arXiv:2601.06375v1 [stat.CO] 10 Jan 2026

PCA引导的基于分位数的抽样，实现高效数据缩减

Hui-Mean Foo[1] 和 Yuan-chin Ivan Chang[1*†]

[1*]中央研究院统计科学研究所，台北市 Academia Road Section 2 128号，中华民国台湾 11529。

*通讯作者。电子邮箱：ycchang@as.edu.tw；贡献作者：huimean87@gmail.com；

†这些作者对本研究的贡献相同。

摘要

在大规模统计建模中，通过子抽样缩减数据规模，对于平衡计算效率与统计准确性至关重要。我们提出一种新方法——主成分分析引导的分位数抽样（PCA-QS）：将数据投影到主成分上，并采用基于分位数的抽样，以保留具有代表性且多样化的子集。与均匀随机抽样、杠杆分数抽样和核心集方法相比，PCA-QS 始终能取得更低的均方误差，更好地保留数据的关键特征，同时也具有较高的计算效率。该方法适用于多种数据场景，在统计计算领域展现出广泛的应用潜力。

关键词：数据缩减、主成分分析、分位数抽样

1 引言

高维数据集的分析是现代数据科学中的一项基础性难题。随着数据集的规模和复杂度不断增长，计算负担也随之加重，常常使传统分析技术不堪重负；这些技术力求兼顾效率与代表性。

主成分分析（PCA）广泛用于降维，能够有效捕捉低维表示中的方差。变分自编码器等替代方法则提供非线性变换，以学习潜在

高维数据中的表示（Vahdat 和 Kautz 2020）。然而，传统的基于 PCA 的方法并未明确确保缩减数据集能够保持原始分布的代表性，因此可能会丢失重要的结构信息。同样，基于分位数的方法提供了稳健的统计技术来处理偏斜或异质数据，但其主要用途是分布建模，而非数据缩减。现有方法通常要么侧重于降维而不进行实例选择，要么侧重于实例选择而不利用数据结构信息。所提出的 PCA-QS 方法通过将 PCA 与基于分位数的抽样相结合，弥补了这一不足，从而实现单独使用任一技术都无法达到的数据缩减效果。

主成分分析捕捉全局方差模式，但会舍弃原始特征空间，通常以牺牲可解释性为代价（Hotelling 1933；Jolliffe 2002）。t-SNE 和 UMAP 等非线性嵌入方法主要用于可视化和特征提取，而非系统性的数据缩减（Becht et al. 2019）。尽管这些方法能有效保留高维数据中的局部结构，却未能提供用于选择实例代表性子集的系统框架。随机抽样和分层抽样等基于抽样的方法旨在构建代表性子集，但可能引入采样偏差，或无法顾及底层数据结构（EasyBA.co2023）。已有研究提出了更复杂的方法（Sarwaret al. 2000；Erokhin et al. 2022；Grosset et al. 2020；AlKarawi and AlJanabi 2022），以平衡效率与代表性，但在确保数据完整性和计算可行性方面仍面临挑战。这些文献指出了具体挑战，但也存在明显的局限性。

与传统主成分分析不同，传统主成分分析会将变量转换为主成分并丢弃原始特征空间，而 PCA-QS 则利用主成分分析来指导实例选择，而非仅仅降低维数。为提高计算效率，研究者提出了随机主成分分析方法（Halko 等，2011），但这些方法主要关注线性投影，而非实例选择。随后，在 PCA变换空间中应用基于分位数的采样，以确保保留子集在结构上能够代表完整数据集。通过结合这两种技术，PCA-QS 在保留可解释性的同时显著提高了计算效率，克服了传统基于主成分分析的降维方法的一项关键局限。

PCA-QS 的优势不仅限于传统的抽样和降维技术。PCA-QS 将主成分分析用作组织工具，而非仅仅作为特征提取机制，从而确保缩减数据集既能保留方差结构，也能维持分布完整性。基于分位数的选择过程能够实现具有统计意义的抽样，捕捉完整数据集的变异性，并确保关键数据子群具有代表性。这种方法提出了一种数据缩减新方案，将结构洞察与代表性抽样相结合，为传统方法提供了切实可行的替代方案。当需要解释原始特征时，PCA-QS 尤其有用，例如在基因组学和社会科学等领域，理解各变量的贡献至关重要（Hotelling1933；Jolliffe 2002）。在机器学习预处理中，PCA-QS 通过支持使用紧凑的数据集进行模型训练，提升了计算可行性

，同时保留数据子集的多样性。此外，在大数据分析中，它为管理大规模数据集提供了有效策略，否则对这些数据集进行完整分析在计算上难以承受。为评估 PCA-QS 的有效性，我们使用线性回归模型和聚类应用开展了实证研究。这些研究通过比较在 PCA-QS 缩减数据集上训练的模型与其他方法，评估了关键性能指标，包括计算效率、预测准确性和信息保留。实证研究还探讨了 PCA-QS 在多种自适应抽样策略中的应用，包括 A-最优、D-最优、G-最优和基于不确定性的抽样，展现了其在不同实例选择框架中的灵活性。此外，我们还将 PCA-QS 应用于 UCI机器学习存储库等数据档案中的常用大规模数据集，展示了其在真实场景中的可扩展性。结果表明，PCA-QS 为监督学习和聚类任务中的高维数据降维提供了一种实用且计算高效的解决方案。

本文其余部分的结构安排如下。第2 节介绍 PCA-QS 框架及其实现。第 3 节讨论实证评估，重点关注计算效率和预测准确性。第 4 节总结主要发现、实际应用及未来研究方向。

2 方法

2.1 PCA-分位数抽样概述

PCA-QS 方法将 PCA 与基于分位数的抽样相结合，以实现高效的数据缩减，同时确保代表性实例选择。其主要目标是在保持统计完整性的同时缩小数据集规模，从而能够对高维数据进行高效建模，而不丢失关键结构属性。PCA-QS框架包含以下关键步骤：

1. *PCA*变换： 将数据投影到前 k 个主成分上，以组织数据结构。2.基于分位数的抽样： 在 PCA 变换空间中将实例划分为分位数组，以确保代表性均衡。3. 样本选择：根据给定的缩减率提取实例子集，同时保留原始特征空间。

Remark 1 面向二元数据的扩展 PCA 虽然主成分分析通常用于连续数据，但本研究使用的一些真实数据集包含二元变量。为适应这种情况，我们采用 `irlba` 软件包 *R*，它能够高效地为大型稀疏矩阵计算截断奇异值分解（SVD）。该软件包尤其适用于高维二元矩阵，因为它无需显式存储稠密矩阵，从而确保计算和内存效率（Lewis 和 Baglama，2021；Baglama and Reichel，2005；Baglama 等，2019，2024）。其他随机化分解方法见（Erichson 等，2016），更多应用见（Ai 等，2021；Taguchi and Turki，2021）。

主成分分析空间中的基于分位数的抽样

PCA-QS 不是随机选择实例，而是采用基于分位数的分层方法来保持数据集代表性。其核心思路是：(a) 沿每个主成分将数据划分为基于分位数的组；(b) 确保分布的各个部分都得到均衡的代表。

作为结构指南的主成分分析

主成分分析是一种基础的降维技术，可将高维数据投影到能够捕获最大方差的低维空间中（Hotelling 1933年；Jolliffe2002）。给定一个数据集 $\mathbf{X} \in \mathbb{R}^{n\times p}$，其中包含 n 个样本和 p个特征，主成分分析通过奇异值分解（奇异值分解）执行：

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top}, \tag{1}$$

其中：$\mathbf{U} \in \mathbb{R}^{n\times p}$ 包含左奇异向量（主成分），$\mathbf{D} \in \mathbb{R}^{p\times p}$是由奇异值构成的对角矩阵，$\mathbf{V} \in \mathbb{R}^{p\times p}$ 包含右奇异向量（主轴）。数据被转换到主成分分析空间中，如下所示：

$$\mathbf{Z} = \mathbf{U}_k\mathbf{D}_k, \tag{2}$$

其中，$\mathbf{Z} \in \mathbb{R}^{n\times k}$ 表示使用前 k个能够捕获大部分方差的主成分后，数据集在新的降维空间中的表示。尽管主成分分析能够捕获数据方差，但它并不执行实例选择，这意味着数据集仍然庞大且分析成本高昂，这也是仅依靠主成分分析仍然不够的原因。PCA-QS通过利用主成分分析进行结构化抽样，弥补了这一不足。

***PCA*变换空间（主成分数 $= 1$）**

为将数据划分为 g 个分位数组（$g \in N$），我们将分位数分层定义如下：

$$[\min, Q_1), [Q_1, Q_2), [Q_2, Q_3), [Q_3, Q_4), \ldots, [Q_{g-1}\,\max], \tag{3}$$

