

$$z_i^{(t+1)} \in \arg\min_{j \in \{1,2\}} \|x_i - \widehat{\mu}_j^{(t)}\|^2, \qquad i \in [n]. \tag{7}$$

求平均步骤。 给定更新后的划分，将质心重新计算为经验均值：

$$\widehat{\mu}_j^{(t+1)} = \frac{1}{|C_j^{(t+1)}|} \sum_{i \in C_j^{(t+1)}} x_i, \qquad j \in \{1,2\}. \tag{8}$$

Lloyd 算法会使损失单调递减（方程 (5)），并保证收敛（例如参见，Slonim等人，2013 年，第 1678 页）。针对本文所讨论的情形，更详细的伪代码见附录 B。

2.2.2. HARTIGAN 算法

Hartigan 算法（Hartigan，1975）是一种用于最小化 k-均值损失（方程（5））的贪心算法。与 Lloyd 的批量重新分配不同，Hartigan 每次只更新一个样本的划分。对于每个样本，Hartigan 算法依据*Hartigan* 加权距离，将该样本重新分配给最近的质心：

$$\Delta^2_{\mathrm{H}}(x_i, C_j^{(t)}) := \begin{cases} \frac{|C_j^{(t)}|}{|C_j^{(t)}|-1} \|x_i - \widehat{\mu}_j^{(t)}\|^2, & \text{if } i \in C_j^{(t)}, \\ \frac{|C_j^{(t)}|}{|C_j^{(t)}|+1} \|x_i - \widehat{\mu}_j^{(t)}\|^2, & \text{if } i \notin C_j^{(t)}. \end{cases} \tag{9}$$

算法反复遍历所有样本，直到不再接受任何重新分配；此时，输出划分对于单样本移动而言是局部最优的（即 1-交换局部最优）。

Hartigan 算法会使损失（方程（5））单调递减，并且保证收敛（Slonim等人，2013 年，第 1678 页）。针对本文所讨论情形的更详细伪代码见附录 B。

2.2.3. 补充符号

为了陈述下一节的主要结果，介绍以下定义会很方便。

定义 2.4（类别比例、纯度和正确性）。令 $S^\star_1$，$S^\star_2 \subseteq [n]$ 表示真实类别，使得$S^\star_1 \cap S^\star_2 = \emptyset$ 且 $S^\star_1 \cup S^\star_2 = [n]$。令 $\mathcal{P} = \{C_1, C_2\}$ 为将 $[n]$ 划分为两个非空簇的（当前）划分。我们定义

(i) *Class proportions*: $R^1 := \frac{|S^\star_1|}{n}, R^2 := \frac{|S^\star_2|}{n} = 1 - R^1$.

(ii) 簇 C_j 相对于类别 ℓ的纯度系数： R_j^ℓ: $= |C_j \cap S_\ell^\star|/|C_j|$。

(iii) 如果划分 $\mathcal{P}$ 在置换意义下与真实类别一致，则称其正确，也就是说，满足以下任一条件：$(C_1, C_2) = (S_1^\star, S_2^\star)$，或 $(C_1, C_2) = (S_2^\star, S_1^\star)$。

对于簇 j 或类别 ℓ，我们用 j 或 C_j 表示另一个簇，用 ℓ 或 $S_\ell^\star$ 表示另一个真实类别，因此$\bar{1} = 2$、$\bar{2} = 1$、$\bar{C}_1 = C_2$、$\bar{C}_2 = C_1$、$\bar{S}_1^\star = S_2^\star$ 和 $\bar{S}_2^\star = S_1^\star$。

2.3. 近似平衡划分

所有划分中有很大一部分在“平衡”意义上大致如此，即其大小接近 $n/2$。这可以通过概率论证证明，例如参见附录中的事实 A.10 。我们固定 n，并考察簇大小与 q 相差不超过 $n/2$个标准差的所有划分。

定义 2.5(q-近似平衡划分)。固定$n \in \mathbb{N}$ 和参数 $q > 0$。若 $\mathcal{P} = \{C_1, C_2\}$是 $[n]$ 的二划分（即 $C_1 \cap C_2 = \emptyset$ 和 $C_1 \cup C_2 = [n]$），且两个簇的大小均为$|C_k| > 2$，并且其内部标准差与 $n/2$相差不超过 q，即 $n/2 - q\sqrt{n/4} \le |C_k| < n/2 + q\sqrt{n/4}$，则称其为 q-近似平衡的，其中 $k \in \{1, 2\}$.

当 q 较大时，所有 q-近似平衡的 $[n]$ 二划分几乎包含所有二划分（仅有指数级小比例的例外），因此涵盖了“典型”情形。论证也可调整为允许 q 依赖于 n，使得当 $n \to \infty$时，该陈述对趋近于 1 的划分比例成立，见备注C.4。

3. 分析工具

本文旨在展示一种基本的高维情形：尽管底层双簇模型十分简单，Lloyd 算法仍会以很高的概3.4率失败。定理刻画了单个样本被 Lloyd 算法重新分配到新簇的概率。附录C.3中证明的直觉是，考虑一种划分：除一个误分类样本外，其余部分都与真实划分一致；人们可能直觉上会认为，这种几乎正确的初始化会在下一次 Lloyd 分配步骤中立即得到纠正。然而，我们证明，在高噪声、高维情形下，该误分类样本仍会被错误分类，因为它（以很高的概率）到其当前错误簇的经验质心的距离，比到正确簇的经验质心更近。直觉上，这种单个样本误分类是算法面临的“最有利情形”，因此可作为一个界限：附录C.3 将这一直觉转化为严谨的论证。在确立了单样本持续误分类这一现象后，我们在定理3.6中将其推广到广泛的划分族，并对其统一加以证明，然后利用并集界论证定理3.8中所有近似平衡划分（几乎涵盖所有划分）里的所有样本。

3.1. 到两个聚类中心的距离

我们对 Lloyd 和 Hartigan 算法的分析，都取决于样本与经验聚类质心之间平方距离的分布。在各向同性高斯模型下，这类平方距离（除确定性缩放外）服从自由度为 d 的卡方分布。接下来的两个引理将形式化描述该模型下这些距离的分布。引理 3.1 刻画样本到其当前所分配簇的质心的距离分布，而引理 3.2 刻画样本到另一个（竞争）簇质心的距离分布。证明见附录 C.2。

引理 3.1（到当前聚类质心的距离）。 考虑模型 *2.1* 的设定。固定 $i \in [n]$。令 $j \in \{1, 2\}$为满足下式的（当前）簇索引： $i \in C_j$，并令$\ell \in \{1, 2\}$ 为满足 $i \in S_\ell^\star$ 的真实类别索引。那么，样本 x_i 与簇 $\widehat{\mu}_j$ 的质心 C_j 之间的距离服从

$$\begin{aligned} &\|x_i - \widehat{\mu}_j\|^2 \sim \alpha_{\mathrm{cur}} \chi_d^2 \ , \\ &\alpha_{\mathrm{cur}} = 2\tau^2 (1 - R_j^\ell)^2 + (1 - 1/|C_j|)\, \sigma^2 \ . \end{aligned} \tag{10}$$

其中 R_j^ℓ 是定义 *2.4* 中定义的簇纯度。

引理 3.2（到另一个簇质心的距离）。 在引理 *3.1*的设定下，样本 x_i与簇 C_j 的质心 $\widehat{\mu}_{\bar{j}}$ 之间的距离服从如下分布：

$$\begin{aligned} &\left\|x_i - \widehat{\mu}_{\bar{j}}\right\|^2 \sim \alpha_{\mathrm{alt}} \chi_d^2 \ , \\ &\alpha_{\mathrm{alt}} = 2\tau^2 \left(1 - R_{\bar{j}}^\ell\right)^2 + \left(1 + 1/|C_{\bar{j}}|\right) \sigma^2 . \end{aligned} \tag{11}$$

我们强调，在这两个引理中，真实潜在类别标签（模型 2.1（c））和当前的簇分配（因此也包括定义 $\mathcal{P} = \{C_1, C_2\}$ 2.2中的二划分）都视为固定。唯一相关的随机性来源是真实中心和加性噪声（模型 2.1（a）–（b）），它们与类别标签和簇标签相互独立。

接下来，对上述引理中距离的比较可归结为涉及缩放后的 χ_d^2随机变量之差的事件（缩放因子分别为 α_{cur} 和α_{alt}）。以下引理给出了这类差值的一个便捷尾界；其证明基于 Chernoff 方法（Chernoff，1952），见附录 C.1。

引理 3.3。 固定 $b_1 > b_2 > 0$和 $m \in \mathbb{R}$。设 $Y_1 \sim b_1 \chi_d^2$和 $Y_2 \sim b_2 \chi_d^2$ 服从具有 d自由度的缩放卡方分布，但它们不一定相互独立。则，

$$\mathbb{P}\left(Y_1 - Y_2 \le m\right) \le \exp\left(m \frac{b_1 - b_2}{8 b_1 b_2}\right) \rho^{d/4}, \tag{12}$$

其中 $\rho = 1 - \big((b_1 - b_2)/(b_1 + b_2)\big)^2 < 1$。

3.2. Lloyd 算法分析

介绍完技术工具后，我们按照本节开头所述的策略继续进行。

3.2.1. 单样本重新分配概率

本节研究固定样本x_i 在下一次 Lloyd 迭代中改变分配的概率。设$j \in \{1,2\}$ 为其当前簇索引（因此 $i \in C_j$），并设 j表示另一个簇。当且仅当 $\|x_i - \widehat{\mu}_j\|^2 < \|x_i - \widehat{\mu}_j\|^2$时，Lloyd 会将 x_i 重新分配至 C_j；等价地，也就是差值$\|x_i - \widehat{\mu}_j\|^2 - \|x_i - \widehat{\mu}_j\|^2$ 为负时；否则，该样本在下一次迭代中仍留在 C_j。

定理 3.4 (Lloyd 算法：单样本)。 我们考虑模型 *2.1*中的设定，其中固定 $i \in [n]$当前所属簇为 C_j，另一个簇为 C_j，且满足 $i \in C_j$。我们将簇大小记为 $c := |C_j|$ 和 $c := |C_j|$。

如果噪声水平 $\sigma > 0$ 满足

$$\sigma \;>\; \frac{\sqrt{2\bar{c}}\,\tau\,(c-1)}{\sqrt{c(c+\bar{c})}}\,, \tag{13}$$

那么，我们有

$$\mathbb{P}\Big(\big\|x_i - \widehat{\mu}_{\bar{j}}\big\|^2 < \big\|x_i - \widehat{\mu}_j\big\|^2\Big) \;\le\; \rho^{\,d/4}, \tag{14}$$

其中 ρ 定义为

$$\rho(\sigma,\tau,c,\bar{c}) = \frac{4\sigma^2(c-1)c^2\bar{c}(\bar{c}+1)\big(c\big(\sigma^2+2\tau^2\big)-2\tau^2\big)}{\big(-c(\sigma^2+4\tau^2)\bar{c}+c^2(\sigma^2+2(\sigma^2+\tau^2)\bar{c})+2\tau^2\bar{c}\big)^2}, \tag{15}$$

并且满足 $0 \le \rho < 1$。

定理 3.4 的证明见附录 C.3。该证明结合了引理 3.1 和 3.2中的距离刻画，以及引理3.3中关于缩放卡方变量之差的尾界。所得估计对于所有具有规定簇大小的划分都是一致的，尤其适用于“最有利”的错误初始化，即该划分除单个误分类样本外均与真实划分一致。

备注 3.5。 方程（13）中关于噪声的条件，是指这样一个阈值：即使在“最有利”或“最容易修正”的错误初始化下，样本到其当前簇的期望距离仍会超过到另一簇的期望距离。从技术层面来说，该条件是满足引理 3.3要求所必需的条件（详见证明）。

3.2.2. 近似平衡划分中的样本

以下定理在附录C.4中证明，为所有 q-近似平衡划分（定义2.5）提供了一个统一界，从而无需显式给出簇大小。为简化该定理中的记号，我们不失一般性地固定 $\tau = 1$。

定理3.6 (q-近似平衡划分的统一界)。 考虑模型 *2.1*的设定，其中固定一个$i \in [n]$，其当前簇索引为 j。另一个簇为 j。固定一个划分不平衡因子 $q > 1$，并假设划分$\mathcal{P} = \{C_1, C_2\}$ 是 q-近似平衡的（定义 *2.5*）。固定 $\tau = 1$，并固定 $\beta > 1$，使得

$$\sigma = \beta\frac{(\sqrt{n}\,q + n - 2)}{\sqrt{2}\sqrt{\sqrt{n}\,q + n}}. \tag{16}$$

于是，

$$\mathbb{P}\left(\big\|x_i - \widehat{\mu}_{\bar{j}}\big\|^2 - \big\|x_i - \widehat{\mu}_j\big\|^2 < 0\right) \le \rho_q^{\,d/4}, \tag{17}$$

其中

$$\rho_q = \frac{\sigma^2\big(\sqrt{n}q+n-2\big)\big(\sqrt{n}q+n\big)\big(\sqrt{n}q+n+2\big)\big(\sqrt{n}\big(\sigma^2+2\big)\big(\sqrt{n}+q\big)-4\big)}{\Big(n\sigma^2\big(\sqrt{n}+q\big)^2+\big(\sqrt{n}q+n-2\big)^2\Big)^2}. \tag{18}$$

注释 3.7（定理 3.6的渐近性质）。 当 n 较大时（固定 q 和 β），定理 3.6 中的表达式更容易理解。将方程（16）中的表达式平方，并按 n展开，得到

$$\sigma^2 = \frac{\beta^2 n}{2} + \frac{\beta^2 q\sqrt{n}}{2} - 2\beta^2 + \frac{2\beta^2}{n} + O\left(n^{-3/2}\right). \tag{19}$$

将方程（16）代入方程（18），并按 n展开，得到

$$\rho_q = 1 - \frac{4\big(\beta^2-1\big)^2 n^{-2}}{\beta^4} + \frac{8\big(\beta^2-1\big)^2 n^{-5/2} q}{\beta^4} + O\left(n^{-3}\right). \tag{20}$$

定理 3.6 表明，在适当的条件下，给定样本在近似平衡划分中于 Lloyd k-均值算法的下一次迭代中转移到另一个簇的概率很小，并且会随着维度 d 增大而降低。

3.2.3. 所有近似平衡划分都是 Lloyd 算法的不动点

回顾第 2.3 节可知，绝大多数划分几乎是平衡的，并且通过设定适当的q，所有划分都是 q-近似平衡划分（簇大小为 2 或更小的划分除外）。下面将定理3.6推广为关于任意划分不是 Lloyd 算法不动点之概率的陈述（排除条件与之前相同）。

推论 3.8（主要结果：Lloyd 算法）。考虑模型2.1的设定。固定不平衡参数 $q > 1$，并假设噪声水平满足方程（16）。则存在一个 q-近似平衡划分（见定义 2.5），且该划分不是 *Lloyd* k-均值更新方案之不动点的概率，其上界为

$$\mathbb{P}(\exists \text{ 近似平衡但不是不动点的划分}) \le 2^n n \rho_q^{d/4}, \tag{21}$$

其中 ρ_q 定义见（18）。

证明见附录C.5。其思路是利用并集界扩展定理3.6：先对单个划分中的多个样本取并集界，再对多个划分取并集界。

考察证明中的距离分布，有助于阐明 Lloyd 算法失效背后的机制。当 $i \in C_j$ 时，质心 $\widehat{\mu}_j$ 是根据包含 x_i 的均值计算得出的，因此 x_i 会产生不可忽略的“自影响”，并将 $\widehat{\mu}_j$ 拉向自身，偏移量约为 $1/|C_j|$。在高噪声、高维情形下，这种自影响可能压过微弱的类别分离信号：即使划分中的其余部分都正确，分错的样本仍可能离当前（错误）簇的质心比离另一个质心更近。在足够极端的情形下，这种效应强到足以使 Lloyd 算法的更新以很高的概率连一个错误分类都无法纠正。我们还指出，在质心初始化下，同样的结论也成立：第一次分配步骤产生的划分已经是不动点（以高概率成立，且有相同的除外条件），因此后续不会再有进展。

3.3. Hartigan 算法的分析

对于 Hartigan 算法，我们采用类似的工具，但方向相反。在定理3.9中，我们界定样本 无法 迁移的概率：当其当前所在的簇中，该样本所属真实类别的比例不高于另一簇时。随后，我们在推论 3.11中界定错误划分是不动点的概率，在推论 3.11中对更广泛的参数范围建立统一界，并在推论 3.12中推知，以高概率，Hartigan 动力学不存在错误划分的不动点。

定理 3.9(Hartigan 算法：单个样本)。在模型 2.1的设定下，固定一个索引 $i \in [n]$，并令 $\ell \in \{1,2\}$为其真实类别（因此 $i \in S_\ell^\star$）。令 $j \in \{1,2\}$ 为其当前簇索引（因此 $i \in C_j$），并令 j 表示另一个簇。若簇纯度系数（定义 2.4）满足

$$0 < R_j^\ell \le R_{\bar{j}}^\ell \le 1, \tag{22}$$

然后，

$$\mathbb{P}\Big(\Delta_H^2(x_i, C_j) \le \Delta_H^2(x_i, C_{\bar{j}})\Big) \;\le\; \rho^{\,d/4}, \tag{23}$$

w此处 Δ_H 定义于方程 (9) 且 ρ给出 通过

$$\rho = 1 - \left(\frac{\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 - \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right)}{\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 + \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right)+\sigma^2} \right)^2. \tag{24}$$

并满足 $0 \le \rho < 1$。

证明推迟至附录D.2。其思路与 Lloyd 分析相同，但将通常的平方距离替换为 Hartigan 重缩放距离 $\Delta_H^2(\cdot,\cdot)$，该距离在方程（9）中定义；因此，我们使用针对 Hartigan 加权定制的引理 3.1 和 3.2 的对应版本。

特别地，划分 $\mathcal{P} = \{C_1, C_2\}$ 只有在以下条件成立时，才可能是 Hartigan 动力学的不动点：每个样本都更倾向于（按 Hartigan 的定义）归入其当前簇，而非另一个簇；因此，只需找到一个指标 i，使得一次 Hartigan 移动能够严格改善结果。因此，下一个推论可直接由定理3.9得出。

推论 3.10。若满足定理3.9 的条件，则划分 $\mathcal{P}$ 是 *Hartigan* 算法不动点的概率至多为

$$\mathbb{P}(\mathcal{P} \textit{ is a fixed point}) \;\le\; \rho^{\,d/4}, \tag{25}$$

其中 ρ由方程 (24)给出。

定理 3.9 中的界取决于具体的簇大小及其对应的纯度系数。以下推论在附录D.3中得到证明，将这一逐点估计推广为一致性结论：它给出了一个界，该界同时适用于所有可容许的簇大小，以及所有对应于不与真实类别（在置换意义下）一致的划分的纯度配置。

推论 3.11。考虑模型2.1中的设定。进一步假设 $n \ge 4$。如果当前划分 $\mathcal{P}$ 非空且为错误划分（定义 2.4），那么，$\mathcal{P}$ 为不动点的概率有如下界：

$$\mathbb{P}\left(\mathcal{P} \textit{ is a fixed point}\right) \le \rho_h^{\,d/4}, \tag{26}$$

其中 ρ_h 由下式给出

$$\rho_h = 1 - \left(\frac{4\tau^2 (R^\star)^2 n^{-1}}{3\tau^2 + \sigma^2}\right)^2 < 1\,, \tag{27}$$

且 $R^\star = \min(R^1, R^2)$ 是最小真实类别的相对大小。

推论3.11 中的这一一致界使得我们可以对错误的二分划分族进行联合界，并引出下一个推论（证明见附录D.4），该推论给出了任意与真实划分（在置换意义下）不匹配的划分是该算法不动点的概率上界。

推论 3.12（主要结果：Hartigan 算法）。考虑模型2.1的设定。假设 $n \geq 4$。则存在任一非空错误划分（定义 2.4）是 *Hartigan* 算法不动点的概率有上界

$$\mathbb{P}(\exists \mathcal{P} \text{ 非空错误划分}) \leq 2^n \rho_h^{d/4}, \tag{28}$$

其中 ρ_h 由下式给出（27）。

4. 数值结果

本节报告数值实验，展示 Lloyd k-均值算法在本文研究的高噪声、高维情形中的失效。除了 Lloyd 方法，我们还评估了 Hartigan 算法和几种现代替代方法：常见的高维启发式方法主成分分析+k-均值，即先应用主成分分析，再在降维后的空间中运行 Lloyd（Zha 等人，2001年；Ding 和 He，2004）；现代 k-均值半定规划松弛（Mixon 等人，2016年）以及谱聚类（Shi 和 Malik，2000年；Ng 等人，2001年），均通过标准的 scikit-learn 实现运行（Pedregosa 等人，2011年）。

更多数值结果见附录 F。实现细节见附录 E。我们的代码可在以下地址免费获取：
h
ttps://github.com/Lederman-Group/Catastrophic_Failure_KMeans。

4.1. 合成高斯混合模型

第一组实验展示了本文的理论发现与 Hartigan 和 Lloyd 的 k-均值算法的性能之间的联系，并将它们与其他聚类算法进行了比较。我们从模型 2.1（推广到 $K \geq 2$）所定义的高斯混合模型中采样，设置不同的维度 d 和噪声方差 σ^2，取值为 $K = 2$、5和 10 个簇、$n = 20 \times K$ 个样本，以及 $\tau^2 = 1$。我们通过两种方式衡量各算法的性能：i）根据标准化互信息得分（定义A.13）评估其恢复真实标签的效果；ii）计算解的 k-均值损失。

在每次实验中，我们都会生成一个新的数据集，并运行每种聚类算法，且只进行一次随机初始化。我们对每种取值组合重复实验 100 次，其中取值为 d、σ^2 和 K。对于 Lloyd 和 Hartigan 算法，以及 PCA+k-均值算法，我们评估了三种不同初始化策略的性能：i)“随机划分”初始化，即随机选取大小相同的簇，并将初始中心设为相应簇的均值；ii)“随机中心”初始化，即指定 K 随机样本作为中心；以及 iii）流行的 k-*means*++ 初始化（Arthur &Vassilvitskii，2006）。

在图1中，我们展示了各方法与真实划分之间的 NMI；在此图中，我们仅关注 k均值++：对于 Lloyd 算法，使用该算法的情况；对于 Hartigan 算法，则使用随机平衡划分。总体而言，Hartigan 的 k均值和谱聚类在所有K取值下都达到最高准确率。半定规划松弛在 $K = 2$时表现良好，但随着 K 增大，在较高噪声水平下性能下降更为明显。Lloyd 算法在高噪声、高维度情况下表现最差，不过，简单的主成分分析预处理步骤（PCA+ k均值）显著提高了其准确率。补充图2见附录F.2。

我们注意到 k-*means*++用于图1中的结果所采用的 Lloyd 算法，通过将质心作为初始猜测来为算法提供初始值，而我们对 Lloyd 算法的分析则考虑划分。附录F.2中展示了对其他初始化策略的初步实证评估，结果表明 k-*means*++作为 Lloyd 算法的初始化方式，优于随机划分。我们将基于质心的初始化详细研究留待未来工作；这里仅作非正式说明：在某些情况下，良好的初始中心可以产生良好的第一个划分，而该划分已经是算法的不动点。我们认为，这种情况下算法本身“什么也没做”，只是直接根据初始化结果进行了划分。图1说明，即使采用这种有利的初始化方式，Lloyd 算法的表现仍不如其他算法。

4.2 真实世界数据集

我们在四个真实世界数据集上比较这些聚类算法：包含 40 个人在 10 种不同姿势下图像的奥利维蒂人脸数据集（Samaria & Harter，1994年），以及从 20 Newsgroups数据集（Mitchell，1997年）衍生出的三个数据集，分别记为 20NG-A、20NG-B 和 20NG-C，其 $K = 2$、5和分别为10。关于这些数据集的更多详情见附录E。

对于每个数据集，我们运行 Lloyd 和 Hartigan 的 k均值算法、谱聚类和半定规划。对于每种算法（确定性的半定规划除外），我们从 k次独立初始化中，选取产生最小 k均值损失（方程（5））的结果。由于半定规划是确定性的，因此只运行一次，直到收敛或达到最多 2000 次迭代。在所有实验中，我们使用 k均值++对 Lloyd 的k均值算法和谱聚类进行初始化，并使用随机平衡划分对 Hartigan 的 k均值算法进行初始化。表1总结了每个数据集的实验结果。

我们注意到，尽管在奥利维蒂示例中，谱聚类的 NMI 表现优于 Hartigan 算法，但 Hartigan 算法取得了更低的损失，而这正是该算法旨在优化的指标。

5. 讨论与结论

我们对 Lloyd k-均值算法的分析，为其在高维、高噪声区域中的经验性失效提供了具体解释。在这种情形下，对于抽样数据集，以高概率而言，每个近似平衡划分都是 Lloyd 更新映射的不动点。因此，除非初始划分极度不平衡，否则算法在第一次更新后便会停止，并返回初始划分（若采用质心初始化，则返回由初始中心决定的划分），几乎不会从起点取得任何进展。尽管人们早已知道 k-均值算法对初始化很敏感（Balanov 等人，2025），我们的结果揭示了一种极端情形：初始化几乎就是最终结果——Lloyd 算法几乎不会从起点取得任何进展，并返回初始划分。

我们所得到的理论界较为保守，旨在证实这一现象的存在，而非精确确定尖锐阈值。未来工作的一个有趣方向是理解这种不动点激增如何随噪声或维度的降低而减弱，并量化即使并非所有划分都是不动点时，Lloyd 动力学陷入次优不动点的概率。

与之形成鲜明对比的是，Hartigan 算法并不表现出那种使 Lloyd 更新陷入困境的不动点激增。在我们的双分量高斯分布模型中，一旦维度足够大（此时聚类任务变得更容易），我们证明 Hartigan 的贪心局部改进动力学以高概率不存在错误不动点，因而会终止于正确划分（在标签置换意义下）。

与 Lloyd 分析一样，我们的保证较为保守，维度阈值也很可能远未达到紧确的程度。从经验上看，Hartigan 算法在低得多的维度下仍然稳健得多，其性能似乎常常可与 Lloyd 方法的强大现代替代方案相媲美。这些替代方案中有许多在扩展性方面要困难得多；例如，半定规划松弛需要处理 $n \times n$ 矩阵，而 Hartigan 算法与 Lloyd 方法一样，只依赖反复比较样本与经验质心。要仔细研究运行时间与准确率的权衡，需考虑实现细节，这超出了本文的范围；不过，我们的发现引出了一个理论问题：能否针对高维混合模型及更广泛的场景，为 Hartigan 算法获得更精确的复杂度保证和类似半定规划的恢复保证？本文中的保守界限能否大幅收紧？

尽管我们的理论聚焦于双簇情形，第4节中的实证结果表明，Lloyd 与 Hartigan 之间同样的定性分离在簇数量更多时依然存在。将分析推广到一般的 K看来是可行的：对于 Lloyd，不动点机制预计只需稍作修改即可延续，而对 Hartigan 的相应推广则需要额外注意（例如，处理多个相互竞争的移动以及跨 K 个簇的簇规模效应）。我们将对两种算法的完整 $K > 2$理论处理留待未来工作。

例如，正如 Bottou &Bengio (1994年) 和 Bishop& Nasrabadi (2006年) 所指出的，Lloyd 算法与期望最大化（EM）算法（Dempster 等人，1977）密切相关；后者也是统计学习和数据分析中的一项基础工具（Wu 等人，2008）。我们研究的这一效应似乎也适用于 EM 算法；在高噪声、高维情境下，EM 同样可能陷入初始化附近。关于这一问题的详细分析也留待未来工作。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能产生许多社会影响，但我们认为没有哪一项需要在此特别强调。

致谢

作者谨此感谢 Amit Singer、Fred Sigworth、Sheng Xu、Zhou Fan 和 Yihong Wu 提供的宝贵讨论。作者谨此感谢耶鲁大学研究计算中心（YCRC）

表*1.* **使用 Lloyd 算法和 Hartigan 算法、 k-均值、谱聚类和半定规划对真实世界数据集进行聚类的结果。** 每个数据集针对 Lloyd 的 k-均值、Hartigan 的 k-均值和谱聚类进行 500 次随机初始化；半定规划因具有确定性而只运行一次。随着 d 与 n 之比增大，Lloyd 算法的表现会比其他算法更差。对于每种算法，我们报告对应于达到最低 k-均值损失的划分的值（方程（5））和 NMI（见定义 A.13）。 k-均值损失。

数据集参数				k-均值损失				NMI			
	n	d	K	Lloyd	Hartigan	SDP	谱聚类	Lloyd	Hartigan	SDP	谱聚类
奥利维蒂	400	4096	40	8.53	**8.11**	8.85	8.95	0.74	0.77	0.72	**0.83**
20NG-A	200	5000	2	193.72	**193.46**	193.54	193.48	0.27	**0.54**	0.48	0.52
20NG-B	500	5000	5	484.04	**481.72**	484.29	482.89	0.24	**0.44**	0.34	0.37
20NG-C	1000	5000	10	957.43	**951.96**	956.88	953.67	0.23	**0.31**	0.27	0.27

提供计算资源与支持。本研究得到美国国立卫生研究院/美国国立普通医学科学研究所（1R35GM157226）、阿尔弗雷德·P·斯隆基金会（FG-2023-20853）和西蒙斯基金会（1288155）的支持。

参考文献

Aloise, D., Deshpande, A., Hansen, P. 和 Popat, P. 欧几里得平方和聚类的 NP 难度。《机器学习》，75:245–248，2009年。

Arthur, D. 和 Vassilvitskii, S. k-均值++：谨慎进行种子初始化的优势。技术报告，斯坦福大学，2006年。

Balanov, A., Bendory, T. 和 Huleihel, W. 高斯混合模型中的确认偏误。《*IEEE* 信息论汇刊》，2025年。Bishop, C. M. 和 Nasrabadi, N. M.《模式识别与机器学习》，第4卷。施普林格，2006年。Bock, H.-H. 聚类分析中 k 均值算法的起源与发展。《概率与统计史电子期刊》，4(2):1–18，2008年。Bottou, L. 和 Bengio, Y. k 均值算法的收敛性质。《神经信息处理系统进展》，第7卷，1994年。Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S. 和 Zhang, Q. JAX：可组合的 Python+NumPy 程序变换，2018年。网址：http://github.com/jax-ml/jax。Chen, X. 和 Yang, Y. 高斯混合模型精确恢复的截断阈值。《*IEEE* 信息论汇刊》，67(6):4223–4238，2021年。Chernoff, H. 一种基于观测值之和的假设检验渐近效率度量。《数学统计年刊》，第493–507页，1952年。Dempster, A. P., Laird, N. M. 和 Rubin, D. B. 通过期望最大化算法从不完整数据中进行最大似然估计。《英国皇家统计学会会刊：*B*辑（方法论）》，39(1):1–22，1977年。Diamond, S. 和 Boyd, S. CVXPY：一种嵌入 Python 的凸优化建模语言。《机器学习研究期刊》，17(83):1–5，2016年。Ding, C. 和 He, X. 通过主成分分析进行 k-均值聚类。载于第二十一届国际机器学习会议论文集，第29页，2004年。