其中，Q_i 表示经PCA变换的数据集的第 i分位数。最终的缩减数据集通过从每个分位数组中抽取固定比例的数据构建而成，从而确保代表性，同时显著降低计算复杂度。

完成PCA变换后，会应用基于分位数的分层来组织数据。当主成分数 $= 1$ 为 $g = 5$时，方法沿第一个主成分将数据划分为基于分位数的组，如式（3）：

$$[\min, Q_1), [Q_1, Q_2), [Q_2, Q_3), [Q_3, Q_4), [Q_4, \max] \tag{4}$$

其中，Q_i 表示第 i个分位数。该划分确保每个分位数组约包含20%的样本，从而促进均衡且具有代表性的抽样。

num pcs > 1 的 ***PCA***变换空间

当保留多个主成分时（主成分数量为 > 1），经 PCA 变换后的空间将成为一个多维子空间。基于分位数的抽样不再局限于单一轴，而是需要采用结构化划分策略，以确保缩减数据集在多个主成分维度上保持代表性。

多个主成分的分位数分配

对于 $k > 1$，每个主成分在变换空间中定义一条轴，并且会针对每个主成分独立分配分位数。当 $g = 5$ 时，分位数分配遵循以下步骤：

1. **逐分量分位数：** 每个主成分 z_j 都会根据其分布划分为分位数，这与 num pcs $= 1$ 的情况类似。例如，划分为五个分位数时：

$$\text{Quantile Groups for Component } j : [\text{Min}, Q_1), [Q_1, Q_2), \ldots, [Q_4, \text{Max}].$$

2. **复合分位数组：** 根据每个样本在所有 k 个主成分上的分位数归属，为其分配一个复合分位数组：

$$\text{Quantile Group: } \{q_{i1}, q_{i2}, \ldots, q_{ik}\},$$

其中，q_{ij} 表示样本 i 在主成分 j上的分位数组。

3. **唯一组标识符：** 复合分位数组被编码为一个唯一标识符（例如，$k = 3$ 对应的“2-4-1”），其中每个数字代表样本在相应主成分上的分位数分配。

从分位数组中采样

为保持均衡代表性，将从每个分位数组中随机选择样本。随机选择可避免确定性偏差，同时确保按比例代表各个由分位数定义的子组。每组所选样本的数量由以下公式确定：

$$\text{Sample Size per Group} = \min\left(\lceil \text{retention rate} \times N_g \rceil, N_g\right),$$

其中，N_g 是组 g 中的样本数量，而保留率 （例如 0.05）控制保留的数据比例。这种方法在降低计算复杂度的同时保留了结构多样性。

虽然随机抽样能够确保公平性，但其他选择策略可以进一步优化样本的代表性。其他选择策略，例如基于实验设计准则（如 A-和 D-最优设计）以及基于不确定性的抽样，也可以纳入 PCA-QS 框架。在接下来的数值研究（第 3.3 节）中，我们将说明如何将自适应抽样策略有效地整合到 PCA-QS 中。

说明性示例

考虑一个包含 $n = 10{,}000$ 个样本和 $p = 50$ 个特征的数据集，并对其应用主成分分析，保留 $k = 2$ 个主成分。假设保留率设为 $\delta = 0.05$。PCA-QS 流程包括以下步骤:

1. 主成分分析将原始数据集转换为低维表示，得到一个 10，000×2矩阵**Z**，其中每一行代表主成分空间中的一个二维样本。2. 每个主成分划分为 5 个分位数。如果样本 i 属于第一个主成分的第 3 个分位数和第二个主成分的第 5 个分位数，则将其分配到复合分位数组 (3，5)。3. 由于共有 $5^2 = 25$ 个复合分位数组，每个样本只会被分配到一个组。4. 抽样：如果复合分位数组 (3，5)包含 $N_{3,\ 5} = 400$ 个样本，则该组的保留样本量为：**保留的样本** $= \lceil 0.05 \times 400 \rceil = 20$.5。对全部 25 个复合分位数组应用此流程，形成最终的保留数据集，确保主成分空间中的代表性。

PCA-QS的算法实现

为提供清晰的计算框架，我们介绍 PCA-QS 算法。

算法 1 PCA-分位数抽样（PCA-QS）框架

1: **输入：** 数据集 $\mathbf{X} \in \mathbb{R}^{n \times p}$，包含 n 个样本和 p 个特征。2: **预处理：** 将特征标准化为零均值和单位方差。3: **步骤 1：** 对 **X**应用主成分分析，将其转换到主成分空间。4: **步骤 2：** 保留能够解释足够方差的前 k 个主成分。5: **步骤 3：** 根据每个样本转换后的主成分分析坐标，将其分配到分位数组中。6: **步骤 4：** 根据给定的**保留率** δ，从每个分位数组中选择代表性实例。7: **步骤 5：** 构建用于下游分析的最终**保留数据集** $\mathbf{X}_{\text{retentive}}$。8: **输出：** 保留数据集 $\mathbf{X}_{\text{retentive}}$，同时保留代表性结构。

2.2 PCA-QS的数学性质

设 $X \in \mathbb{R}^{n \times p}$ 是一个多元正态随机矩阵： $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$，其中 Σ 是总体协方差矩阵。主成分分析通过计算样本协方差矩阵的特征向量，将数据集变换到一个正交基中:

$$S = \frac{1}{n} X^T X. \tag{5}$$

S 的特征向量定义了主成分（PC）的方向，而相应的特征值则表示其方差贡献。PCA变换

表示为： $Z = XV$，其中 V 是 Σ的特征向量矩阵。第一个主成分满足： $Z_1 \sim \mathcal{N}(0, \lambda_1)$，其中 λ_1 是最大特征值。

主成分分析与基于分位数的抽样

主成分分析会自然地将数据变换到一个不相关基中，从而能够准确地进行分位数抽样。即使原始特征之间存在相关性（$\Sigma \neq I$），主成分分析也能确保变换后的主成分彼此不相关。这为在主成分分析空间中进行基于分位数的分层提供了依据，因为每个主成分都可以视为独立的正态变量：

$$Z_j \sim \mathcal{N}(0, \lambda_j). \tag{6}$$

因此，无论特征依赖关系如何，基于主成分分析的分位数抽样都能保持稳健。

Remark 2 主成分分析会直接在变换空间中将协方差矩阵对角化，因此无需再进行显式的对角化步骤。如果原始协方差矩阵已经是对角矩阵，主成分分析只会对特征进行重新缩放，而 PCA-QS 会直接作用于这些重新缩放后的变量。人为对角化会消除特征之间有意义的关系，并降低主成分分析的效果。

主要主成分的分布

在低维情形（$p \ll n$）下，前 k 个主成分服从多元正态分布：$(Z_1, Z_2, ..., Z_k) \sim \mathcal{N}(0, \Lambda_k)$，其中 Λ_k =diag$(\lambda_1, \lambda_2, ..., \lambda_k)$ 是由最大的特征值构成的对角矩阵。

由于每个主成分都服从正态分布： $PC_j \sim \mathcal{N}(0, \lambda_j)$,其中 λ_j 是对应于 PC j的特征值，其分位数分布为：

$$Q_j(p) = \Phi^{-1}(p) \cdot \sqrt{\lambda_j}, \tag{7}$$

其中 $\Phi^{-1}(p)$ 是标准正态分位数函数。

使用前 k 个主成分进行分位数抽样

使用前 k 个主成分进行基于分位数的分层时，每个主成分的分位数都会按其特征值进行缩放。若使用前 5 个主成分进行抽样，则分位数分布如下：

$$Q_j(p) = \Phi^{-1}(p) \cdot \sqrt{\lambda_j}, \quad j = 1, ..., 5. \tag{8}$$

这确保 PCA-QS 方法根据经过方差缩放的正态分位数选择样本，从而保留原始分布结构。

示例：前 ***5*** 个主成分的理论分位数分布

使用 5 个分位数组，我们取：

$$(0\% \text{ (min)}, 20\%(Q_1), 40\%(Q_2), 60\%(Q_3), 80\%(Q_4), 100\% \text{ (max)}).$$

对于标准正态分布，这些值对应于：

$$\Phi^{-1}(0.2) = -0.84, \quad \Phi^{-1}(0.4) = -0.25, \quad \Phi^{-1}(0.6) = 0.25, \quad \Phi^{-1}(0.8) = 0.84.$$

然后将这些值乘以 $\sqrt{\lambda_j}$进行缩放，得到以下经验分位数：

分位数	PC1	PC2	PC3	PC4	PC5
20%	$-0.84\sqrt{\lambda_1}$	$-0.84\sqrt{\lambda_2}$	$-0.84\sqrt{\lambda_3}$	$-0.84\sqrt{\lambda_4}$	$-0.84\sqrt{\lambda_5}$
40%	$-0.25\sqrt{\lambda_1}$	$-0.25\sqrt{\lambda_2}$	$-0.25\sqrt{\lambda_3}$	$-0.25\sqrt{\lambda_4}$	$-0.25\sqrt{\lambda_5}$
60%	$0.25\sqrt{\lambda_1}$	$0.25\sqrt{\lambda_2}$	$0.25\sqrt{\lambda_3}$	$0.25\sqrt{\lambda_4}$	$0.25\sqrt{\lambda_5}$
80%	$0.84\sqrt{\lambda_1}$	$0.84\sqrt{\lambda_2}$	$0.84\sqrt{\lambda_3}$	$0.84\sqrt{\lambda_4}$	$0.84\sqrt{\lambda_5}$