Forgy, E. W. 多元数据的聚类分析：分类的效率与可解释性。生物统计学，21：768–769，1965年。Gao, C. 和 Zhang, A. Y. 用于离散结构恢复的迭代算法。《统计年鉴》，50(2)：1066–1094，2022年。Hartigan, J. A.聚类算法。John Wiley & Sons, Inc.，1975年。Hartigan, J. A. 和 Wong, M. A. 算法 AS 136：一种 k-均值聚类算法。《皇家统计学会杂志 *C* 辑（应用统计学）》，28(1)：100–108，1979年。Lam, S. K.、Pitrou, A. 和 Seibert, S. Numba：基于 LLVM 的 Python 即时编译器。载于第二届高性能计算 *LLVM* 编译器基础设施研讨会论文集，第 1–6 页，2015年。Lloyd, S. 脉冲编码调制中的最小二乘量化。《*IEEE* 信息论汇刊》，28(2)：129–137，1982年。Lu, Y. 和 Zhou, H. H. Lloyd 算法及其变体的统计与计算保证。*arXiv* 预印本 *arXiv:1612.02099*，2016年。MacQueen, J. 多元观测。载于第五届伯克利数学统计与概率研讨会论文集，第 1 卷，第 281–297 页，1967年。Mitchell, T. 20个新闻组。UCI 机器学习知识库，1997年。DOI：https://doi.org/10.24432/C5C323。Mixon, D. G.、Villar, S. 和 Ward, R. 通过半定规划对次高斯混合进行聚类，2016年5月。Mousavi, N. Chernoff 界有多紧？未发表的手稿，2010年。网址 https://ece.uwaterloo.ca/~nmousavi/Papers/Chernoff-Tightness.pdf。Ndaoud, M. 双成分高斯混合模型中的精确最优恢复。《统计年鉴》，50(4)：2096–2126，2022年。Ng, A.、Jordan, M. 和 Weiss, Y. 关于谱聚类：分析与算法。《神经信息处理系统进展》，14，2001年。Pedregosa, F.、Varoquaux, G.、Gramfort, A.、Michel, V.、Thirion, B.、Grisel, O.、Blondel, M.、Prettenhofer, P.、Weiss, R.、Dubourg, V.、Vanderplas, J.、Passos, A.、Cournapeau, D.、Brucher, M.、Perrot, M. 和 Duchesnay, E. Scikit-learn：用于 Python 的机器学习。《机器学习研究期刊》，12：2825–2830，2011年。

Peng, J. 和 Wei, Y. 通过半定规划近似 k-均值类聚类。*SIAM* 优化期刊，18(1)：186–205，2007年。

Samaria, F. S. 和 Harter, A. C. 人脸识别随机模型的参数化。载于 *1994* 年 *IEEE* 计算机视觉应用研讨会论文集，第 138–142 页。IEEE，1994年。Shi, J. 和 Malik, J. 归一化割与图像分割。*IEEE* 模式分析与机器智能汇刊，22(8):888–905，2000 年 8 月。ISSN 1939-3539。doi：10.1109/34.868688。Slonim, N.、Aharoni, E. 和 Crammer, K. Hartigan 的 k-均值与 Lloyd 的 k-均值——是时候改变了吗？载于第 *23* 届国际人工智能联合会议（*IJCAI*）论文集，2013 年。Steinley, D. k-均值聚类：半世纪综述。英国数学与统计心理学杂志，59(1):1–34，2006年。Telgarsky, M. 和 Vattani, A. Hartigan 方法：无需 Voronoi 的 k-均值聚类。载于第十三届人工智能与统计国际会议论文集，第 820–827 页。《JMLR》研讨会与会议论文集，2010年。Wu, X.、Kumar, V.、Ross Quinlan, J.、Ghosh, J.、Yang, Q.、Motoda, H.、McLachlan, G. J.、Ng, A.、Liu, B.、Yu, P. S. 等。数据挖掘十大算法。知识与信息系统，14(1):1–37，2008 年。Zha, H.、He, X.、Ding, C.、Gu, M. 和 Simon, H. k-均值聚类的谱松弛。《神经信息处理系统进展》，14，2001年。

附录

附录结构。我们首先在附录A中汇总若干标准结果，以便完整呈现并供后续参考。附录B 以伪代码形式介绍本文研究的聚类算法。与 Lloyd 算法相关的主要结果证明见附录 C，而 Hartigan 算法的相应证明见附录 D。其他实现细节，包括辅助聚类方法的定义和数据集预处理步骤，见附录 E。最后，附录 F 报告了进一步的数值实验和补充讨论，用以补充第 4节中的结果。

A. 标准结果与定义

以下是统计学中的标准教科书事实。

事实A.1（高斯变量的相加）。设 ξ_1和 ξ_2为服从正态分布 ξ_1，$\xi_2 \sim \mathcal{N}(0,\ 1)$ 的独立同分布变量，并设a_1，a_2，b_1，$b_2 \in \mathbb{R}$。则，

$$a_1 + b_1\xi_1 + a_2 + b_2\xi_2 \sim \mathcal{N}(a_1 + a_2, b_1^2 + b_2^2) \tag{29}$$

事实A.2。设 $X \sim \mathcal{N}(0,\ \sigma^2)$ 服从正态分布。设 $-\infty < t < 1/2$。则，

$$\mathbb{E}\left(\exp\left(tX^2\right)\right) = \frac{1}{\sqrt{1-2t}}. \tag{30}$$

事实 A.3（Cochran定理的特殊情形）。 设 $X \sim \mathcal{N}(0, \sigma^2 I_d)$ 是一个 d 维高斯随机向量，其均值为零，协方差矩阵为 $\sigma^2 I_d$，其中 I_d 是 $d \times d$ 单位矩阵。那么

$$\|X\|^2 \sim \sigma^2 \chi_d^2, \tag{31}$$

其中 χ_d^2 表示自由度为 d 的卡方分布。

T上述事实可用于计算 χ^2 分布的矩母函数。

事实 A.4（ χ_d^2 的矩母函数）。 设 $X \sim a\chi_d^2$。 那么 $\mathbb{E}(X) = da$ ，且 $\mathrm{Var}(X) = 2da^2$。设 $t < 1/2$。那么

$$M_X(t) = \mathbb{E}\left(\exp\left(tX\right)\right) = (1-2t)^{-d/2}. \tag{32}$$

*F*事实 A.5（马尔可夫不等式）。 设 X 是一个非负随机变量。那么对于任意 $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}. \tag{33}$$

事实 A.6（柯西-施瓦茨不等式）。 设 X 和 Y 是随机变量。那么，

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2). \tag{34}$$

事实 A.7（切比雪夫不等式）。 设 X 是一个具有有限方差 $\sigma^2 > 0$ 和有限均值的可积随机变量。那么对于任意 $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a\sigma) \leq \frac{1}{a^2}. \tag{35}$$

事实 A.8（霍夫丁不等式）。 设 X_1，X_2，……，X_n是满足 $a_i \leq X_i \leq b_i$几乎必然成立的独立同分布随机变量。令$S = \sum_{i=1}^n X_i$。那么对于任意 $t > 0$,

$$\mathbb{P}(|S - \mathbb{E}(S)| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{36}$$

事实 A.9（计数分割）。 将 n 个元素划分为两个有标记集合，共有 2^n 种方式。第一个集合恰好含有 S 个元素（另一个集合含有 $n-S$ 个元素）的划分所占比例为 $\binom{n}{S}/2^n$。因此，在均匀随机划分下，簇大小 S 的分布为二项分布（n，$1/2$），均值为 $n/2$，方差为 $n/4$。等价地，均匀随机地选择一个划分，等同于以概率 $1/2$ 独立地将每个元素分配到第一个集合（否则分配到第二个集合）。

事实 A.10（计数典型划分）。 对于 n 的如下划分：将其划分为2 个已标识集合，且第一个集合恰好包含 S 个元素（另一个集合包含 $n-S$ 个元素），这些划分所占的比例由应用于二项分布的霍夫丁不等式（事实A.8）给出上界：

$$\mathbb{P}(|S-n/2| \geq q\sqrt{n}/2) \leq 2\exp\left(-\frac{q^2}{2}\right) \tag{37}$$

更通俗地说，对于较大的 n， n划分为 2 个已识别集合的划分中，几乎所有划分的每个集合都包含约 $n/2$ 个元素。事实 A.11（并集界）。设 $A_1, A_2, \ldots, A_n$ 为事件。那么，

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{38}$$

定义 A.12 （二项比例置信区间的 Wilson 区间）。 本文中比例估计的误差线使用 Wilson 区间计算。我们选择这种方法来计算置信区间，因为当预测比例接近 1 或 0 时，它仍然稳健；在这种情况下，无论样本数量多少，其他计算置信区间的方法都会给出宽度为零的区间。置信区间定义如下：

$$CI = (\text{center} - \text{width}, \text{center} + \text{width}) \tag{39}$$

$$\text{center} = \frac{n_s + \frac{1}{2}z_\alpha^2}{n + z_\alpha^2} \tag{40}$$

$$\text{width} = \frac{z_\alpha}{n + z_\alpha^2}\sqrt{\frac{n_s n_f}{n} + \frac{z_\alpha^2}{4}}\,, \tag{41}$$

其中， n 是实验次数， n_s 和 n_f 分别是成功和失败的次数。z_α 的值是标准正态分布的 $1-\frac{\alpha}{2}$。在使用 Wilson 区间的图中，我们绘制实际估计比率n_s/n，并省略 Wilson 区间的中心值。

定义 A.13 (归一化互信息**)**。 互信息量化两个随机变量之间的依赖关系。在本文中，我们将其应用于离散随机变量，不过也存在针对连续随机变量的定义。设 X, Y为两个离散随机变量，其联合概率密度函数为 $P_{(X,Y)}$。对于离散情形，互信息定义如下：

$$I(X;Y) = \sum_{x\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \mathbb{P}_{(X,Y)}(x,y)\log\left(\frac{\mathbb{P}_{(X,Y)}(x,y)}{\mathbb{P}_X(x)\mathbb{P}_Y(y)}\right)\,, \tag{42}$$

其中， P_X 和 P_Y 分别是 X 和 Y的边缘概率密度函数。

为了将通过 k-均值获得的划分与真实划分进行比较，我们使用归一化互信息（NMI）。我们使用 scikit-learn 提供的实现计算 NMI，该实现将互信息（方程 (42)）归一化到 0 至 1 的范围，其中 1 表示完全相关，0 表示不存在依赖关系。

B. k-均值算法

B.1. Lloyd 的 k-均值算法

算法 1 描述了 Lloyd 的 k-均值算法（Lloyd，1982）。

注释 B.1（Lloyd 算法初始化）。我们注意到，初始化该算法有两种方法：使用初始划分（如本文所述），或使用初始猜测中心。对于基于初始中心的初始化，算法的第一次迭代会生成划分：本文中的论证证明，在本文所讨论的设置下，算法生成的第一个划分就是该算法的不动点（以高概率成立，但极不平衡的划分可能例外）。因此，除非初始猜测足够好，能够立即生成正确划分（或者可能通过某些异常不平衡的划分做到这一点），否则该算法不会收敛到正确划分。

算法 1 Lloyd 的 k均值算法

1：输入：数据集 $X = \{x_1, x_2, \ldots, x_n\}$，簇的数量 K，基于索引的初始划分 $C = \{C_1, C_2, \ldots, C_K\}$
2 $\widehat{\mu}_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad j = 1 \ \ldots \ K$：计算初始簇质心：对于 , , ：重复 4：
3
已更改 ←false 5：对于数据集 X 中的每个样本 i , 执行 6：设 C_m 为包含 i 的簇 7：对于每个簇 C_j , 执行 8：计算到簇质心的距离 9：$\Delta^2_{\text{Lloyd}}(x_i, C_j) = \|x_i - \widehat{\mu}_j\|^2$10：结束循环 11：查找 $j_{\text{new}} = \arg\min_j \Delta^2_{\text{Lloyd}}(x_i, C_j)$ 12：如果 $j_{\text{new}} \neq m$，则 13：将 i 从 C_m 移至 $C_{j_{\text{new}}}$14：已更改 ←true 15：结束条件 16：结束循环 17：更新簇质心：$\widehat{\mu}_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$对于 $j = 1, \ldots, K$18：直到已更改 = false 19：返回：簇 $C_1, C_2, \ldots, C_K$ 和中心 $\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_K$

—

B.2. Hartigan k-均值算法

Hartigan k-均值算法在算法 2 中进行了描述。我们注意到，当其中一个子集为空时，该算法未定义；与 Lloyd 算法不同，若初始化不含空簇，Hartigan 算法无法达到空簇状态。

基于方程（9）中加权距离的 Hartigan 分配准则，事实上等价于在当前簇下所有可能的分配中（不重新分配任何其他样本）进行贪心重新分配，以最小化 k-均值损失（方程（5））；Hartigan 准则在计算上比直接朴素地计算该损失更为高效。详情参见（Hartigan，1975）。该算法的一个更高效版本见（Hartigan & Wong，1979）。

算法 2 Hartigan k-均值算法

输入： 数据集 $X = \{x_1, x_2, \ldots, x_n\}$，簇的数量 K，初始划分 $C = \{C_1, C_2, \ldots, C_K\}$ 基于 the 索引。2 3 $\widehat{\mu}_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i \quad j = 1 \ldots K$: 计算初始簇质心: 对于 , , : 重复执行 4: 已更改 $\leftarrow$false 5: 对于数据集 X 中的每个样本 i，执行 6: 令 C_m 为包含 i 的簇 7: 如果 $|C_m| > 1$，则 8: 对于每个簇 C_j，执行 9: 如果 $m = j$，则 10: 计算相对于当前簇 j 的 Hartigan 加权距离 11: $\Delta^2_{\text{Hartigan}}(x_i, C_j) = \frac{|C_j|}{|C_j|-1}\|x_i - \widehat{\mu}_j\|^2$12: 否则 13: 计算相对于备选簇 j 的 Hartigan 加权距离 14: $\Delta^2_{\text{Hartigan}}(x_i, C_j) = \frac{|C_j|}{|C_j|+1}\|x_i - \widehat{\mu}_j\|^2$15: 结束如果 16: 结束循环 17: 令 $j_{\text{new}} =$ 为 $\arg\min_j \Delta^2_{\text{Hartigan}}(x_i, C_j)$ 18: 如果 $j_{\text{new}} \neq m$ 且 $\Delta^2_{\text{Hartigan}}(x_i, C_{j_{\text{new}}}) < \Delta^2_{\text{Hartigan}}(x_i, C_m)$，则 19: 将 i 从 C_m 移至 $C_{j_{\text{new}}}$20: 更新簇质心 $\widehat{\mu}_m$ 和 $\widehat{\mu}_{j_{\text{new}}}$21: 已更改 $\leftarrow$true 22: 结束如果 23: 结束如果 24: 结束循环 25: 直到已更改 $=$ false 26: 返回: 簇 $C_1, C_2, \ldots, C_K$ 和中心 $\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_K$

C. Lloyd 算法证明

本节包含正文中与 Lloyd 算法相关的定理、引理和推论的证明，以及一些额外的辅助结果和说明。为方便起见，我们在证明之前重述这些定理、引理和推论。因此，文本和编号会有一些冗余——某些方程编号可能看起来不连贯——但这样或许能让证明更易于理解。

C.1. 引理3.3的证明

我们重述引理3.3 并给出证明。

引理 3.3。 固定 $b_1 > b_2 > 0$ 和 $m \in \mathbb{R}$。设 $Y_1 \sim b_1\chi^2_d$ 和 $Y_2 \sim b_2\chi^2_d$ 服从具有 d 个自由度的尺度卡方分布，二者不必相互独立。

个自由度，不一定相互独立。接下来，

$$\mathbb{P}\left(Y_1 - Y_2 \leq m\right) \leq \exp\left(m\frac{b_1 - b_2}{8b_1b_2}\right)\rho^{d/4}, \tag{12}$$

其中 $\rho = 1 - \left((b_1 - b_2)/(b_1 + b_2)\right)^2 < 1$。

证明。 有

$$\begin{aligned}&\mathbb{P}\left(Y_1 - Y_2 - m \leq 0\right)\\ &= \mathbb{P}\left(-(Y_1 - Y_2 - m) \geq 0\right)\\ &= \mathbb{P}\left(\exp\left(-t(Y_1 - Y_2 - m)\right) \geq 1\right) \quad \text{for all } t > 0.\end{aligned} \tag{43}$$

根据马尔可夫不等式（方程（33））：

$$\begin{aligned}\mathbb{P}\left(Y_1 - Y_2 - m \leq 0\right) \leq &\mathbb{E}\left(\exp\left(-t(Y_1 - Y_2 - m)\right)\right)\\ &= \exp\left(tm\right)\mathbb{E}\left(\exp\left(-tY_1\right)\exp\left(tY_2\right)\right)\end{aligned} \tag{44}$$

根据柯西-施瓦茨不等式（方程（34））：

$$\mathbb{P}\left(Y_1 - Y_2 - m \leq 0\right) \leq \exp\left(tm\right)\sqrt{\mathbb{E}\left(\exp\left(-2tY_1\right)\right)\mathbb{E}\left(\exp\left(2tY_2\right)\right)}\,. \tag{45}$$

接下来，根据方程（32），若进一步假设 $-2tb_1 < 1/2$且 $2tb_2 < 1/2$ （我们将证明在最优 t时满足这些条件），则有：

$$\begin{aligned}\mathbb{P}\left(Y_1 - Y_2 - m \leq 0\right) &\leq \exp\left(tm\right)\sqrt{\left(1 + 4tb_1\right)^{-d/2}\left(1 - 4tb_2\right)^{-d/2}}\\ &= \exp\left(tm\right)\left(\left(1 + 4tb_1\right)\left(1 - 4tb_2\right)\right)^{-d/4}\end{aligned} \tag{46}$$

表达式 $(4b_1t + 1)\,(1 - 4b_2t)$ 在以下位置达到最大值

$$t_{\max} = \frac{b_1 - b_2}{8b_1b_2}. \tag{47}$$

对 $t_{\max}$ 处的表达式进行简单代数运算，我们得到：

$$\left(\left(1 + 4t_{\max}b_1\right)\left(1 - 4t_{\max}b_2\right)\right) = \frac{\left(b_1 + b_2\right)^2}{4b_1b_2}\,, \tag{48}$$

and

$$\left(\frac{\left(b_1 + b_2\right)^2}{4b_1b_2}\right)^{-1} = 1 - \left(\frac{b_1 - b_2}{b_1 + b_2}\right)^2\,. \tag{49}$$

将方程 (47)、(48) 和 (49) 代入方程 (46，即可得到所需结果。

还需检查方程 (46) 的假设在 $t_{\max}$处是否成立。第一个假设 $-2tb_1 < 1/2$ 成立，因为 $b_1 > b_2 > 0$:

$$-2t_{\max}b_1 = -\frac{b_1 - b_2}{4b_2} = \frac{b_2 - b_1}{4b_2} = \frac{1}{4} - \frac{b_1}{4b_2} < 1/2. \tag{50}$$

第二个假设 $2tb_2 < 1/2$成立，因为 $b_1 > b_2 > 0$:

$$2t_{\max}b_2 = \frac{b_1 - b_2}{4b_1} < \frac{b_1}{4b_1} < 1/2. \tag{51}$$

□

注记 C.1（直观）。 在引理 3.3 的假设下，我们有变量 $Y_1 \sim b_1\chi^2_d$ 与 $Y_2 \sim b_2\chi^2_d$。 $\tilde{Z}_1$ 的期望值是 $\mathbb{E}\left(\tilde{Z}_1\right) = d \cdot b_1$， $\tilde{Z}_2$ 的期望值是 $\mathbb{E}\left(\tilde{Z}_2\right) = d \cdot b_2$。由于 $b_1 > b_2$，~ ~Z_1 的质心大于 Z_2的质心。随着 d 增大，这些分布更加集中在其均值附近，因此我们预期差值 $Y_1 - Y_2$ 为正的概率会随 d增大。该引理指出，负值的概率按 $\left(1 - \left(\frac{b_1 - b_2}{b_1 + b_2}\right)^2\right)^{d/4}$ 递减。换言之，由于 $\left(1 - \left(\frac{b_1 - b_2}{b_1 + b_2}\right)^2\right) < 1$，我们便可使 $Y_1 - Y_2 \leq m$ 的概率任意小，只需增大 d。

C.2. 第 3.1 节的证明：3.1到聚类中心的距离分布

C.2.1. 到当前簇中心的距离

我们重述引理 3.1 并给出证明。

引理 3.1 (到当前聚类质心的距离**)**。 考虑模型 *2.1*的设定。固定 $i \in [n]$。令 $j \in \{1,2\}$ 为满足 $i \in C_j$的（当前）簇索引，并令 $\ell \in \{1,2\}$ 为满足 $i \in S_\ell^\star$的真实类别索引。那么，样本 x_i 与簇 C_j 的质心 $\widehat{\mu}_j$ 之间的距离服从

$$\begin{aligned} &\|x_i - \widehat{\mu}_j\|^2 \sim \alpha_{\mathrm{cur}} \chi_d^2 \ , \\ &\alpha_{\mathrm{cur}} = 2\tau^2 (1 - R_j^\ell)^2 + (1 - 1/|C_j|)\, \sigma^2 \ . \end{aligned} \tag{10}$$

其中 R_j^ℓ 是定义 *2.4* 中定义的簇纯度 *2.4*。

证明。 考虑模型 2.1的设定，并固定真实簇分配 $\{z_i^\star\}$，如模型 2.1(c) 所述，同时固定当前簇分配 z，如定义 2.2所述。不失一般性（W.L.O.G.），假设当前被分配到簇 C_j的样本 属于真实类别 ℓ，即 $z_i^\star = \ell$。

样本 x_i 与其当前所属簇 C_j 的质心 $\widehat{\mu}_j$ 之间的差为:

$$x_i - \widehat{\mu}_j = \mu_\ell^\star + \xi_i - \widehat{\mu}_j \tag{52}$$

$$= \mu_\ell^\star + \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j} x_k \tag{53}$$

$$= \mu_\ell^\star + \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j} \left(\mu_{z_k^\star}^\star + \xi_k \right) \tag{54}$$

$$= \left(1 - \frac{|C_j \cap S_\ell^\star|}{|C_j|} \right) \mu_\ell^\star - \frac{|C_j \cap S_{\overline{\ell}}^\star|}{|C_j|} \mu_{\overline{\ell}}^\star + \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j} \xi_k \tag{55}$$

使用定义（2.4）中引入的记号 $R_j^\ell = |C_j \cap S_\ell^\star| / |C_j|$，方程（55）可改写为

$$x_i - \widehat{\mu}_j = (1 - R_j^\ell) \mu_\ell^\star - R_j^{\overline{\ell}} \mu_{\overline{\ell}}^\star + \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j} \xi_k \tag{56}$$

$$= (1 - R_j^\ell) - (1 - R_j^\ell) \mu_{\overline{\ell}}^\star + \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j} \xi_k \tag{57}$$

$$= (1 - R_j^\ell)(\mu_\ell^\star - \mu_{\overline{\ell}}^\star) + \left(1 - \frac{1}{|C_j|} \right) \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j \setminus \{i\}} \xi_k . \tag{58}$$

在方程（57）中，我们使用了 $R_j^\ell = 1 - R_j^{\overline{\ell}}$；方程（58）则是通过将求和拆分为 $\sum_{k \in C_j} \xi_k = \xi_i + \sum_{k \in C_j \setminus \{i\}} \xi_k$ 得到的。根据模型 2.1(a) 中 $\mu_\ell^\star$的定义，方程（58）中的第一项的方差为

$$\mathrm{Var}\left[(1 - R_j^\ell)(\mu_\ell^\star - \mu_{\overline{\ell}}^\star) \right] = (1 - R_j^\ell)^2 \, \mathrm{Var}\left(\mu_\ell^\star - \mu_{\overline{\ell}}^\star \right) = 2(1 - R_j^\ell)^2 \tau^2 . \tag{59}$$

此外，由于 ξ_i 在模型 2.1中定义，且噪声项独立同分布，方差为 σ^2，方程（58）中的第二项（噪声项）满足

$$\begin{aligned} \mathrm{Var}\left[\left(1 - \frac{1}{|C_j|} \right) \xi_i - \frac{1}{|C_j|} \sum_{k \in C_j \setminus \{i\}} \xi_k \right] &= \left(1 - \frac{1}{|C_j|} \right)^2 \sigma^2 + \frac{1}{|C_j|^2} \left(|C_j| - 1 \right) \sigma^2 \\ &= \left(1 - \frac{1}{|C_j|} \right) \sigma^2 . \end{aligned} \tag{60}$$

因此，根据方程（29）、（59）–（60），以及模型 2.1（a）–（b）中的独立性假设，（58）中的随机向量服从高斯分布，具有零均值和各向同性协方差。具体而言，

$$x_i - \widehat{\mu}_j \sim \mathcal{N}\left(0,\ \left(2(1-R_j^\ell)^2\tau^2 + \left(1-\frac{1}{|C_j|}\right)\sigma^2\right) I_d\right). \tag{61}$$

利用方程（31），即可得到该引理。 □

下面的引理指出，在“最有利情形”下仅有一次误分配，此时 α_{cur} 的值最大。

引理 C.2 （“最有利情形”）。 考虑模型 *2.1*。固定 $i \in [n]$ 并令 $i \in C_j \cap S_\ell^\star$，回忆一下定义 R_j^ℓ中的*2.4*，以及方程 α_{cur} （10）中的 α_{cur} 在可行的最小纯度 $R_j^\ell = 1/|C_j|$处取最大值（等价地，$|C_j \cap S_\ell^\star| = 1$）。此时，

$$\|x_i - \widehat{\mu}_j\|^2 \sim \omega_{\text{cur}}\,\chi_d^2, \qquad \textit{where} \quad \omega_{\text{cur}} := \Big(1-\frac{1}{|C_j|}\Big)\sigma^2 + \frac{2\tau^2(|C_j|-1)^2}{|C_j|^2}, \tag{62}$$

对于所有可行的 R_j^ℓ，我们都有

$$\alpha_{\text{cur}} \le \omega_{\text{cur}} = \sigma^2\left(1-\frac{1}{|C_j|}\right) + \frac{2\tau^2\left(|C_j|-1\right)^2}{|C_j|^2}. \tag{63}$$

证明。 由于 $i \in C_j \cap S_\ell^\star$蕴含 $R_j^\ell \ge 1/|C_j|$，且 $r \mapsto (1-r)^2$在 $[0,1]$,上递减，因此我们得到（$1-R_j^\ell$）$^2 \le$（$1-1/|C_j|$）^{2}，从而 $\alpha_{\text{cur}} \le \omega_{\text{cur}}$。 代入 $R_j^\ell = 1/|C_j|$，即可得到 ω_{cur} 的表达式。 □

C.2.2. 到另一个簇中心的距离

首先，我们考虑样本到其不属于的簇的距离分布。我们重述引理 3.2并给出证明。

引理 3.2 （到另一个簇质心的距离）。 在引理 *3.1*的设定下，样本 x_i 与簇 C_j 的质心 $\widehat{\mu}_j$ 之间的距离服从如下分布

$$\begin{aligned} &\left\|x_i - \widehat{\mu}_{\bar{j}}\right\|^2 \sim \alpha_{\text{alt}}\chi_d^2\ , \\ &\alpha_{\text{alt}} = 2\tau^2\left(1-R_{\bar{j}}^\ell\right)^2 + \left(1+1/|C_{\bar{j}}|\right)\sigma^2. \end{aligned} \tag{11}$$

证明。 考虑模型 2.1中的设定，并固定真实簇分配 $\{z_i^\star\}$，如模型 2.1(c) 中所示，以及定义 z 2.2中的当前聚类分配。不失一般性，假设样本 i 当前被分配到簇 C_j，且属于真实类别 ℓ，即 $z_i^\star = \ell$，并考虑簇 C_j，该簇不包含样本 i，因此 $i \notin C_j$。

样本 x_i 到簇 C_j 的质心 $\widehat{\mu}_j$ 的距离为:

$$
\begin{aligned}
x_i - \widehat{\mu}_{\bar{j}} &= \mu^\star_\ell + \xi_i - \widehat{\mu}_{\bar{j}} && (64)\\
&= \mu^\star_\ell + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in C_{\bar{j}}} x_k && (65)\\
&= \mu^\star_\ell + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in C_{\bar{j}}} \left(\mu^\star_{z^\star_k} + \xi_k\right) && (66)\\
&= \mu^\star_\ell - \frac{1}{|C_{\bar{j}}|}\sum_{k\in C_{\bar{j}}} \mu^\star_{z^\star_k} + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in |C_{\bar{j}}|} \xi_k && (67)\\
&= \mu^\star_\ell - \frac{|C_{\bar{j}}\cap S^\star_\ell|}{|C_{\bar{j}}|}\mu^\star_\ell - \frac{|C_{\bar{j}}\cap S^\star_{\bar{\ell}}|}{|C_{\bar{j}}|}\mu^\star_{\bar{\ell}} + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in |C_{\bar{j}}|} \xi_k && (68)\\
&= \left(1 - \frac{|C_{\bar{j}}\cap S^\star_\ell|}{|C_{\bar{j}}|}\right)\mu^\star_\ell - \frac{|C_{\bar{j}}\cap S^\star_{\bar{\ell}}|}{|C_{\bar{j}}|}\mu^\star_{\bar{\ell}} + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in C_{\bar{j}}} \xi_k. && (69)
\end{aligned}
$$

使用 $R^\ell_j = |C_j \cap S^\star_\ell|/|C_j|$ 定义（2.4）中引入的记号，可得

$$
\begin{aligned}
x_i - \widehat{\mu}_{\bar{j}} &= (1 - R^\ell_{\bar{j}})\,\mu^\star_\ell - R^{\bar{\ell}}_{\bar{j}}\mu^\star_{\bar{\ell}} + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in C_{\bar{j}}} \xi_k\\
&= (1 - R^\ell_{\bar{j}})(\mu^\star_\ell - \mu^\star_{\bar{\ell}}) + \xi_i - \frac{1}{|C_{\bar{j}}|}\sum_{k\in C_{\bar{j}}} \xi_k. && (70)
\end{aligned}
$$

其中使用了 $R^\ell_j = 1 - R^{\bar{\ell}}_j$。

利用方程（29）、（59）以及模型2.1 （a）–（b）中所定义的随机变量的独立性，可知该差值服从如下分布

$$
x_i - \widehat{\mu}_{\bar{j}} \sim \mathcal{N}\left(0,\ \left(2(1 - R^\ell_{\bar{j}})^2\tau^2 + \left(1 + \frac{1}{|C_{\bar{j}}|}\right)\sigma^2\right) I_d\right), \tag{71}
$$

利用假设$i \notin C_j$。根据方程（31），可得结论。 □

引理 C.3 （另一个质心的“最有利情形”）。 考虑模型 *2.1* 的设定。固定 $i \in [n]$ 并假设$i \in C_j \cap S^\star_\ell$，回顾 R^ℓ_j 定义 *2.4* 中的 α_{alt} 以及方程（ 11 ）中的 $|C_j|$。对于固定的 α_{alt}，尺度 $R^\ell_j = 1$ 在可行的最大纯度 $C_j \subseteq S^\star_\ell$ 处达到最小值（等价地，$C_j \subseteq S^\star_\ell$）。在此情况下，

$$
\|x_i - \widehat{\mu}_{\bar{j}}\|^2 \sim \omega_{\text{alt}}\,\chi^2_d, \qquad \textit{where} \quad \omega_{\text{alt}} := \left(1 + \frac{1}{|C_{\bar{j}}|}\right)\sigma^2, \tag{72}
$$

并且对于所有可行的 R^ℓ_j，我们有

$$
\alpha_{\text{alt}} \geq \omega_{\text{alt}} = \left(1 + \frac{1}{|C_{\bar{j}}|}\right)\sigma^2. \tag{73}
$$

证明。显然，我们有 $\tau^2\left(1 - R^\ell_j\right)^2 \geq 0$，因此

$$
\alpha_{\text{alt}} = 2\tau^2\left(1 - R^\ell_{\bar{j}}\right)^2 + \left(1 + \frac{1}{|C_{\bar{j}}|}\right)\sigma^2 \ \geq\ \left(1 + \frac{1}{|C_{\bar{j}}|}\right)\sigma^2 = \omega_{\text{alt}}. \tag{74}
$$

代入 $R^\ell_j = 1$ 可得 ω_{alt}的所述表达式，而所声称的分布可由 $\|x_i - \widehat{\mu}_j\|^2 \sim \alpha_{\text{alt}}\chi^2_d$（方程（11））得出。

□

C.3. 定理 3.4 的证明3.4

我们重述定理3.4 并给出证明。

定理 3.4 (Lloyd 算法：单个样本)。 我们考虑模型 *2.1* 中的情形，其中固定 $i \in [n]$ 当前簇C_j 和另一个簇 C_j 满足 $i \in C_j$。我们将簇大小记为 $c := |C_j|$ 和 $c := |C_j|$。

如果噪声水平 $\sigma > 0$ 满足

$$\sigma \;>\; \frac{\sqrt{2\bar{c}}\,\tau\,(c-1)}{\sqrt{c(c+\bar{c})}}\,, \tag{13}$$