表1：前5个主成分的分位数值

如果 λ_j 较小，分位数值仍接近于零，这表明相应的主成分未能捕捉数据中有意义的变异性。因此，从此类主成分进行基于分位数的抽样可能会引入偏差，且无法保留原始数据结构。为缓解这一问题，可将用于确定保留主成分数量（k）的经典方法应用于 PCA-QS。一种常见方法是累积方差阈值法，即保留主成分，直至其解释的总方差比例达到足够高的水平，通常为 95%。另一种广泛使用的技术是碎石图分析，它通过可视化方式识别方差解释量开始趋于平缓的节点。此外，交叉验证误差最小化可根据预测性能，以数据驱动的标准选择最优主成分数量。或者，也可以采用统计假设检验（如断棒法），通过评估各主成分解释的方差是否高于随机模型下的预期，来确定适当的维数。

Remark 3 （未来研究主题：高维 $p \sim n$ 情形）在维度 p以与 n相同的速率增长的超高维情形下，经典的主成分分析假设不再成立。具体而言，样本协方差矩阵的最大特征值不再服从正态分布，而是呈现非高斯波动。在这种情况下，最大特征值服从特雷西-威多姆分布，该分布描述了大型随机矩阵的极端特征值（Forrester 2010）：

$$\lambda_1 \approx \mu_\lambda + n^{-2/3} \cdot TW_1, \tag{9}$$

其中 TW_1 表示特雷西-威多姆分布（高斯正交系综情形）（Craig 和 Harold，1994、1996），而 μ_λ 则由马尔琴科-帕斯图尔定理（Marchenko 和 Pastur，1967）推导而来。这一结果是随机矩阵理论的基石，该理论刻画了大协方差矩阵中的特征值行为（Mehta，2004；Baik 等，1999；Johnstone，2001）。特雷西-威多姆定律对统计信号处理、无线通信和基因组学具有重要意义（Baik 等，1999）。尽管本研究未将 PCA-QS 扩展至高维情形 $p \sim n$，但该方法可以针对这类设置进行调整，这将是未来研究的一个很好的主题。

计算优势与挑战

PCA-QS 具有较高的计算效率，因为它只关注前几个主成分，从而同时降低内存占用和计算成本。它尤其适用于大规模数据分析，并且可以在并行或分布式计算框架中高效实现（Halko 等，2011；Martinsson 和 Tropp，2020；Erichson 等，2016；Liu 等，2017）。

在许多应用中，主成分分析通常用于连续数据。然而，一些真实世界数据集包含二元变量，因此需要对标准主成分分析进行调整。多个 R 包提供了处理此类情况的解决方案。例如，`irlba`包执行截断奇异值分解（SVD），尤其适用于二元矩阵，因为它无需显式存储大型矩阵，从而提高了内存效率和计算性能（Lewis 和 Baglama，2021）。另请参阅 Baglama and Reichel (2005)；Baglama 等 (2019、2024)，了解矩阵分解技术的更多详情。此外，随机化分解方法可为大规模问题提供高效近似，相关讨论见（Erichson 等，2016）。这些技术的更多应用见（Ai 等，2021；Taguchi 和 Turki，2021）。

为了突出 PCA-QS 的优势，我们根据现有文献，将其与广泛使用的数据缩减方法进行比较，包括随机抽样（Provost 等，1998；Shao，2003）、分层抽样（Shao，2003）、标准主成分分析（Jolliffe2002；Shlens，2014）、核主成分分析（Sch¨olkopf 等，1998；Mika 等，1999；Chang 等，2008；Li 等，2009）、t-SNE/UMAP（van der Maaten 和 Hinton，2008；McInnes 等，2018；Linderman 和 Steinerberger，2017；Belkin 等，2001）、杠杆分数抽样（Drineas 等，2006；Mahoney，2011a）和核心集抽样（Har-Peled 和 Mazumdar，2004；Feldman 和 Langberg，2011）。与可能丢弃关键数据结构的随机抽样，或计算成本高且缺乏可解释性的 t-SNE/UMAP 不同，PCA-QS 能高效保留数据结构和原始特征，同时可扩展至大型数据集。也就是说，PCA-QS 是一种数据删减工具，能够生成具有代表性的保留型子数据集，供下游分析使用。由于其子数据集中保留了所有特征，最终结果的可解释性取决于对保留数据集所采用的后续分析工具。与这些传统方法相比，PCA-QS 在计算效率、可解释性和可扩展性之间实现了良好平衡。下表总结了主要比较结果：

尽管表 2 中包含 t-SNE、UMAP 和核主成分分析，但需要注意的是，t-SNE、UMAP 和核主成分分析主要用于降维，而非数据缩减。因此，直接比较可能并不恰当，因为它们的目标与 PCA-QS 有着本质区别。不过，未来的研究可以将这些非线性嵌入技术整合到 PCA-QS框架中，以进一步提升实例选择和表示。

2.3 实际考量：挑战与权衡

尽管 PCA-QS 具有多项优势，但仍有一些重要的权衡因素需要考虑。一个关键挑战是分位数组数量呈指数级增长。每个主成分对应

方法	计算效率	可解释性	可扩展性
PCA-QS（所提出的方法）	High	High	High
随机抽样	High	Low	High
分层抽样	中等	High	中等
杠杆分数抽样	High	中等	High
核心集抽样	Low	中等	中等
标准主成分分析	中等	Low	High
核主成分分析	Low	非常低	Low
t-SNE/UMAP	Low	极低	Low

表 2：PCA-QS 与一些降维和抽样方法在计算效率、可解释性和可扩展性方面的比较。

5 个分位数，且保留 k 个主成分时，分组总数为 5^k。例如，当 $k = 3$ 时，共有 125 个可能的组，这会显著增加计算成本。此外，某些分位数组中的样本可能很少，甚至没有样本，尤其是在数据集较小或 k 较大时。这种稀疏性可能会降低基于采样数据集训练的模型的统计稳定性。

此外，尽管增加 k 可以保留更多方差并提高代表性，但也会增加计算开销。要实现有效的数据缩减，必须在准确性和效率之间取得适当平衡。另一个挑战是，沿主成分分布的数据高度偏斜，导致分位数组不平衡。在这种情况下，可能需要进行额外调整，以确保抽样在不同子组中仍具有代表性。

应用 PCA-QS 时，应仔细评估三个关键因素：保留的主成分数量（k）、最佳保留率（δ），以及处理稀疏分位数组的策略。较小的 k 可降低计算成本，但可能无法保留关键方差，从而导致模型性能下降。相反，较大的 k 可提高代表性，但会增加分位数组的数量，可能造成稀疏。根据累计解释方差选择 k 这一传统方法仍是有用的指导原则。

保留率应在计算效率与预测准确性之间取得平衡。典型取值范围为 0.01 到 0.1，具体取决于数据集的大小和维度。如果某些分位数组中的样本很少，可以采用过采样或合并相邻组等备用策略，以保持代表性。这些参数的选择取决于数据集的特征。以下章节中的数值研究将进一步阐述这些考量，并为选择合适的参数值提供实践指导。

Remark 4（历史与统计基础）既有研究已对主成分分析和基于分位数的方法的统计性质进行了广泛研究。主成分分析最初由Hotelling （1933年）提出，其理论性质已有充分记载，包括在正态性假设下的一致性和渐近效率（Jolliffe 2002）。主成分分析已广泛应用于数据科学中的降维、特征提取和噪声滤波，并在基因组学（Ringn´er 2008年）、图像识别（Turk and Pentland 1991）和推荐系统（Sarwar et al. 2000）等领域取得了显著应用。

同样，基于分位数的方法可追溯至Galton （1874），长期以来被广泛用于对偏态或异方差数据进行建模。此类方法的稳健性——例如对单调变换的不变性以及抵御离群值的能力——使其成为统计推断和预测建模中的重要工具（Hampel etal. 1986；Koenker and Bassett 1978）。这些基础性质凸显了 PCA-QS 作为一种混合方法的可靠性，它充分发挥了两种方法各自的优势。

3 数值研究

3.1 与常用抽样方法比较

数据生成

合成数据按如下方式生成。构造了一个高斯分布混合来表示协变量空间。具体而言，样本数设为 $n = 10,000$，特征数设为 $p = 50$。共定义了 K 个主成分，每个主成分的均值向量均从标准正态分布中抽取，并乘以五倍的系数；协方差矩阵为单位协方差矩阵：