那么，我们有

$$\mathbb{P}\Big(\big\|x_i - \widehat{\mu}_{\bar{j}}\big\|^2 < \big\|x_i - \widehat{\mu}_j\big\|^2\Big) \;\le\; \rho^{\,d/4}, \tag{14}$$

其中 ρ 定义为

$$\rho(\sigma,\tau,c,\bar{c}) = \frac{4\sigma^2(c-1)c^2\bar{c}(\bar{c}+1)\big(c\big(\sigma^2+2\tau^2\big)-2\tau^2\big)}{\big(-c(\sigma^2+4\tau^2)\bar{c}+c^2(\sigma^2+2(\sigma^2+\tau^2)\bar{c})+2\tau^2\bar{c}\big)^2}, \tag{15}$$

并满足 $0 \le \rho < 1$。

证明。 将 $b_1 = \alpha_{\text{alt}}$ （方程（11））和 $b_2 = \alpha_{\text{cur}}$ （方程（10））代入引理 3.3，会得到一个稍微复杂一些且依赖于其他参数的表达式。我们注意到，表达式 $\rho = 1 - \left(\frac{b_1-b_2}{b_1+b_2}\right)^2$

引理 3.3中的表达式会 b_1 减小且在 b_2 增大时增大，因为

$$\partial_{b_1}\left\{-\left(\frac{b_1-b_2}{b_1+b_2}\right)^2\right\} = -2\frac{b_1-b_2}{b_1+b_2}\frac{2b_2}{(b_1+b_2)^2} < 0, \tag{75}$$

以及

$$\partial_{b_2}\left\{-\left(\frac{b_1-b_2}{b_1+b_2}\right)^2\right\} = 4\frac{(b_1-b_2)b_1}{(b_1+b_2)^3} > 0. \tag{76}$$

由于方程（75）和（76）的单调性，我们可以将 b_1 和 b_2 替换为其“最有利情形”下的取值 $\omega_{\text{alt}} \le \alpha_{\text{alt}}$ 和 $\omega_{\text{cur}} \ge \alpha_{\text{cur}}$ （辅助引理中的方程（73）和（63）），从而得到一个比代入 更宽松所得的界 b_1 和 b_2要松的界。特别地，将 $b_1 = \omega_{\text{alt}}$ 和 $b_2 = \omega_{\text{cur}}$ 代入$\rho = 1 - \left(\frac{b_1-b_2}{b_1+b_2}\right)^2$ 可得

$$\mathbb{P}\big(\|x_i - \widehat{\mu}_{\bar{j}}\|^2 < \|x_i - \widehat{\mu}_j\|^2\big) \le \rho^{\,d/4}, \tag{77}$$

with

$$\rho = \frac{4\sigma^4(1+|C_{\bar{j}}|^{-1})(1-|C_j|^{-1}) + 8\tau^2\sigma^2(|C_j|-1)^2(1+|C_{\bar{j}}|^{-1})|C_j|^{-2}}{\Big(\sigma^2(1+|C_{\bar{j}}|^{-1}) + \sigma^2(1-|C_j|^{-1}) + 2\tau^2(|C_j|-1)^2|C_j|^{-2}\Big)^2}. \tag{78}$$

这正是方程 (15)。

引理3.3 要求 $b_1 > b_2 > 0$ （其中 $m = 0$）满足方程 (13) 给出的噪声条件，因为后者正是为此而设计的。具体而言，方程 (13) 对应于条件$\omega_{\text{alt}} > \omega_{\text{cur}}$，因此引理3.3 中的要求 $b_1 > b_2 > 0$ 成立。

□

C.4. 定理3.6的证明

现在我们重述定理 3.6 并给出证明。

定理 3.6 (q-近似平衡划分的统一界)。 考虑模型 *2.1* 的设定，其中固定$i \in [n]$，其当前簇索引为 j，另一个簇为 j。固定划分不平衡因子 $q > 1$，并假设划分$\mathcal{P} = \{C_1, C_2\}$ 是 q-近似平衡的（定义 *2.5*）。固定 $\tau = 1$，并固定 $\beta > 1$，使得

$$\sigma = \beta\frac{(\sqrt{n}\,q + n - 2)}{\sqrt{2}\sqrt{\sqrt{n}\,q + n}}. \tag{16}$$

那么，

$$\mathbb{P}\left(\left\|x_i-\widehat{\mu}_{\bar{j}}\right\|^2-\left\|x_i-\widehat{\mu}_j\right\|^2<0\right)\leq\rho_q^{d/4}, \tag{17}$$

其中

$$\rho_q=\frac{\sigma^2\left(\sqrt{n}q+n-2\right)\left(\sqrt{n}q+n\right)\left(\sqrt{n}q+n+2\right)\left(\sqrt{n}\left(\sigma^2+2\right)\left(\sqrt{n}+q\right)-4\right)}{\left(n\sigma^2\left(\sqrt{n}+q\right)^2+\left(\sqrt{n}q+n-2\right)^2\right)^2}\,. \tag{18}$$

证明。 该定理旨在重现定理 3.4 的一个版本，该版本统一适用于所有 q-近似平衡划分。我们的思路是推导方程 (16) 中的噪声阈值（其中 $\beta=1$），该阈值对于 13) 中定理 3.4 所述的簇大小而言，大于方程 (q-近似平衡划分中的噪声阈值；类似地，推导方程 ρ_q (18) 中的 ρ，它大于定理 15) 中的 3.4，适用于 q-近似平衡划分中的簇大小。为此，我们将对定理3.4 的证明及其组成部分稍作修改后予以重复。

我们回顾定理 3.4的证明可知， ρ 是将 $b_1=\omega_{\text{alt}}$ （方程 (73)）和 $b_2=\omega_{\text{cur}}$ （方程(63)）代入引理 3.3后得到的，从而得出方程 (78。我们将重复这一推导，并将其推广到 q-近似平衡划分。

将 $|C_j|=n/2+q\sqrt{n}/4$ 和 $|C_j|=n/2+q\sqrt{n}/4$ 的最大值代入 $b_1=\omega_{\text{alt}}$ （方程（73））和 $b_2=\omega_{\text{cur}}$ （方程（63）），再代入引理 3.3 的方程（12），即可得到公式 (18)；接下来还需证明 ρ_q 为 ρ 的值提供了一个上界。为此，我们将证明，在范围 $n/2-q\sqrt{n}/4<|C_k|<n/2+q\sqrt{n}/4$ 内，对于 $k\in\{1,2\}$，方程（15）中的表达式关于 $|C_j|$ 和 $|C_j|$ 均为单调函数，即使不要求 $n=|C_j|+|C_j|$ 也成立。由此，我们可以用上界替换 $|C_j|$ 和 $|C_j|$。

接下来，我们需要证明，方程（16）中噪声的选取是有效的，且不依赖于任何簇大小。尤其是，由于噪声在确立 ρ的单调性时起着作用，因此重要的是证明，将 $|C_j|=n/2\pm q\sqrt{n}/4$ 和 $|C_j|=n/2+q\sqrt{n}/4$ 的最大值代入 $b_1=\omega_{\text{alt}}$ （方程（73））和 $b_2=\omega_{\text{cur}}$（方程（63））后，单调性结论仍然成立。

为此，我们将证明，方程（13）中 σ 的依赖簇大小的下界，在相关定义域内关于 $|C_j|$和 $|C_j|$ 均为单调函数。因此，可以求出其最大值，并与以下数值进行比较

$$\beta\frac{(\sqrt{n}\,q+n-2)}{\sqrt{2}\sqrt{\sqrt{n}\,q+n}} \tag{79}$$

我们在定理3.6的陈述中固定的值。

ρ的单调性。我们将计算 ρ 关于 $|C_j|$ 和 $|C_j|$的导数。为简化推导，我们利用链式法则，并以 ρ 是将 $b_1=\omega_{\text{alt}}$ （方程（73））和 $b_2=\omega_{\text{cur}}$ （方程（63））代入引理 3.3所得这一事实为基础。这样可以使表达式更简洁、清晰。回顾定理 3.4 的证明，对于 $i\in\{1,2\}$，

$$\partial_{b_i}\rho(b_1,b_2)=\partial_{b_i}\left(1-\frac{(b_1-b_2)^2}{(b_1+b_2)^2}\right)=\frac{4(b_{\bar{i}}-b_i)b_{\bar{i}}}{(b_1+b_2)^3}, \tag{80}$$

与 $i=2$，若 $i=1$，反之亦然。此外，利用 ω_{alt} 和 ω_{cur}的精确形式，

$$\partial_{|C_{\bar{j}}|}\omega_{\text{alt}}(|C_{\bar{j}}|)=\partial_{|C_{\bar{j}}|}\,(1+|C_{\bar{j}}|^{-1})\sigma^2=\frac{-\sigma^2}{|C_{\bar{j}}|^{-2}} \tag{81}$$

and

$$\partial_{|C_{\bar{j}}|}\omega_{\text{cur}}(|C_{\bar{j}}|)=\partial_{|C_j|}\,\left\{\sigma^2(1-|C_j|^{-1})+2\tau^2(|C_j|-1)^2|C_j|^{-2}\right\}=|C_j|^{-2}(\sigma^2+4\tau^2-4\tau^2|C_j|^{-1}). \tag{82}$$

这些就是应用链式法则所需的全部部分，

$$\partial_{|C_{\bar{j}}|}\rho\;=\;\partial_{\omega_{\text{alt}}}\rho(\omega_{\text{cur}},\omega_{\text{alt}})\;\partial_{|C_{\bar{j}}|}\omega_{\text{alt}}(|C_{\bar{j}}|). \tag{83}$$

由于 $\partial_{|C_{\bar{j}}|}\omega_{\text{alt}}(|C_{\bar{j}}|)=-\sigma^2|C_{\bar{j}}|^2$ 为负，要证明 ρ 随 $|C_{\bar{j}}|$增加，需满足 $\partial_{\omega_{\text{alt}}}\rho(\omega_{\text{cur}},\omega_{\text{alt}})$ 也为负。根据（80），这等价于条件

$$\sigma^2(1-|C_j|^{-1})+2\tau^2(|C_j|-1)^2|C_j|^{-2}<\sigma^2(1+|C_{\bar{j}}|^{-1}). \tag{84}$$

简单整理可得，

$$\frac{2\tau^2(|C_j|-1)^2|C_j|^{-2}}{|C_{\bar{j}}|^{-1}+|C_j|^{-1}} < \sigma^2. \tag{85}$$

请注意，这是先前对方差所述的条件。然而，在当前定理的背景下，我们希望这一界限对任意 q-近似平衡划分都成立。因此，我们要根据 $|C_j|$ 和 $|C_{\bar{j}}|$ 找出最坏情形，并证明定理陈述中 σ 的取值足以满足要求。

现在来证明 ρ 随 $|C_j|$ 增大。此时的链式法则为

$$\partial_{|C_j|}\rho \;=\; \partial_{\omega_{\mathrm{cur}}}\rho(\omega_{\mathrm{cur}},\omega_{\mathrm{alt}})\;\partial_{|C_j|}\omega_{\mathrm{cur}}(|C_j|). \tag{86}$$

注意到 $\partial_{|C_j|}\omega_{\mathrm{cur}}(|C_j|) = |C_j|^{-2}(\sigma^2+4\tau^2-4\tau^2|C_j|^{-1})$ 为正，因为 $1-|C_j|^{-1}>0$。由于方程 (80) 的分子中，当 $\partial_{\omega_{\mathrm{cur}}}\rho(\omega_{\mathrm{cur}},\omega_{\mathrm{alt}})$ 时出现的差，与 $\partial_{\omega_{\mathrm{alt}}}\rho(\omega_{\mathrm{cur}},\omega_{\mathrm{alt}})$时的差互为相反数，因此，根据方程 (85) 中的条件，同样可得$\partial_{\omega_{\mathrm{cur}}}\rho(\omega_{\mathrm{cur}},\omega_{\mathrm{alt}}) > 0$。由此即可得到所需的单调性。

单调性： σ的下界。由（85）可见， σ 的下界作为 $|C_{\bar{j}}|$的函数而增大，因此我们转向另一个自变量。计算

$$\partial_{|C_j|}\,\frac{(|C_j|-1)^2|C_j|^{-2}}{|C_{\bar{j}}|^{-1}+|C_j|^{-1}} = \frac{\big(2(|C_j|-1)|C_j|^{-2}-2(|C_j|-1)^2|C_j|^{-3}\big)\big(|C_{\bar{j}}|^{-1}+|C_j|^{-1}\big)+(|C_j|-1)^2|C_j|^{-4}}{(|C_{\bar{j}}|^{-1}+|C_j|^{-1})^2}$$

并注意到

$$\begin{aligned}
&\big(2(|C_j|-1)|C_j|^{-2}-2(|C_j|-1)^2|C_j|^{-3}\big)\big(|C_{\bar{j}}|^{-1}+|C_j|^{-1}\big)+(|C_j|-1)^2|C_j|^{-4}\\
&=(|C_j|-1)|C_j|^{-2}\Big(2|C_{\bar{j}}|^{-1}+2|C_j|^{-1}-(|C_j|-1)|C_j|^{-1}\big(2|C_{\bar{j}}|^{-1}+2|C_j|^{-1}-|C_j|^{-1}\big)\Big)\\
&=(|C_j|-1)|C_j|^{-2}\Big(|C_j|^{-1}+(|C_j|-1)|C_j|^{-1}\big(2|C_{\bar{j}}|^{-1}+|C_j|^{-1}\big)\Big)>0.
\end{aligned}$$

因此，我们已证明：在近似平衡划分的假设下，回顾定义 2.5，将 $|C_j| = n/2+q\sqrt{n}/4$ 与 $|C_j| = n/2+q\sqrt{n}/4$ 的最大值代入方程（13）和（15），可得到一个上界，并进一步得到本定理中的表达式。

我们注意到，在 $|C_j|$ 与 $|C_j|$的范围内最大化 ρ 与 σ，而不是分别最大化 $|C_j|$ 与 $|C_j|$ 的值，可以得到略优的表达式。 □

注 C.4。 我们强调，我们的陈述在精确意义上排除了极端不平衡的划分：我们要求每个簇至少包含两个样本。因此，不平衡容差参数 q 可以随 n增长。在这种情况下，事实A.10给出

$$\mathbb{P}(|S-n/2|\ge q\sqrt{n}/2) \le 2\exp\left(-\frac{q^2}{2}\right) \tag{87}$$

这样就能确保所覆盖的划分比例（如 $n\to\infty$）趋近于 1。

为确保我们仍不会取 q 过大（以满足划分至少包含 2 个样本的要求），取 $q\le 0.6\sqrt{n}$ 便足够了（当 $n\to\infty$渐近趋于$\mathrm{O}(n^2)$。事实上，大小至多为 2 的划分数量的阶为$\mathrm{O}(n^2)$；因此它们所占的比例约为$\mathrm{O}(n^2 2^{-n})$。于是可以使用类似Mousavi(2010年)给出的反集中结果，断言

$$\mathbb{P}\left((S_n-n/2)>t\right) \ge \frac{1}{4}\exp(-2t^2/n). \tag{88}$$

接下来，只需将我们选取的 $q\le 0.6\sqrt{n}$ 代入上面的公式，即可看到问题划分渐近地得以正确避开。另请注意，$S_n-n/2$ 关于零对称。

C.5. 推论3.8的证明

我们重述推论 3.8 并给出证明。

推论 3.8（主要结果：Lloyd 算法）。 考虑模型*2.1*。固定不平衡参数 $q > 1$ 并假设噪声水平满足方程 (16)。则存在一个 q-近似平衡划分（见定义*2.5*）且不是 *Lloyd* 的 k-均值更新方案的不动点的概率上界为

$$\begin{aligned}&\mathbb{P}\left(\exists \textit{ approx. balanced partition that is not a fixed point}\right)\\&\quad\leq 2^n n\rho_q^{d/4}\,,\end{aligned} \tag{21}$$

其中 ρ_q 定义于(18)。

证明。首先，我们考虑满足推论条件的一个划分 $\mathcal{P}$。根据并集界（方程（38））， 我们有

$$\mathbb{P}\left(\exists i\in[n]:\left\|x_i-\widehat{\mu}_{\bar{j}}\right\|^2<\left\|x_i-\widehat{\mu}_j\right\|^2|\mathcal{P}\right)\leq\sum_{i=1}^n\mathbb{P}\left(\left\|x_i-\widehat{\mu}_{\bar{j}}\right\|^2<\left\|x_i-\widehat{\mu}_j\right\|^2|\mathcal{P}\right). \tag{89}$$