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{0}, 5^2\mathbf{I}_p), \quad \boldsymbol{\Sigma}_k = \mathbf{I}_p.$$

混合权重从对称狄利克雷分布中抽取：

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}_K).$$

根据这些权重为每个观测抽取分量分配，并按如下方式生成特征：

$$\mathbf{x}_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

随后，通过抽取一个系数向量 $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ 并进行计算，生成了一个响应变量：

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1).$$

（随机种子固定为 42，以确保结果可复现。）

在深入讨论 PCA-QS 方法的详细性能之前，我们先进行一项简要的数值研究，比较通过杠杆分数抽样（Drineas 等，2006；Mahoney，2011a）、核心集抽样（Har-Peled 和 Mazumdar，2004；Feldman 和 Langberg，2011）和简单随机抽样获得的保留数据集。表A1总结了这四种数据缩减技术的表现：基于高斯混合合成数据集（包含 100000 个样本和 50 个特征），采用原始数据中 10% 的样本，对线性回归模型进行 1000 次重复试验评估。从表中可见，PCA-QS 始终在分布保真度与模型准确度之间实现了良好的平衡。它在所有方法中取得了最低的能量距离、马氏距离、KL散度和最大均值差异，表明其对整体数据结构的保留效果更佳。其均方误差（MSE）略

高于杠杆分数抽样和核心集抽样，但仍具竞争力。由于主成分分析分解和分位数分层，运行时间适中。杠杆分数抽样（Drineas 等，2006；Mahoney，2011a）的 MSE 仅次于核心集抽样，最低，这凸显了它在保留对线性模型拟合至关重要的影响力数据点方面的优势。然而，其较高的能量距离、马氏距离、KL散度以及异常大的最大均值差异，表明原始数据分布遭到了显著扭曲，这是因为该方法倾向于抽取极端杠杆点。核心集抽样（Har-Peled 和 Mazumdar，2004；Feldman 和 Langberg，2011）的 MSE 最低且方差为零，体现了其在概括簇均值方面的稳定性。不过，正如较高的分布距离和聚类迭代过程导致的最长运行时间所表明的那样，它在保留局部变化方面表现不佳。简单随机抽样是稳健的基准方法，执行速度快，且各项指标表现尚可。它不利用任何结构信息，但能提供无偏子样本，较好地近似原始分布。（图1 展示了本研究所用各项指标的箱线图，印证了数值研究的结果。为便于复现，所用 Python 代码可应要求提供。）

实践建议

由于这些方法的设计考虑和用途各不相同，杠杆采样会选取杠杆值极高的行，可能导致分布偏差，但有利于线性模型估计。PCA-QS 会有意在降维空间中扩大选取范围，因此能更好地匹配分布。核心集采用簇中心，对于降低各簇均值的误差很有效，但不利于呈现细节分布。简单随机抽样则进行均匀抽样，通常无偏，但并未针对任何一项任务进行优化。因此，在实践中，如果希望兼顾全局数据分布的保留和可解释性，PCA-QS 是最合适的方法。对于只关注线性模型预测准确性的任务，杠杆采样表现出色，但代价是分布相似性降低。对于基于聚类的汇总，核心集抽样最为适用，因为此时信号近似比保留分布细节更重要。当计算简便性至关重要时，简单随机抽样仍是实用的备用方案。

表 3：1000 次重复试验中各抽样方法的性能比较。每个均值后面的括号内为标准差。每项指标中最小（最佳）的值以粗体显示。

方法	PCA-QS	杠杆	核心集	SRS
MSE	1.0749 (0.0100)	**1.0719** (0.0107)	**1.0718** (≈0)	1.0745 (0.0103)
执行时间（秒）	0.211 (0.041)	0.092 (0.049)	4.615 (0.235)	**0.015** (0.020)
能量距离	**0.0138** (0.0031)	6.0695 (0.344)	0.9771 (≈0)	0.0325 (0.0164)
马氏距离	**0.2020** (0.0210)	1.1431 (0.0369)	0.4229 (≈0)	0.2111 (0.0216)
KL散度	**0.6781** (0.0573)	2.9623 (0.1508)	1.5117 (≈0)	0.6991 (0.0643)
MMD	**0.1785** (0.0890)	177.9425 (10.493)	29.2147 (≈0)	0.7427 (0.4920)

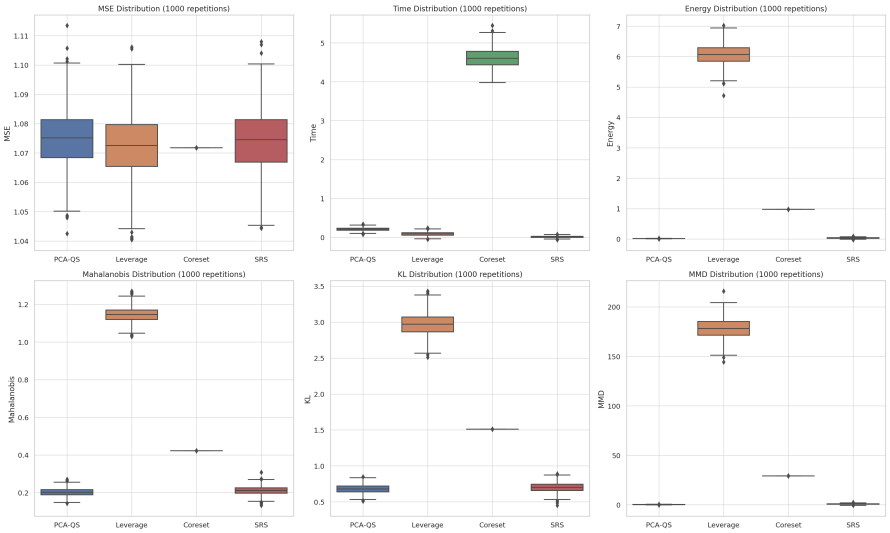


图1：比较基于高斯混合数据生成的数据集上，各种抽样方法得到的保留数据集与原始数据集之间的“距离”。此处结果基于 10% 抽样率下的 1000 次重复试验。

3.1.1 线性模型示例

为了评估 PCA-QS 在线性模型中的表现，我们生成了一个样本量大、特征维数高的合成数据集，并将 PCA-QS 抽样的结果与基于完整数据集得到的结果进行比较。

具体而言，我们使用MASS 程序包中的，000 个样本和 $p = 500$ 个特征，通过 mvrnorm 函数生成数据。协变量矩阵 X 服从均值为零且协方差矩阵 Σ具有等相关性的多元正态分布。将 Σ 的对角线元素设为 1（单位方差），所有非对角线元素设为常数相关系数 $\rho = 0.2$:

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}$$

回归系数向量 β 服从正态分布$N(0, 0.1_2)$，响应变量定义如下:

$$Y = X\beta + \varepsilon, \quad \text{where } \varepsilon \sim N(0, 0.5^2).$$

这一受控设置能够客观评估 PCA-QS 在保留与线性回归分析相关的信息方面的表现。

计算效率与预测性能

表4 和图2 全面比较了在 PCA-QS 保留子集上训练的线性回归模型的性能。评估涵盖了不同的 PCA 配置，包括固定数量的主成分（PC2、PC5、PC10）、动态选择策略（Dyn），以及两种基准方法：杠杆分数和核心集抽样。所有模拟均在特征相关性设为 $\rho = 0.2$、样本量为 $n = 10000$ 且特征维度为 $p = 500$的条件下进行。

计算效率

运行时间结果表明，与基线方法相比，PCA-QS 可显著节省计算资源。具体而言，PCA-QS 抽样的平均 CPU时间为 0.82 秒（PC2）至 0.86 秒（PC10），而动态变体（Dyn）由于需要自适应确定主成分数，计算耗时略高，为 2.27 秒。相比之下，杠杆分数抽样约需 0.69 秒，而核心集抽样的成本则显著更高，平均每次重复需 100.3 秒。因此，PCA-QS——尤其是固定主成分数的配置——在速度与表示质量之间实现了良好的权衡，同时避免了核心集方法带来的过高运行时间开销。

预测性能

就模型准确度而言，PCA-QS 在所有配置下都能实现较低的均方误差（MSE）和较高的 R^2 值。值得注意的是，动态变体（Dyn）在所有方法中表现最佳，MSE 为 0.238，R^2 为 0.953；它仅使用部分数据，性能却与完整模型的表现十分接近。杠杆分数方法和核心集方法的 MSE 均较高（分别为 0.464 和 0.475），R^2 值则较低（0.908 和 0.905）。固定主成分数的 PCA-QS 变体（PC2、PC5、PC10）表现稳定，R^2 值介于 0.902 至 0.920 之间。这些结果证实，PCA-QS 在大幅减少计算时间和数据量的同时，仍能保持较高的预测准确性。

计算效率

当主成分数量较少（PC2 和 PC5）时，PCA-QS 的相对 CPU 时间仍然最低，但随着纳入更多主成分，运行时间会增加，这反映了额外的计算成本。动态主成分方法由于需要额外确定最佳主成分数量的步骤，因此耗时更长；而核心集方法的总体运行时间最高。在所有方法中，低主成分数量下的 PCA-QS 之后，杠杆采样的执行速度最快。附录中对 PCA-QS、杠杆采样和核心集抽样的理论特性及计算权衡进行了综合比较 A.5。

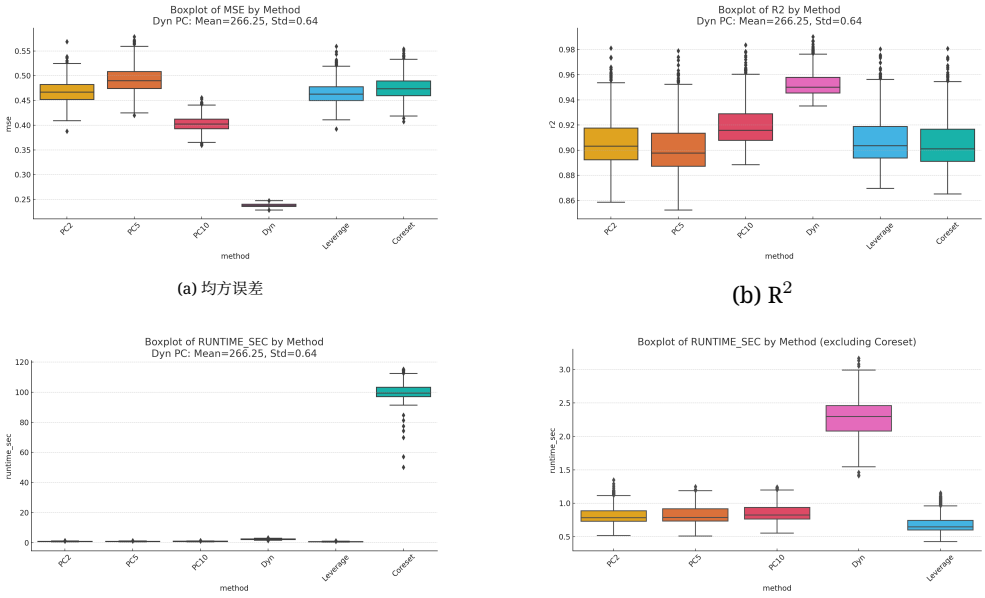
最佳主成分数量

这些发现表明，对于 PCA-QS，使用 2 到 5 个主成分能够在计算效率与预测性能之间实现良好平衡。虽然增加主成分（例如 PC10 或动态主成分）可能会略微提升模型准确度，但计算成本会显著上升。总体而言，PCA-QS

表4：在采样子集上训练的线性回归模型的性能比较。每个单元格显示 1000 次运行的均值（标准差）。`Dyn` （采用动态主成分选择的 PCA-QS）准确性最高。杠杆 速度最快，而核心集的计算开销最大。

方法	MSE	R2	运行时间（秒）
PC2	0.4674 (0.0219)	0.9069 (0.0196)	0.8192 (0.135)
PC5	0.4912 (0.0252)	0.9021 (0.0207)	0.8267 (0.1383)
PC10	0.4029 (0.0151)	0.9197 (0.0167)	0.86 (0.1415)
Dyn	**0.238 (0.0034)**	**0.9526 (0.0097)**	2.2668 (0.273)
核心集	0.4747 (0.0221)	0.9054 (0.0198)	100.3096 (5.3999)
杠杆	0.4637 (0.0211)	0.9076 (0.0193)	**0.6922 (0.1451)**

在此表中，Dyn 表示基于 70% 主成分方差载荷确定主成分数量的情形。使用动态主成分选项时实际使用的主成分数量：均值：266.248，标准差：0.6395，最小值：264，最大值：268。



(a) 均方误差

(b) R^2

(c) 运行时间（秒）

(d) 运行时间（不包括核心集）

图 2：不同抽样方法下模型性能指标的箱线图。Dynamic70 组归入标签`Dyn`，所选主成分数量的均值为 $= 266.25$（标准差为 $= 0.64$）。（动态主成分选项下实际使用的主成分数量：均值：266.25，标准差：0.64，最小值：264，最大值：268。）由于核心集方法的计算成本不同，运行时间分别报告了包含和不包含核心集方法的情况。

在保持较高预测准确性的同时，数据缩减效果和运行时间节省均显著提升；在本模拟设置下，采用约 2–5 个主成分的配置似乎最为理想。

Remark 5 主成分数量的选择可依据方差解释比例，通常基于传统的 PCA 分析选择诸如 95% 这样的阈值（Jolliffe 2002；Jackson 1991）。或者，计算成本或可解释性等实际约束可能要求选择较少的主成分。关于 PCA 方差载荷和成分选择的深入讨论，可参见相关文献（Jolliffe 2002；Abdi and Williams 2010a）。为简明起见，本研究未实施这一自动选择步骤。

大规模线性模型模拟。
为考察 PCA-QS 在更大规模数据设置下的可扩展性，我们在保持所有其他参数不变的情况下，增大样本量至 $n = 100,\,000$，并重复了第 3.1.1 节中的线性模型模拟。结果表明，即使在这一大规模场景下，PCA-QS 仍能保持较强的预测性能，并显著减少数据规模和计算时间。

表（5）大规模线性模型模拟（$n = 100,\,000$）中不同抽样方法的性能指标汇总。

方法	能量距离	马哈拉诺比斯	KL散度	MMD
PCA-QS	0.00005 (≈ 0)	0.0154 (0.0001)	0.00117 (≈ 0)	0.0003 (≈ 0)
杠杆	0.0789 (0.0084)	0.5221 (0.0275)	1.9505 (0.1267)	0.4361 (0.0493)
核心集	0.2342 (≈ 0)	0.8566 (≈ 0)	3.587 (≈ 0)	1.336 (≈ 0)
SRS	0.0018 (0.0003)	0.0911 (0.0100)	0.3192 (0.1240)	0.0084 (0.0022)

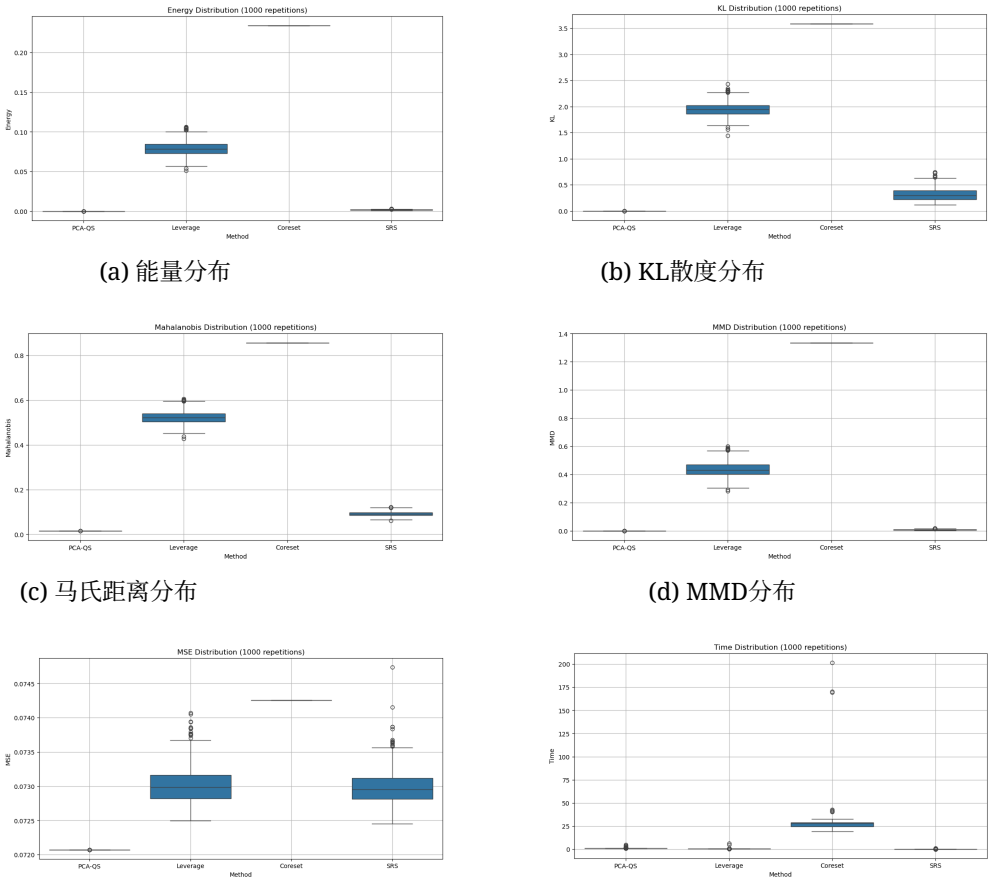
(a) 距离和散度度量

方法	MSE	运行时间（秒）
PCA-QS	0.0720 (≈ 0)	0.8615 (0.1757)
杠杆	0.0730 (0.0003)	0.1549 (0.2389)
核心集	0.0742 (≈ 0)	27.4645 (8.9609)
SRS	0.0729 (0.0002)	0.0131 (0.0335)