代入方程（17）， 得到

$$\mathbb{P}\left(\exists i\in[n]:\left\|x_i-\widehat{\mu}_{\bar{j}}\right\|^2<\left\|x_i-\widehat{\mu}_j\right\|^2\,|\mathcal{P}\right)\leq n\rho_q^{d/4}\,. \tag{90}$$

接下来，我们考 虑所有满足推论条件的划分。再次使用并集界，我们有 e

$$\mathbb{P}\left(\exists i\in[n],\mathcal{P}:\left\|x_i-\widehat{\mu}_{\bar{i}}\right\|^2<\left\|x_i-\widehat{\mu}_j\right\|^2\right)\leq\sum_{\mathcal{P}}\mathbb{P}\left(\exists i\in[n]:\left\|x_i-\widehat{\mu}_{\bar{j}}\right\|^2<\left\|x_i-\widehat{\mu}_j\right\|^2\,|\mathcal{P}\right)\leq\sum_{\mathcal{P}}n\rho_q^{d/4}\,. \tag{91}$$

回顾事实 A.9，样本可能的划分至多有 2^n 种，因此满足定理要求的划分至多有 2^n 种，由此得到所需结果。

□

D. Hartigan 算法证明

本节包含 正文中与 Hartigan 相关的定理、引理和推论的证明 's 算法相关的 定理、引理和推论的证明，以及一些额外的辅助结果和说明。

D.1. 辅助引理：Hartigan 加权距离的分布

将 Hartigan 加权距离应用于方程（9）中引理 3.1和引理 3.2 给出的距离，可立即得到以下引理。

推论 D.1（Hartigan 加权距离至不同簇）。 在引理 *3.2*的设定下，考虑一个来自真实类别 $x_i \in S_\ell^\star$的样本 ℓ，并设 C_j 为 x_i 当前未被分配到的簇（$i \notin C_j$）。那么，方程（9）中定义的 *Hartigan* 加权距离服从

$$\Delta_H^2(x_i,C_{\bar{j}})=\frac{\|x_i-\hat{\mu}_{\bar{j}}\|^2}{1+1/|C_{\bar{j}}|}\sim\eta_{C_{\bar{j}}}\,\chi_d^2, \tag{92}$$

其尺度参数为

$$\eta_{C_{\bar{j}}}=2\tau^2\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2+\sigma^2. \tag{93}$$

随机化 涉及噪声和真实类别中心，如模型 *2.1(a)–(b)*所规定。 .

推论 D.2 (Hartigan 到当前簇的加权距离)。 在引理 *3.1*的设定下，考虑样本 $x_i \in S_\ell^\star$与其当前所属簇 C_j 的质心 之间的$(i \in C_j)$。根据方程 (9) 定义的 *Hartigan* 加权距离服从如下分布

$$\Delta_H^2(x_i,C_j)=\frac{|C_j|}{|C_j|-1}\,\|x_i-\widehat{\mu}_j\|^2\sim\eta_{C_j}\,\chi_d^2, \tag{94}$$

其尺度参数为

$$\eta_{C_j}=2\tau^2\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2+\sigma^2. \tag{95}$$

随机化 n 再次表示噪声和真实类别中心，具体定义见模型*2.1(a)–(b)* .

观察D.3。 根据推论D.2 和推论D.1，Hartigan 加权平方距离的期望值为

$$\mathbb{E}\Delta_H^2(x_i, C_j) = \begin{cases} d\left(2\tau^2 \frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 + \sigma^2\right) & \text{if } i \in C_j \\ d\left(2\tau^2 \frac{|C_j|}{|C_j|+1}(1-R_j^\ell)^2 + \sigma^2\right) & \text{otherwise .} \end{cases} \tag{96}$$

有趣的是，当前簇的加权平方距离与另一簇的加权平方距离之间的期望差异并不取决于噪声水平 σ。事实上，

$$\begin{aligned} &\mathbb{E}\left(\Delta_H^2(x_i, C_j) - \Delta_H^2(x_i, C_{\bar{j}})\right) = \mathbb{E}\left(\Delta_H^2(x_i, C_j)\right) - \mathbb{E}\left(\Delta_H^2(x_i, C_{\bar{j}})\right) \\ &= 2\tau^2 d\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 - \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right) . \end{aligned} \tag{97}$$

这一性质不适用于未加权距离或其他看似自然的权重。回想一下， χ_d^2会随着 d 增大而集中在均值附近，仔细考察方程（97）可知，如果另一簇中来自该样本真实类别的样本比例更高，那么 Hartigan 算法往往会将该样本从当前簇重新分配到另一簇。证明的后续步骤将使这一观察更加严谨。

D.2. 定理3.9的证明

我们重述定理3.9 并给出证明。

定理 3.9 (Hartigan 算法：单个样本)。 在模型 *2.1*的设定下，固定一个索引 $i \in [n]$，并令 $\ell \in \{1,2\}$ 为其真实类别（因此 $i \in S_\ell^\star$）。令 $j \in \{1,2\}$ 为其当前簇索引（因此 $i \in C_j$），并令 j 表示另一簇。若簇纯度系数（定义*2.4*）满足

$$0 < R_j^\ell \le R_{\bar{j}}^\ell \le 1, \tag{22}$$

那么，

$$\mathbb{P}\Big(\Delta_H^2(x_i, C_j) \le \Delta_H^2(x_i, C_{\bar{j}})\Big) \;\le\; \rho^{\,d/4}, \tag{23}$$

其中 Δ_H 定义于方程(9)，且 ρ 由下式给出

$$\rho = 1 - \left(\frac{\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 - \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right)}{\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 + \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right) + \sigma^2}\right)^2 . \tag{24}$$

并满足 $0 \le \rho < 1$。

证明。 根据推论 D.2 和推论D.1，

$$\Delta_H^2(x_i, C_j) \;\sim\; b_1\,\chi_d^2, \qquad \Delta_H^2(x_i, C_{\bar{j}}) \;\sim\; b_2\,\chi_d^2, \tag{98}$$

with

$$b_1 \;=\; 2\tau^2\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 + \sigma^2, \qquad b_2 \;=\; 2\tau^2\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2 + \sigma^2, \tag{99}$$

由于 $\frac{|C_j|}{|C_j|-1} > 1$ 和 $\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1} < 1$，并根据方程（22）中的假设，我们可以得出 $b_1 - b_2 > 0$。事实上，

$$\begin{aligned} b_1 - b_2 &= 2\tau^2\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2 - 2\tau^2\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2 \\ &> 2\tau^2\Big[(1-R_j^\ell)^2 - (1-R_{\bar{j}}^\ell)^2\Big] \\ &= 2\tau^2\Big[\big(2-R_j^\ell-R_{\bar{j}}^\ell\big)\big(R_{\bar{j}}^\ell - R_j^\ell\big)\Big] \ge 0. \end{aligned}$$

因此，这些分布满足引理3.3的要求。将（99）代入引理 3.3 得到 ds

$$\mathbb{P}\left(\Delta_H^2(x_i, C_j) - \Delta_H^2(x_i, C_{\bar{j}}) < 0\right) \leq \rho^{d/4} \tag{100}$$

其中

$$\rho = 1 - \left(\frac{b_1 - b_2}{b_1 + b_2}\right)^2 = 1 - \left(\frac{\tau^2\left[\frac{|C_j|}{|C_j|-1}(1 - R_j^\ell)^2 - \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1 - R_{\bar{j}}^\ell)^2\right]}{\tau^2[\frac{|C_j|}{|C_j|-1}(1 - R_j^\ell)^2 + \frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1 - R_{\bar{j}}^\ell)^2] + \sigma^2}\right)^2 \tag{101}$$

且 $\rho < 1$。 □

定理3.9 的一个直接推论如下。

推论 3.10。 如果满足定理 *3.9* 的条件，那么该划分成为 $\mathcal{P}$ *Hartigan* 算法不动点的概率有如下界：

$$\mathbb{P}(\mathcal{P} \textit{ is a fixed point}) \;\leq\; \rho^{d/4}, \tag{25}$$

其中 ρ 由方程(24)给出。

D.3. 推论3.11的证明

我们重新陈述推论3.11 并给出证明。

推论3.11。 考虑模型*2.1*的设定。进一步假设 $n \geq 4$。如果当前划分 $\mathcal{P}$ 非空且是错误划分（定义*2.4*），那么， $\mathcal{P}$ 是不动点的概率上界为

$$\mathbb{P}\left(\mathcal{P} \textit{ is a fixed point}\right) \leq \rho_h^{d/4}, \tag{26}$$

其中 ρ_h 由下式给出

$$\rho_h = 1 - \left(\frac{4\tau^2(R^\star)^2 n^{-1}}{3\tau^2 + \sigma^2}\right)^2 < 1\,, \tag{27}$$

且 $R^\star = \min(R^1, R^2)$ 是最小真实类别的相对大小。

证明。 我们回顾定义 2.4中的比例记号，

$$R_j^\ell = \frac{|S_\ell^\star \cap C_j|}{|C_j|}, \quad R^\ell = \frac{|S_\ell^\star|}{n}\,. \tag{102}$$

步骤 1：选择 C_j 和 $S_\ell^\star$。 由于 $\mathcal{P}$ 不等于真实划分，因此至少有一个簇不是纯的。因此，存在 $j \in \{1, 2\}$ 使得 $0 < R_j^1 < 1$ 且 $0 < R_j^2 < 1$。固定这样一个索引 j。此外，由于 $R_j^1 + R_j^2 = 1$ 且$R^1 + R^2 = 1$， $R_j^1 > R^1$ 和 $R_j^2 > R^2$ 不可能同时成立。因此，对于这个（混合）簇 C_j，存在一个索引 $\ell \in \{1, 2\}$ 使得

$$R_j^\ell \leq R^\ell. \tag{103}$$

此后我们固定这样一个 ℓ，并相应地取一个样本 $i \in C_j \cap S_\ell^\star$，满足 $C_j \cap S_\ell^\star \neq \emptyset$（这是成立的，因为 C_j 包含来自两个类别的点）。

步骤 2：的表达式 R_j^ℓ。为方便起见，令 $n_j := |C_j|$ 和 $n_{\bar{j}} := |C_{\bar{j}}|$，于是 $n = n_j + n_{\bar{j}}$。

根据定义2.4，我们有 $|S_\ell^\star| = R^\ell n$ 和 $|C_j \cap S_\ell^\star| = R_j^\ell n_j$。由于

$$|C_{\bar{j}} \cap S_\ell^\star| = |S_\ell^\star| - |C_j \cap S_\ell^\star| = R^\ell n - R_j^\ell n_j, \tag{104}$$

由此可得

$$R_{\bar{j}}^\ell = \frac{|C_{\bar{j}} \cap S_\ell^\star|}{n_{\bar{j}}} = \frac{R^\ell n - R_j^\ell n_j}{n - n_j}. \tag{105}$$

注意，由方程（105），方程（103）的选择意味着 $R_j^\ell \geq R^\ell$（因为 $n - n_j > 0$）。这对不等式，

$$R_j^\ell \leq R^\ell \leq R_{\bar{j}}^\ell, \tag{106}$$

将在下文中使用。

步骤 3：推导 ρ。样本 i 满足定理3.9 和推论3.10，尤其是由于方程（22）而满足方程（106）。因此，划分为不动点的概率也满足方程（25），其中

$$\begin{aligned}\rho &= 1-\left(\frac{\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2-\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right)}{\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2+\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right)+\sigma^2}\right)^2 \\ &= 1-\left(\frac{\tau^2\left(\frac{n_j}{n_j-1}(1-R_j^\ell)^2-\frac{n_{\bar{j}}}{n_{\bar{j}}+1}(1-\frac{R^\ell n-R_j^\ell n_j}{n-n_j})^2\right)}{\tau^2\left(\frac{n_j}{n_j-1}(1-R_j^\ell)^2+\frac{n_{\bar{j}}}{n_{\bar{j}}+1}(1-\frac{R^\ell n-R_j^\ell n_j}{n-n_j})^2\right)+\sigma^2}\right)^2 \\ &= 1-\left(\frac{\tau^2\left[\left(\frac{n_j}{n_j-1}\right)(1-R_j^\ell)^2-\frac{n-n_j}{n-n_j+1}\left(1-\frac{R^\ell n-R_j^\ell n_j}{n-n_j}\right)^2\right]}{\tau^2\left[\frac{n_j}{n_j-1}(1-R_j^\ell)^2+\frac{n-n_j}{n-n_j+1}\left(1-\frac{R^\ell n-R_j^\ell n_j}{n-n_j}\right)^2\right]+\sigma^2}\right)^2.\end{aligned} \tag{107}$$

S第 4 步：的上界 ρ。 首先，我们注意到方程（107）中的分母满足 s

$$\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2+\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right)+\sigma^2\le 3\tau^2+\sigma^2\,. \tag{108}$$

接下来，我们考虑（107）中的分子。

然后，一起 结合 $R_j^\ell\le R^\ell\le R_j^\ell$ 在方程（106）中的事实，我们得到分子的下界 or

$$\begin{aligned}\tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R_j^\ell)^2-\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R_{\bar{j}}^\ell)^2\right) &\ge \tau^2\left(\frac{|C_j|}{|C_j|-1}(1-R^\ell)^2-\frac{|C_{\bar{j}}|}{|C_{\bar{j}}|+1}(1-R^\ell)^2\right) \\ &= \tau^2\left(\frac{n_j}{n_j-1}(1-R^\ell)^2-\frac{n-n_j}{n-n_j+1}(1-R^\ell)^2\right) \\ &= \tau^2(1-R^\ell)^2\,\frac{n}{(n_j-1)(n-n_j+1)} \\ &\ge \tau^2(1-R^\ell)^2\,\frac{4n}{n^2}.\end{aligned} \tag{109}$$

为了得到上方显示式中的最后一个不等式，我们使用了如下界

$$(n_j-1)(n-n_j+1)\le n^2/4. \tag{110}$$

事实上，对于任意非平凡划分 $1\le n_j\le n-1$，我们令 $a:=n_j-1\in[0,n-2]$ 并注意到 $(n_j-1)(n-n_j+1)$ $=a(n-a)$。由于 $f(a):=a(n-a)$ 是凹函数，其在 $[0,n-2]$ 上的最大值在 $n/2\le n-2$ 时取于 $a=n/2$，或取于端点 $a=n-2$。根据我们的假设（回顾推论的陈述）， $n\ge 4$，因此最大值取于 $n/2$，从而 $(n_j-1)(n-n_j+1)$ $\le n/2\,(n-n/2)$。

步骤 5： 的一致上界。本证明中定义的类别 ℓ 可能因划分不同而异。为了得到适用于所有相关划分的界，我们将 $1-R^\ell$ 替换为 $R^\star:=\min(R^1,R^2)$，从而得到

$$\rho\le 1-\left(\frac{4\tau^2(R^\star)^2n^{-1}}{3\tau^2+\sigma^2}\right)^2=:\rho_h.$$

回顾一下，对于任意非平凡的类别分配， $0<R^\star<1$，由此完成证明。 □

D.4. 推论的证明3.12

我们重新陈述推论 3.12 并给出证明。

推论 3.12 (主要结果：Hartigan 算法)。 考虑模型2.1的设定。假设 $n \geq 4$。那么，存在任何非空错误划分（定义2.4）成为 *Hartigan* 算法不动点的概率至多为

$$\mathbb{P}\left(\exists \mathcal{P} \text{ a non-empty incorrect partition}\right) \leq 2^n \rho_h^{d/4}, \tag{28}$$

其中 ρ_h 由下式给出(27)。

证明。 我们将所有非正确划分且非空的划分集合记为 $\mathcal{Q}$。

根据并集　方程（38）中的界，至少有一个划分不是不动点的概率为

$$\mathbb{P}\left(\exists \text{ non-fixed point in } \mathcal{Q}\right) = \mathbb{P}\Big(\cup_{\mathcal{P}\in\mathcal{Q}} \left(\mathcal{P} \text{ is not a fixed point}\right)\Big) \tag{111}$$

$$\leq \sum_{\mathcal{P}\in\mathcal{Q}} \mathbb{P}\left(\mathcal{P} \text{ is not a fixed point}\right) \tag{112}$$