(b) 性能指标

3.2 PCA-QS 在聚类问题中的应用

除了本文所呈现的数值研究（这些研究聚焦于线性回归和逻辑回归模型，以及应用于各类真实世界数据集的自适应设计）之外，本节还探讨了 PCA-QS 在聚类问题中的应用。



(a) 能量分布

(b) KL散度分布

(c) 马氏距离分布

(d) MMD分布

(e) MSE分布

(f) 时间分布

图 3：不同抽样方法在 1,000 次重复实验中的性能指标分布。

鉴于聚类方法本身固有的复杂性，本文并未对聚类进行全面考察，但仍为未来研究提供了宝贵见解。具体而言，本文将 PCA-QS 应用于森林覆盖类型数据集这一高维真实世界数据集，以评估其作为聚类任务降维技术的实用性。

降维技术，尤其是主成分分析，因其能够在简化数据的同时保持分析准确性而受到广泛研究 Jolliffe (2002); Abdi 和 Williams (2010b). 然而，聚类本身仍然颇具挑战，因为它需要进行参数选择，包括确定最佳簇数、选择合适的相似性度量，以及处理对噪声和离群值的敏感性 Hamerly 和 Elkan (2003); Tibshirani 等人 (2001). 聚类性能通常使用轮廓系数等指标进行评估，该指标衡量簇内凝聚度和分离度 Rousseeuw (1987), 以及戴维斯-博尔丁指数，

该指数评估紧凑度和分离度Davies 和 Bouldin (1979). 尽管这些指标稳健且易于解释，但根据所用数据集和聚类算法的不同，评估结果往往会相互矛盾 Vendramin 等人 (2010).

通过比较使用 K 均值聚类对保留数据集（通过 PCA-QS 生成）和完整数据集进行聚类所得的结果，我们发现两者高度相似。这表明 PCA-QS 有望在保留关键数据结构的同时提高计算效率并减少噪声。这些发现与现有研究一致，后者强调了在聚类任务中保持缩减数据集结构完整性的重要性 Jain (2010)；Ding 和 He (2004)。然而，本研究也承认聚类分析面临着更广泛的挑战，尤其是难以定义普遍最优的参数或性能标准，这使得聚类高度依赖具体情境 Arbelaitz 等人 (2013)。这一补充拓宽了本文的研究范围，将聚类应用纳入讨论，并为今后研究 PCA-QS 在不同分析范式中的适用性奠定了基础。

3.2.1 使用 PCA-QS 方法对森林覆盖类型数据进行聚类

森林覆盖类型数据集概述

我们还使用森林覆盖类型数据集进行说明，尽管它主要用于分类任务。如前所述，该数据集包含源自遥感的制图属性和土壤特征，用于预测不同土地区域的森林覆盖类型。目标变量表示七种不同的森林覆盖类型，例如云杉/冷杉、扭叶松、黄松和山杨。由于具有高维特性，且包含类别变量和连续变量，森林覆盖类型数据集常用于评估机器学习算法、聚类技术和特征选择方法。

聚类实验设置

使用 PCA-QS 从森林覆盖类型数据集中提取代表性子集，然后进行聚类实验。为确保评估结果稳健，实验重复进行了 1,000 次。

聚类过程分为以下三个关键步骤：

1. PCA-QS 抽样：采用分层抽样方法，使用 PCA-QS 从数据集中选取 20%（缩减率 $\delta = 0.2$）。这样可以确保抽样数据保留完整数据集的分布特性。

2. K-Means 聚类：分别对保留（抽样）数据集和完整数据集应用 K-Means 算法，形成七个簇，分别对应七种森林覆盖类型。算法采用 10 组不同的随机初始值进行初始化，以提高稳定性。

3. 性能评估：使用保留数据集的质心预测其余数据点的簇分配，并与使用完整数据集获得的结果进行比较。采用并行化轮廓系数计算方法，分别计算两组数据的轮廓系数，以此评估聚类性能。

这种方法确保了聚类性能评估具备可扩展性和效率，同时评估了 PCA-QS 在保持完整数据集结构完整性方面的有效性。

聚类性能与结果

在 99.5% 的情况下（1,000 个案例中的 995 个），保留数据集与完整数据集的轮廓系数之间的绝对差异小于 0.1。在更严格的 0.05 阈值下，仍有 88.9% 的情况（1,000 个案例中的 889 个）符合标准。这项分析表明，即使在更严格的条件下，这两个指标之间仍具有高度一致性。

图 4 展示了 1,000 个观测值中保留数据集的轮廓系数与完整数据集的轮廓系数之间差异的分布。x 轴表示差异（保留数据集轮廓系数 –完整数据集轮廓系数），而 y 轴表示观测值的数量。

大多数差异集中在零附近，表明两个指标之间具有高度一致性。两条阈值线突出显示了关键的一致性范围：

1. 绿色虚线（± 0.05）：889 个观测值（88.9%）落在此范围内，另有 111 个超出该范围。2. 红色虚线（± 0.1）：995 个观测值（99.5%）处于此阈值内，仅有 5 个超出该阈值。

对称分布进一步表明，正向或负向偏差都较为均衡。该可视化有效展现了保留数据集与完整数据集的聚类结果高度相似，只有少数观测值呈现出显著差异。

数据删减对聚类模型的影响

线性模型和聚类模型对数据删减效应（如缺失值、离群值和数据缩减）的敏感程度存在显著差异。线性模型（包括逻辑回归）尤其容易受到缺失数据的影响，通常需要进行插补，以避免参数估计产生偏差 Little 和 Rubin (2002)。离群值也会扭曲预测结果，因此需要采用稳健回归技术Rousseeuw 和 Leroy(2005)。虽然主成分分析能提升可解释性并缓解多重共线性，但不恰当的特征移除可能会降低预测能力Jolliffe (2002)。

聚类模型对不同算法表现出不同的敏感性。K均值等基于距离的方法难以处理缺失值和高维空间，常会导致聚类结果不准确（Hastie 和 Friedman 2009）。这类方法也容易受到离群值的影响，导致簇质心发生偏移；而 DBSCAN 等基于密度的方法则更加稳健 Ester 等 (1996)。维数灾难进一步增加了聚类的难度，因此通常需要采用特征选择或结构化采样技术，以确保进行具有代表性的实例选择。

总体而言，与线性模型相比，聚类模型往往对数据缩减的影响更为敏感，尤其是在高维场景下。这凸显了应用 PCA-QS 时采用定制化预处理策略的必要性，也进一步体现了该方法在回归和聚类场景中的实用价值。

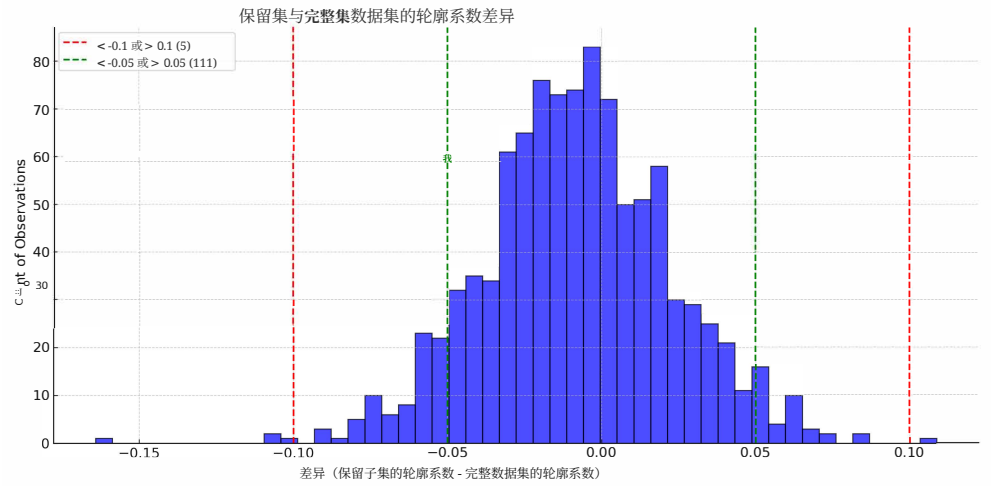


图4：保留集与完整数据集之间差异的轮廓系数分布。直方图展示了保留子集与完整数据集之间轮廓系数的差异。绿色虚线表示阈值± 0.05，涵盖 88.9% 的情况；红色虚线表示阈值 ± 0.1，涵盖 99.5% 的情况。