非正确且非空的划分数量小于所有可能划分的数量 2^n （忽略对称性），由此完成证明。

□

E. 其他实现细节

本节将介绍有关实现和基准测试的其他细节。代码使用 Python 编写，其中 Lloyd 的 k-均值使用 JAX（Bradbury 等人，2018年），Hartigan 的 k-均值使用 Numba（Lam 等人，2015年），谱聚类使用 Python 的 scikit-learn（Pedregosa 等人，2011年），半定规划松弛使用 CVXPY（Diamond 和 Boyd，2016年）。所有实验均在配备 8GB 内存的 Quadro RTX 400 GPU 上运行。

E.1. 聚类算法

本节将进一步介绍我们用于与 Lloyd 的和 Hartigan 的 k-均值算法进行比较的主要基准算法。

"PCA + Lloyd"。这是一种适用于高维数据中 k-均值的常用启发式方法：在执行 k-均值之前，先将 PCA 应用于数据作为预处理步骤（Ding 和 He，2004年）。需要注意的是，在展示涉及 k-均值损失的结果时，我们使用该算法生成的划分来计算原始 数据的损失，而不是降维数据的损失。

半定规划松弛 k-均值问题。我们实现了（Mixon 等人，2016年）中提出的半定规划，它是最先进的一系列半定规划之一。该系列中的一些算法已有正式的最优性结果（Peng 和 Wei，2007年）。需要指出的是，半定规划出了名地难以扩展。

谱聚类。这是一种广受欢迎的聚类算法，它利用数据相似度矩阵的谱分解（Shi 和 Malik，2000年；Ng 等人，2001年）。需要指出的是，有些谱聚类方法并不旨在求解 k-均值最小化问题，如方程（5）中所述。扩展谱聚类同样具有挑战性。在实验中，我们使用 Python 的 scikit-learn 库中实现的谱聚类算法（Pedregosa 等人，2011年），采用默认参数，并在聚类步骤中使用 Lloyd 的 k-均值算法。

E.2. Olivetti人脸

我们使用奥利维蒂人脸数据集（Samaria & Harter，1994年），该数据集可通过 Python 的 scikit-learn 库获取（Pedregosa 等人，2011年），并将每张图像归一化，使其欧几里得范数为 1。

E.3. 20 Newsgroups数据集

这里，我们介绍对 20 Newsgroups数据集（Mitchell，1997年）所做的预处理步骤，以创建第 4 节中用于聚类算法基准测试的三个数据集。首先，我们从 20 Newsgroups 数据中移除了邮件头、邮件尾和引文。随后，我们提取了每个数据集中特定类别所对应的文档（即样本）：

- **20NG-A**：“alt.atheism”和“comp.graphics”。
- **20NG-B**：“alt.atheism”、“comp.graphics”、“comp.windows.x”、“misc.forsale”和“rec.autos”。
- **20NG-C**：“alt.atheism”、“comp.graphics”、“comp.windows.x”、“misc.forsale”、“rec.autos”、“rec.motorcycles”、“rec.sport.baseball”、“sci.crypt”、“sci.med”和“sci.space”。

接下来，我们按照 Python 的 scikit-learn（Pedregosa 等人，2011年）中的实现，使用包含 5000 个特征的标准 TF-IDF 对每个样本进行分词。最后，我们针对每个数据集以相同的概率进行无放回随机抽样，抽取 n 个样本：20NG-A 为 200 个，20NG-B 为 500 个，20NG-C 为 1000 个。

F. 更多数值结果

本节将介绍第 4 节所述结果的补充内容。首先，我们介绍 k均值胜率，这是一种基于 k均值损失（方程（5））比较划分的指标。接下来，我们补充比较第 4 节中进行的实验。随后，我们评估 Lloyd 的 k均值算法与一种基于主成分分析的简单算法。然后，我们对 Lloyd 的 k均值算法和 Hartigan 的 k均值算法的计算时间进行基准测试。最后，我们针对定理3.4 和3.9进行数值实验。

F.1. k-均值胜率：通过 k-均值损失比较划分

在以下各节中，我们使用一种不同于归一化互信息的替代指标来比较聚类划分，该指标基于 k-均值损失（方程（5））， 我们将其称为 k-均值胜率。由于原始损失难以在不同参数设置（例如数据维度或噪声方差）之间进行解读，我们定义了一种评分，用于评估各算法相对于真实划分，在 k-均值损失方面得到的划分是优于还是劣于真实划分。如果算法输出划分的损失与真实划分的损失接近（相对差异不超过 10^{-6}），则得分为零；如果输出划分优于真实划分（损失更低），则得分为一；如果真实划分更优，则得分为负一。针对每种参数设置运行多次实验后，我们对所有生成实例的得分取平均，以获得整体性能指标。

F.2. 合成高斯混合模型实验的其他结果

在本节中，我们重新审视第 4.1 节中进行的实验，并根据基于k-均值胜率的指标（见第 F.1节）呈现结果。需要注意的是，当采用降维（“PCA + k-均值”）时，我们使用该算法生成的划分来计算原始数据的损失，而不是降维数据的损失。此外，我们还给出了图 1的补充结果，其中比较了 Lloyd 和 Hartigan k-均值在不同初始化策略下的 NMI。

图 2 展示了各聚类算法——Lloyd 和 Hartigan 的 k-均值，以及第 E节所述算法——相对于真实划分的平均得分，指标为 k-均值胜率。这些结果与图 1 所示的 NMI 得分一致，并表明在高维情况下，Lloyd 的 k-均值算法无法最小化方程（5），并会收敛到次优不动点。

图4 与图3 分别比较了 Lloyd 与 Hartigan 的 k-均值算法在每种初始化策略下的 k-均值胜率与 NMI。我们观察到，Hartigan 的性能在不同初始化方法之间相对一致，而 Lloyd 的性能在使用“随机划分”初始化时显著下降。我们将 Lloyd 的 k-均值结果中的这一差异归因于：当使用随机中心或 k-均值++ 进行初始化时，初始中心有可能对应于不同的真实中心，且在较低的 K取值下概率更高。在这种情况下，于 Lloyd 算法的单次迭代中将样本分配到最近簇可能是有效的。然而，当 K 较大时，从不同真实簇中获得中心的可能性会降低。当使用随机平衡划分进行初始化时，初始中心更有可能是样本的均衡混合，从而导致初始划分成为 Lloyd 的 k-均值的不动点，如定理3.6所述。下一幅图将对此作进一步说明。

图5 展示了对于每个 k和 d 的取值，Lloyd 的均值 σ^2-means 达到收敛所需的平均迭代次数。图中显示，在高维情况下，Lloyd 的均值 k-means 平均只需一次迭代即可收敛。这与我们的理论一致：Lloyd 算法在第一次迭代后便会终止，因为下一次迭代无法更新划分。我们

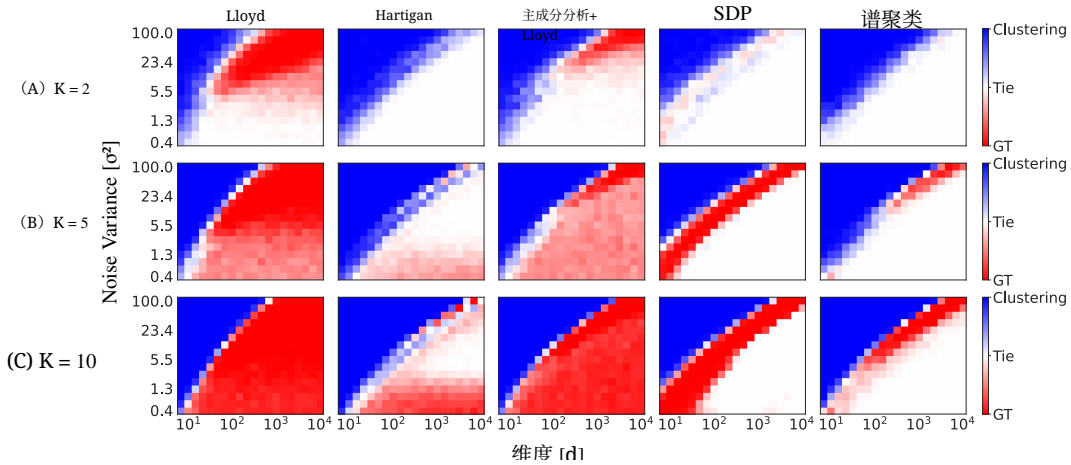


图*2*。 比较 k-均值胜率（见第 F.1节），针对不同 K 值，在合成高斯混合模型数据集上使用各算法得到的结果。每个值对应 100 次独立实验的平均值；对于每个实例，我们从模型 2.1 （推广至 $K \geq 2$）中定义的高斯混合模型采样数据，其中 $\tau^2 = 1$，每类有 20 个样本，并运行各算法直至收敛。

观察到 使用随机划分进行初始化时，这一现象更加明显 s.

图 6比较了第 4 节数值实验中，Lloyd 的均值 k-means、Hartigan 的 k-means、半定规划和谱聚类所获得的 NMI。在低噪声情形下，谱聚类的 NMI 通常略高于其他方法；与 Hartigan 的 0.1 -means 相比，差距很少超过 k。对于 $K = 2$ 和 $K = 5$，谱聚类在最高噪声水平下似乎也更加稳健；不过，随着 K 增大，这一优势会减弱，而 Hartigan 方法则会变得具有竞争力，并在 $K = 10$的情况下表现更优。我们强调，谱聚类通常可扩展性较差，而且根据图构造和归一化方式的不同，它未必优化与 Hartigan 算法相同的 k-means 目标函数。

F.3. Lloyd 的 k-均值算法在“简单问题”上失效

为了说明在某些情况下聚类问题会变得“简单”，我们介绍一种基于主成分分析（PCA）的简化聚类算法，称为“PCA + Split”。该算法专为本文关注的两个簇这一简单情形而设计。在“PCA + Split”算法中，我们计算数据的第一个主成分，并根据样本的主成分系数的符号对样本进行划分。图7 比较了 Lloyd 的 k-均值算法、“PCA + Split”以及真实划分所得到的划分。图 7 (A) 通过 k-均值胜率（见第 F.1节）比较这些划分；图 7 (B) 则通过归一化互信息（见定义 A.13）进行比较。

F.4. Lloyd 与 Hartigan 的 k-均值算法的计算性能比较

虽然计时分析并非本文的主要研究范围，但我们提供了有限的比较，以说明 Hartigan 算法在执行时间方面与 Lloyd 算法表现相近。我们没有将谱聚类和半定规划算法纳入此项分析，因为对于本文所考虑的问题尺度而言，这些算法通常并不实用。为了进行比较，我们使用 Numba（Lam 等人，2015）实现了这两种算法，并在算法 1 和算法2所示的简化伪代码基础上，加入了较为标准的改进。所有实验均在单个 CPU 线程上进行，未使用 GPU 加速。我们在不同维度 (d)、样本量 (n) 和簇数 (K) 下评估了这些算法。数据是从模型2.1 中指定的高斯混合模型（GMM）采样得到的（推广至 $K \geq 2$），其中 $\tau^2 = 1$ 且 $\sigma^2 = 10$。在这种情况下，Lloyd 算法通常经过几次迭代便会终止。每次实验都会生成一个新的数据集，并将随机样本作为中心，分别初始化每种聚类算法，然后各运行一次。对于每组10 的组合，实验均重复 d、 n和 K 取值。图8 （A）和图8（B）分别展示了各算法运行一次迭代以及运行至收敛所需的计算时间。我们观察到，在这一实现中，Lloyd 算法和 Hartigan 算法的计算成本相近；当 n 取值较高时，由于收敛速度更快，Hartigan 方法表现更佳。值得注意的是

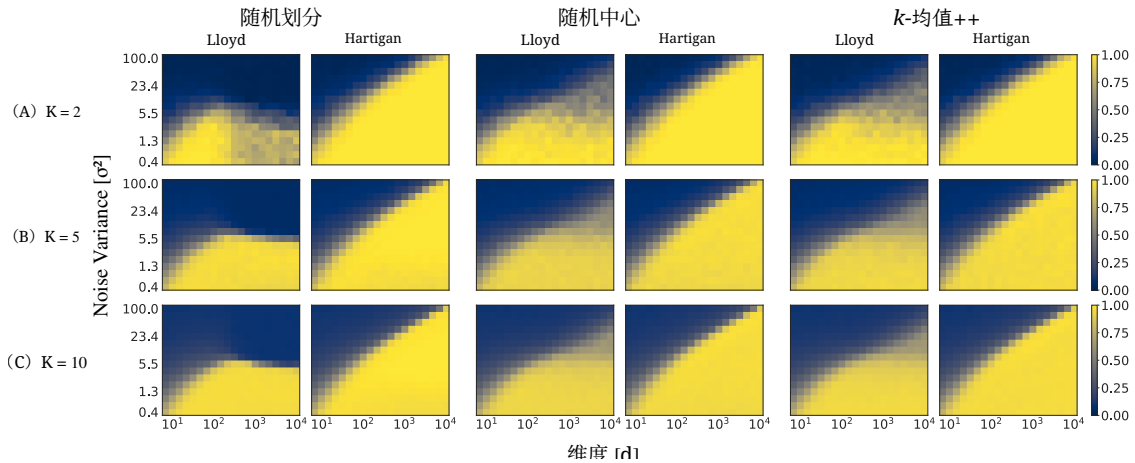


图*3.* Lloyd 和 Hartigan 的k-均值所得簇与真实簇之间的归一化互信息（NMI）。NMI 值为 1 表示完全相关，值为 0 表示两种划分之间没有互信息。每个数值均为 100 次独立实验的平均值；每次实验中，我们从模型 2.1 （推广至 $K \geq 2$）所定义的 $\tau^2 = 1$ 高斯混合模型中抽取 40 个样本，各簇大小相等。该图表明，随着数据维度增加，Lloyd 的 k-均值对初始中心更加敏感，而 Hartigan 的 k-均值则不那么敏感。各初始化策略的详细信息见第 4 节。

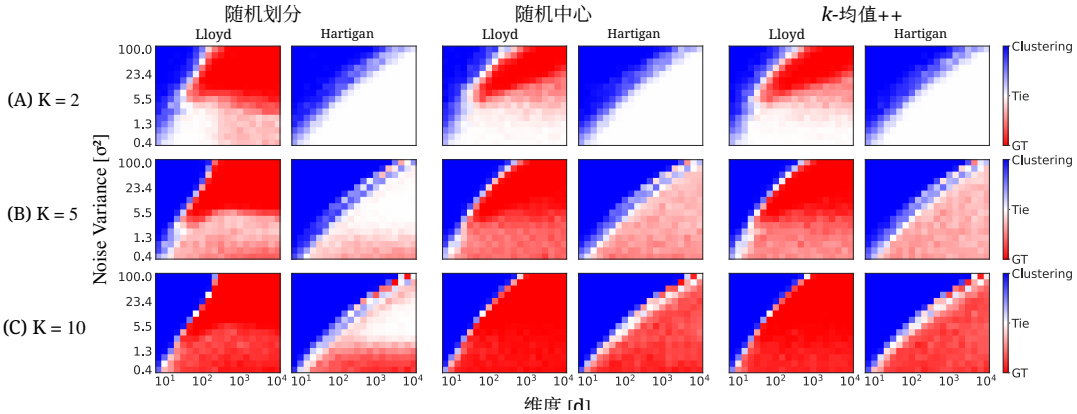


图*4.* k-均值获胜率（见第 F.1 节）比较了 Lloyd 和 Hartigan 的 k-均值在不同初始化策略和类别数 K下的表现。每个数值均为 100 次独立实验的平均值；每次实验中，我们从模型 2.1 （推广至 $K \geq 2$）所定义的高斯混合模型中抽取数据，每类包含 $\tau^2 = 1$ 20 个样本，并运行每种算法直至收敛。