3.3 PCA-QS 与其他基于实验设计的自适应抽样策略

自适应抽样广泛应用于现代大数据分析，以便在大型数据集中高效定位信息丰富的数据点。在此前的数值研究中，PCA-QS 方法采用随机抽样方案，即在主成分分析空间中定义的分位数组内随机选择数据点。由于 PCA-QS 会为数据点分配主成分分析分位数索引，因此该方法可以无缝纳入随机抽样之外的其他抽样策略。本节将评估 PCA-QS 在自适应抽样方案下的表现。该方法采用迭代流程，根据特定优化准则选择能够最大化信息增益的数据点。我们将探究 PCA-QS 是否能从自适应抽样技术中获益，尤其是 A-最优、D-最优、基于 G-最优设计的方法以及不确定性抽样。

数据生成设置与线性模型情形保持一致，其中每种抽样方案均应用于由 100,000 个样本和 500 个变量组成的数据集。结果基于 1,000 次重复试验，以确保稳健性。以下各小节将简要介绍每种自适应抽样方法，随后通过数值评估展示 PCA-QS 如何在不同实验设计准则下保持计算效率和统计完整性。

***A*-最优设计**

A-最优设计通过降低逆费舍尔信息矩阵（费舍尔信息矩阵）的迹，最小化参数估计值的平均方差。该准则提高了

模型参数估计值的平均精度，因此广泛适用于回归建模以及准确参数估计至关重要的场景（Pukelsheim 2006；Fedorov 1972）。

表6展示了 PCA-QS 在 A-最优方案下的计算和预测性能。PCA-QS 显著节省了计算量，保留数据集的 CPU 时间为 0.0207 至 0.0195 秒，而完整数据集则超过 2.8 秒。尽管保留数据集的均方误差略高于完整数据集（≈ 0.3126 对比 0.2498），但相对均方误差保持稳定，约为 1.25，表明参数估计精度得到了良好维持。下表中，我们用 Ret 和 Full 分别表示保留子集和完整数据集。效率得分始终较高，在 PCA =2 时达到峰值 2,423，这表明 PCA-QS 中的 A-最优性有效地平衡了计算效率和统计准确性。

表 6：采用 A-最优设计的 PCA-QS

PCA	主成分分析 = 1	主成分分析 = 2	主成分分析 = 3	主成分分析 = 4	主成分分析 = 5
CPU Ret	0.0207 (0.0038)	0.0196 (0.0026)	0.0192 (0.0022)	0.0195 (0.0021)	0.0195 (0.0022)
CPU Full	2.8298 (0.4576)	2.9187 (0.5514)	2.7643 (0.4902)	2.7573 (0.4783)	2.8035 (0.5369)
MSE Ret	0.3126 (0.0087)	0.3125 (0.0088)	0.3125 (0.0088)	0.3126 (0.0089)	0.3126 (0.0089)
完整均方误差	0.2498 (0.0011)	0.2498 (0.0011)	0.2498 (0.0011)	0.2498 (0.0011)	0.2498 (0.0011)
保留模型的R平方	0.7408 (0.0266)	0.7404 (0.0267)	0.7403 (0.0266)	0.7403 (0.0265)	0.7403 (0.0265)
完整 R 平方	0.7928 (0.0217)	0.7924 (0.0218)	0.7923 (0.0217)	0.7924 (0.0215)	0.7924 (0.0215)
相对均方误差	1.2516 (0.0329)	1.2512 (0.0329)	1.2511 (0.0332)	1.2515 (0.0333)	1.2515 (0.0334)
效率得分	2229.26 (464.85)	2423.26 (544.27)	2332.71 (520.55)	2292.76 (477.98)	2330.95 (516.62)

采用 ***D***最优设计的 ***PCA-QS***

D-最优设计通过最大化费 Fisher 信息矩阵（FIM）的行列式，从而最大化参数估计的置信椭球体积。这一准则广泛应用于实验设计，尤其是在药物试验和工程学等领域，在这些领域中，提高参数估计效率至关重要（Fedorov1972；ChalonerandVerdinelli 1995）。

表7 展示了 PCA-QS 在 D-最优设计下的表现。PCA-QS 保持了较高的计算效率，保留子集的 CPU时间最低仅为 0.0201 秒，而完整数据集则超过 3 秒。尽管保留子集的均方误差（0.3125）仍略高于完整数据集（0.2493），相对均方误差稳定在约 1.25，表明参数估计准确度稳定。PCA= 3的效率得分峰值为 2,581，展现了 PCA-QS 在 D最优准则下的可扩展性和稳健性。

表7：PCA-QS 采用 D-最优方案时的性能

PCA	主成分分析 = 1	主成分分析 = 2	主成分分析 = 3	主成分分析 = 4	主成分分析 = 5
CPU Ret	0.0203 (0.0027)	0.0217 (0.0056)	0.0202 (0.0033)	0.0201 (0.0029)	0.0209 (0.0036)
CPU Full	2.7365 (0.4514)	3.2923 (0.6328)	3.1911 (0.6572)	2.8885 (0.5314)	3.1156 (0.5886)
MSE Ret	0.3122 (0.0158)	0.3126 (0.0165)	0.3125 (0.0165)	0.3125 (0.0164)	0.3126 (0.0165)
完整均方误差	0.2493 (0.0009)	0.2493 (0.0008)	0.2493 (0.0008)	0.2493 (0.0008)	0.2493 (0.0008)
R^2 保留集	0.7508 (0.0205)	0.7515 (0.0191)	0.7516 (0.0191)	0.7515 (0.0191)	0.7515 (0.0191)
R^2 完整集	0.8007 (0.0164)	0.8016 (0.0157)	0.8016 (0.0157)	0.8015 (0.0157)	0.8016 (0.0157)
相对均方误差	1.2521 (0.0629)	1.2539 (0.0652)	1.2539 (0.0653)	1.2537 (0.0649)	1.2540 (0.0654)
效率得分	2183.67 (410.25)	2503.78 (618.14)	2581.74 (673.00)	2336.44 (540.93)	2446.10 (600.01)

采用 ***G-*最优设计的 *PCA-QS***

G-最优设计可使设计空间内预测值的最大方差最小化，从而确保预测误差均匀分布。这种方法在以预测为重点的应用中尤为有用，例如机器学习和响应面优化（Pukelsheim 2006；Cox andReid1958）。

在 PCA-QS 中，G-最优性可在保持计算效率的同时降低极端方差。不同于最小化平均方差的其他准则，G-最优设计确保没有任何单一区域主导预测误差，因此非常适合在整个设计空间内实现均匀的性能。

实证结果表明，在 G-最优性准则下，PCA-QS 能显著节省计算开销：保留子集的 CPU 时间为 0.0819 至 0.0203 秒，而完整数据集的耗时超过 3 秒（见表 8）。不同 PCA 配置下的均方误差略有差异，最高值出现在 PCA=3和 PCA=5 （约为 0.3139）。相对均方误差增幅约为 1.26，这与优先降低最大方差而非优化平均性能的目标一致。效率得分在 PCA=5 时达到峰值（约为 2,521），这表明较高的主成分配置在控制方差的同时可带来更大的计算开销节省。这些结果表明，在 G-最优设计下，PCA-QS 能有效平衡预测均匀性与计算效率。

结合基于不确定性的设计的 ***PCA-QS***

基于不确定性的自适应抽样通过纳入模型不确定性和先验知识，动态更新抽样策略，并逐步降低不确定性。这种方法对于环境建模、机器人学和贝叶斯优化等领域至关重要（Chaloner 和 Verdinelli 1995；MacKay1992）。

表9 显示，在基于不确定性的抽样下，PCA-QS 保持了稳定的计算效率，保留子集的中央处理器耗时为 0.0953 至 0.0967 秒，而完整数据集约为 3 秒。不同 PCA 配置下，保留数据集的均方误差（0.2541）变化极小，与完整数据集（0.2493）非常接近。相对均方误差保持稳定（1.02），

表 8：PCA-QS 在 G-最优设计下的表现

指标	主成分分析 1	主成分分析 2	主成分分析 3	主成分分析 4	主成分分析 5
CPU Ret	0.0819 (0.0117)	0.0203 (0.0032)	0.0197 (0.0027)	0.0212 (0.0042)	0.0203 (0.0032)
CPU Full	2.0987 (0.5749)	2.9489 (0.4652)	3.0283 (0.6183)	3.0811 (0.5421)	3.1424 (0.6203)
MSE Ret	0.2539 (0.0018)	0.3131 (0.0164)	0.3139 (0.0173)	0.3138 (0.0173)	0.3139 (0.0174)
完整均方误差	0.2493 (0.0015)	0.2491 (0.0007)	0.2490 (0.0005)	0.2490 (0.0005)	0.2490 (0.0005)
保留模型的R平方	0.7926 (0.0241)	0.7542 (0.0190)	0.7547 (0.0177)	0.7546 (0.0177)	0.7546 (0.0176)
完整 R 平方	0.7964 (0.0236)	0.8042 (0.0140)	0.8051 (0.0129)	0.8051 (0.0128)	0.8052 (0.0128)
相对均方误差	1.0185 (0.0027)	1.2567 (0.0644)	1.2603 (0.0678)	1.2600 (0.0677)	1.2606 (0.0680)
效率得分	502.78 (110.61)	2369.16 (511.72)	2489.96 (632.38)	2381.87 (596.49)	2521.60 (638.61)