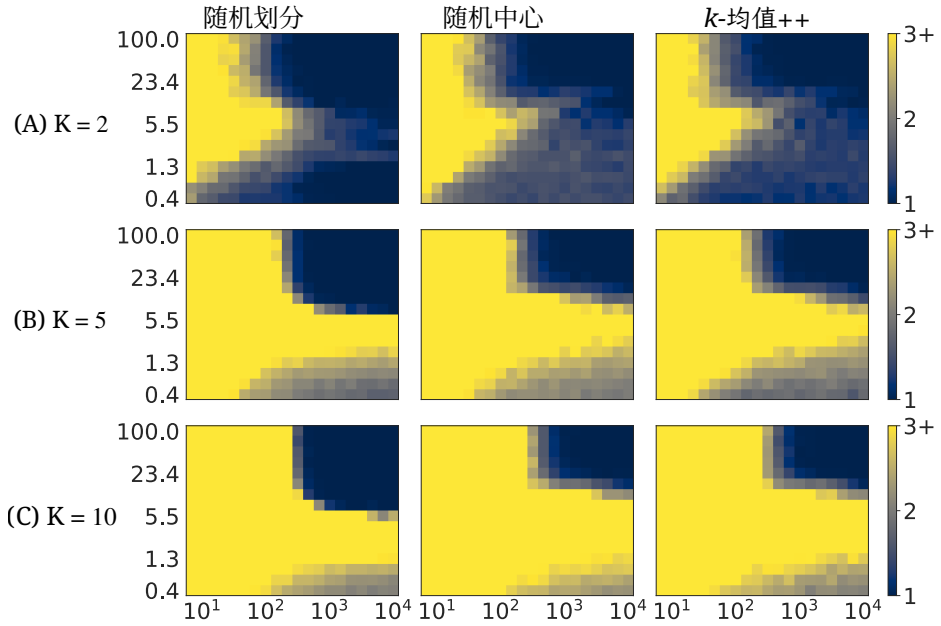


图5. Lloyd k-means 在不同初始化策略和类别数下的迭代次数， K。每个值对应 100 次独立实验的平均值。在每个实验中，我们从模型 2.1 （推广至 $K \geq 2$）中定义的高斯混合模型采样数据，其中 $\tau^2 = 1$ ，每类有 20 个样本，然后运行 Lloyd k-means，直到收敛。我们注意到，对于随机划分初始化的情况，我们排除了初始化本身可能构成不动点的情形。因此，即使划分在第一次迭代后保持不变，迭代次数也为 1。

这些结果可能因实现方式不同而有所变化，尤其是在使用 GPU 等并行计算资源时。

F.5. 算法不动点的不同表现

本节给出数值实验结果，以展示定理3.4 和 3.9。每个实验实例均为一次独立实验，其中数据（质心和样本）从模型 2.1所定义的高斯混合模型中采样，$n = 40$ 且 $\tau^2 = 1$。为将定理 3.4 和 3.9 中的表达式代入具体数值，并满足其假设，我们考虑真实簇和当前簇均衡的情形（$R^{\ell} = R^{\bar{\ell}} = 0.5$且 $|C_1| = |C_2|$） ，但当前簇的定义满足 $R_j^{\ell} = 0.25$ 且 $R_j^{\ell} = 0.75$。也就是说，我们将当前簇定义为：把样本从 C_j 切换到 C_j 会降低 k-means 损失（方程（5））。在每个实验实例中，我们考察 Lloyd k-means 或 Hartigan k-means 是否会将样本 i 移动到另一个簇。

对于几个不同的 10^4 和 σ^2 值，我们分别重复实验 d次，并在图9 中绘出样本 i 在各聚类算法的判据下仍留在当前簇中的实例比例。误差区间不易看清，采用的是 Wilson 区间（见定义A.12）。 σ^2 的取值定义为 $\sigma^2 = \beta\, \sigma_0^2$，其中 σ_0^2 是使定理3.4的假设成立所需的取值（方程（13））。

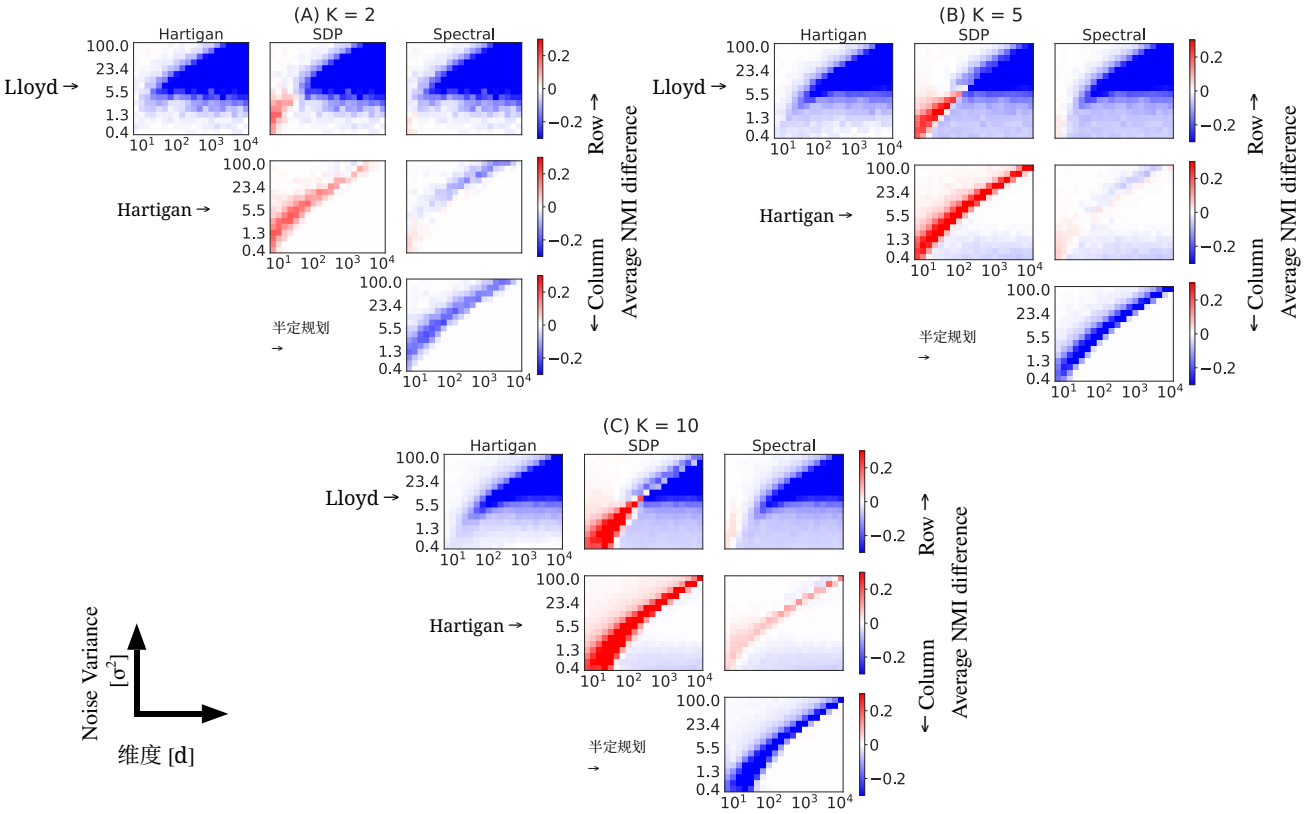


图 *6.* 不同聚类方法之间归一化互信息（NMI）值的比较。每个值表示 100 次独立实验的平均值；在每次实验中，数据均从第 2.1 节定义的高斯混合模型中抽样，$\tau^2 = 1$ 每类样本数为 20。对于每一对聚类方法，我们计算其 NMI 差值，并对所有实验的结果取平均值。色条上的箭头指示哪种颜色代表较好的性能。行对应色条中与向上箭头匹配的方法，列对应与向下箭头匹配的方法。

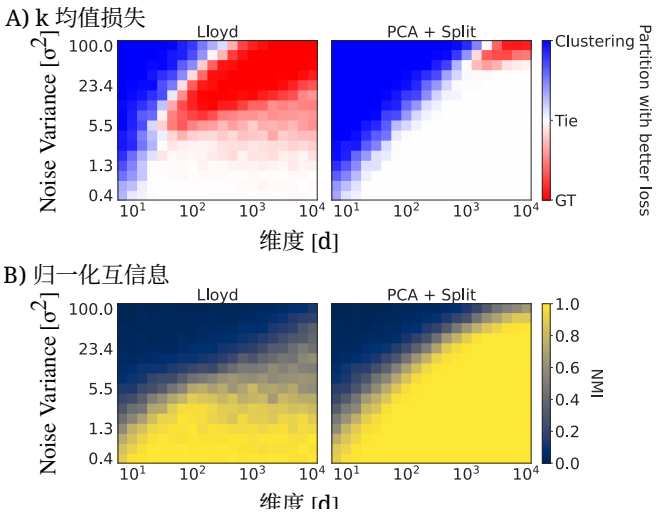


图 7. Lloyd 的 k-means 与“PCA + Split”所得到的划分与真实划分的比较，比较指标为 (A) k-Means 胜率（见第 F.1 节）和 (B) 标准化互信息（ NMI，见定义 A.13 ）。每个值对应 100 次独立实验的平均值；在每次实验中，我们从模型 2.1 定义的高斯混合模型中抽取数据，其中 $\tau^2 = 1.0$，且每类有 20 个样本。抽样时，我们确保真实簇的大小均衡。Lloyd 的 k-means 以 *k-means++* 初始化，并运行至收敛。

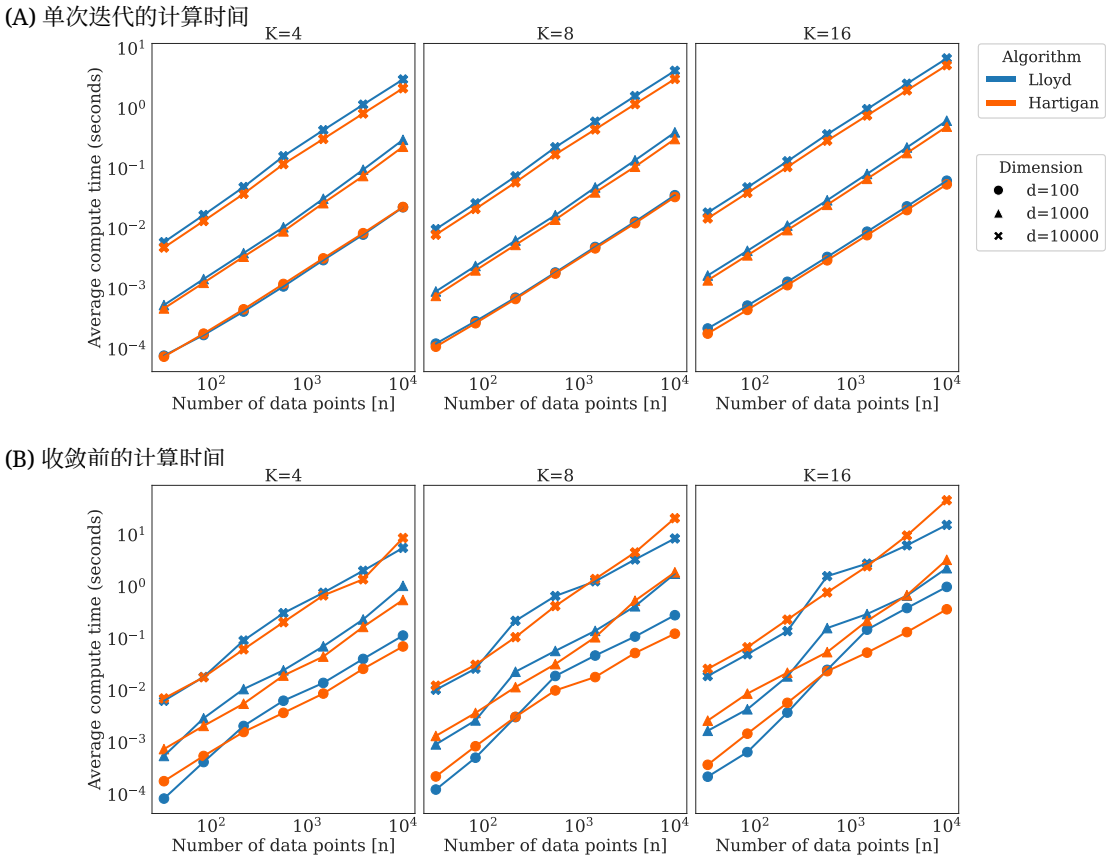


图 8. **Lloyd 和 Hartigan 的 k-Means 算法的计算时间**。每个值表示 10 次独立试验的均值，数据取自模型 2.1 (针对 $K \geq 2$ 推广）所述的高斯混合模型（ GMM ），其中 $\tau^2 = 1$，且 $\sigma^2 = 10$。面板 (A) 显示每种算法单次迭代的平均计算时间，面板 (B) 则显示每种算法达到收敛所需的平均计算时间。结果表明，在此实现中，Lloyd 和 Hartigan 算法的计算成本相近；Hartigan 方法有时收敛更快，因此性能更佳。

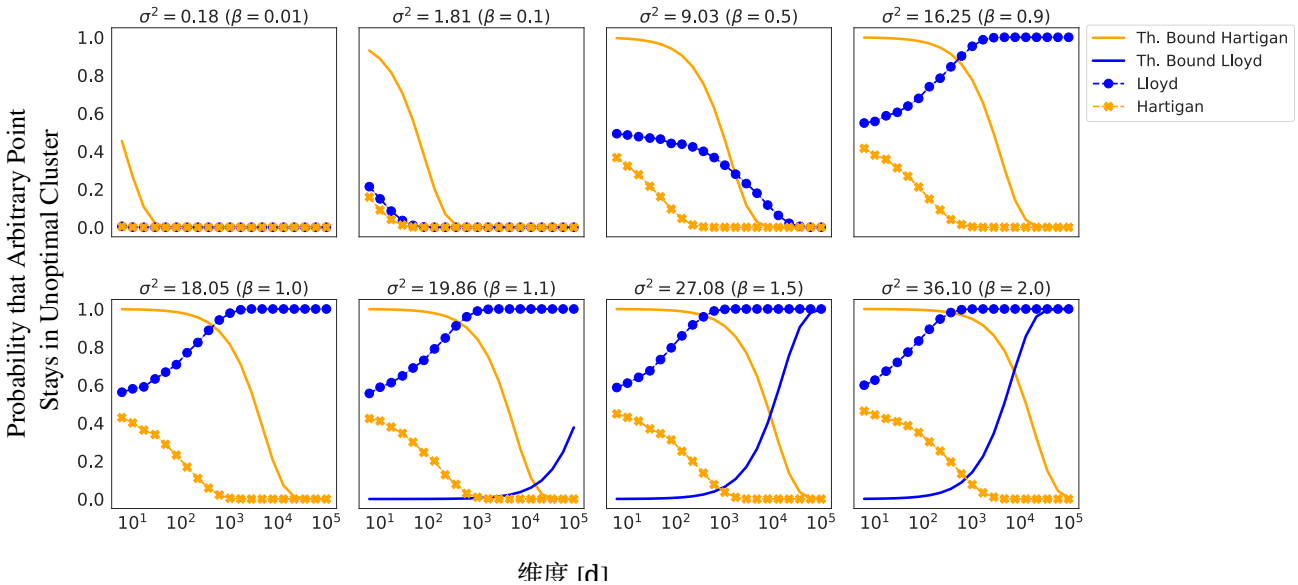


图 *9.* 定理 3.4 和 3.9的数值实验。每次实验中，数据（中心和样本）均从模型 2.1 中抽取，且仅考虑真实划分均衡的特殊情况，$n = 40$ 以及 $\tau^2 = 1$。此外，我们固定“当前簇”，使得对于任意样本 i，其纯度系数为 $R_j^\ell = 0.25$ 和 $R_j^\ell = 0.75$；其中，该样本的真实分配为 $i \in S_\ell^\star$，当前分配为 $i \in C_j$。图中展示了样本 i 在经过一步 Lloyd C_j -均值算法（蓝色圆圈）或 Hartigan k-均值算法（橙色叉号）后仍留在簇 k中的实例比例。图中还以实线标出了对应于各算法的该比例理论界。定理 3.4 的条件由 $\sigma^2 > 18.05$满足。我们观察到，随着维度增加，Lloyd 算法很少将样本移到另一个簇，这与 Hartigan 算法形成鲜明对比。