并且在较高的 PCA 配置下，效率得分超过 600，展现了 PCA-QS 在考虑不确定性的场景中的稳健性。

表 9：PCA-QS 与基于不确定性的抽样

PCA	主成分分析 $= 1$	主成分分析 $= 2$	主成分分析 $= 3$	主成分分析 $= 4$	主成分分析 $= 5$
CPU Ret	0.0953 (0.0119)	0.0966 (0.0174)	0.0975 (0.0173)	0.0966 (0.0186)	0.0967 (0.0162)
CPU Full	2.6915 (0.3769)	2.9057 (0.5384)	2.9637 (0.5448)	3.0313 (0.5672)	2.9112 (0.4859)
MSE Ret	0.2541 (0.0015)	0.2541 (0.0015)	0.2541 (0.0015)	0.2541 (0.0015)	0.2541 (0.0015)
完整均方误差	0.2493 (0.0011)	0.2493 (0.0011)	0.2493 (0.0011)	0.2493 (0.0011)	0.2493 (0.0011)
保留模型的R平方	0.7833 (0.0188)	0.7829 (0.0187)	0.7829 (0.0187)	0.7829 (0.0187)	0.7828 (0.0187)
完整 R 平方	0.7874 (0.0184)	0.7870 (0.0184)	0.7870 (0.0183)	0.7870 (0.0183)	0.7869 (0.0183)
相对均方误差	1.0194 (0.0026)	1.0194 (0.0026)	1.0194 (0.0026)	1.0194 (0.0026)	1.0194 (0.0026)
效率得分	560.58 (95.88)	604.98 (142.98)	611.33 (146.19)	634.99 (159.49)	604.45（135.37）

从上述实验可以看出，在所有自适应抽样方案中，PCA-QS 始终能够在保持统计完整性的同时降低计算成本。该方法能够有效适应各种设计准则，包括方差最小化、行列式最大化和基于不确定性的策略。效率得分和相对均方误差值体现了其可扩展性和多功能性，使 PCA-QS 成为适用于多种建模场景的高维数据缩减实用工具。

PCA-QS 中的分位数分组具有显著的计算和统计优势，尤其适用于基于统计实验设计准则的自适应抽样设计。由于这些方法会在预定义组内评估每个样本，分位数分组能够确保代表性、保留数据异质性并增强模型鲁棒性。通过将数据分层为基于分位数的组，可以高效分配计算工作量，减少冗余计算并

提高可扩展性。这种结构化方法有助于系统地探索特征空间，从而提高迭代抽样和决策过程中的估计准确性与稳定性。

在每个分位数组内均衡分配样本，可使自适应算法有效优化实验设计准则，减少偏差，并提升参数估计和预测性能。在高维线性回归模型（500 个特征、100,000 个样本）上的 PCA-QS 验证表明，保留数据集与完整数据集的协方差和响应向量并无显著差异，这一点已通过柯尔莫哥洛夫-斯米尔诺夫检验和 t 检验得到验证。该方法适用于更广泛的应用。

此外，PCA-QS 与并行和分布式计算高度兼容。通过在 PCA变换空间中对数据分组，它能够在各组内独立抽样，从而显著降低自适应抽样方案（如 A-最优性和 D-最优性方案）的计算复杂度。

3.4 真实世界数据集的相似性分析

以下分析的主要目标是考察每种抽样方法在数据缩减后保留原始数据统计分布的效果。因此，除合成数据实验外，我们还在 UCI机器学习存储库中的多个真实世界数据集上评估 PCA-QS、杠杆分数抽样、核心集抽样和简单随机抽样的性能。我们使用多种距离度量评估分布相似性，包括能量距离、马氏距离、Kullback–Leibler（KL散度）和最大均值差异（MMD）。这些指标从全局结构和局部变异性两方面，量化保留数据集对原始数据分布的近似程度。本研究选用的数据集涵盖多个领域和不同维度：

CAPS 数据集：用于心理健康研究的生理和行为测量数据集合。

在线新闻受欢迎度：与在线文章及其社交媒体 a
受欢迎度相关的特征。
共享单车数据集（小时）：自行车租赁的小时记录，包括 g
时间、天气和节假日特征。
应用预测数据集：预测应用安装或使用情况的人口统计和行为变量。

对于每个数据集，我们采用统一的 10% 保留率，重复进行所有实验 1,000 次，并报告每个指标的均值和标准差。表10 展示了 CAPS、在线新闻受欢迎度、自行车共享和器具能量距离的结果。PCA-QS 和简单随机抽样（SRS）通常在大多数指标上取得最低值，其中简单随机抽样在自行车共享和家电能耗数据集上表现最佳，而 PCA-QS 在 CAPS 和在线新闻受欢迎度数据集上表现出色。每种距离度量中最小（最佳）的值以粗体标出。杠杆分数和核心集抽样方法的结果则有所不同按数据集来看。尽管未报告执行时间，简单随机抽样在四种方法中是最快的方法，而 PCA-QS 通常也比其余方法更高效。
请注意，对于本表所示结果，PCA-QS 使用动态数量的

主成分，每个数据集均保留不少于总方差的 70%。

表10：基于 1000 次重复运行，真实世界数据集中保留子集与原始数据之间分布相似性指标和运行时间的均值与标准差。

数据集	方法	能量距离	马哈拉诺比斯	KL	MMD
CAPS	PCA-QS	**0.013** (0.007)	**0.032** (0.0087)	0.157 (0.258)	**0.0004** (0.0003)
	杠杆	0.485 (0.102)	0.773 (0.0161)	1.666 (0.022)	2.437 (0.120)
	核心集	0.367 (0.083)	0.766 (≈ 0)	**1.402** (≈ 0)	1.897 (≈ 0)
	SRS	0.014 (0.008)	0.041 (0.0099)	0.164 (0.248)	0.0016 (0.0016)
在线	PCA-QS	1.428 (3.604)	**0.040** (0.006)	**0.001** (≈ 0)	8×10^{-6} (≈ 0)
	杠杆	1.233 (30.679)	0.203 (0.0366)	0.756 (0.016)	5.131 (0.101)
	核心集	240.79 (350.92)	0.153 (0.029)	0.655 (≈ 0)	5.403 (≈ 0)
	SRS	**0.016** (0.023)	**0.040** (0.006)	0.112 (0.011)	3×10^5 (2.276×10^5)
自行车共享	PCA-QS	0.023 (0.007)	0.096 (4.9×10^{-4})	**0.015** (2.8×10^{-4})	0.015 (1.4×10^{-4})
	杠杆	0.087 (0.020)	0.513 (0.031)	0.441 (0.036)	0.329 (0.040)
	核心集	0.036 (0.009)	0.274 (≈ 0)	0.213 (≈ 0).	0.087 (≈ 0)
	SRS	**0.020** (0.005)	**0.084** (0.016)	0.055 (0.025)	**0.007** (0.003)
器具	PCA-QS	0.0029 (0.0009)	0.106 (0.0155)	0.294 (0.0143)	≈ 0 (≈ 0)
	杠杆	0.0240 (0.0043)	0.276 (0.0189)	0.483 (0.0335)	≈ 0 (≈ 0)
	核心集	0.0085 (0.0007)	0.208 (0.0077)	0.382 (0.0146)	≈ 0 (≈ 0)
	SRS	0.0032 (0.0011)	0.108 (0.0152)	0.299 (0.0151)	≈ 0(≈ 0)

“在线新闻受欢迎度”和“家电能耗”分别指在线新闻受欢迎度和家电能耗数据集。自行车共享数据集分为两个子集——小时和天，这里给出的是按小时统计的结果。此表未列出执行时间。不过，正如预期，简单随机抽样所需时间少于其他所有方法，而 PCA-QS 通常比核心集和杠杆分数抽样方法耗时更少。请注意，这里的 PCA-QS 基于动态主成分（PC），因此各数据集所使用的 PC 数量有所不同。

在所有数据集中，与其他方法相比，PCA-QS 始终呈现较低的分布距离，这证实了它在大幅缩减数据的同时保留原始数据统计特征的能力。杠杆分数抽样在某些建模任务中表现出色，但往往会造成较大的分布失真，表现为距离更大。核心集抽样的均方误差较低，但在保留局部变化方面存在局限，这体现在能量距离和最大均值差异（MMD）分数偏高。SRS 提供了快速且无偏的基线方法，但未能利用结构信息，因此在各项指标上的表现中等。

这些结果进一步证明，PCA-QS 是一种有效的现实数据降维工具，适用于既需要保持分布相似性、又要求具备计算效率以支持下游分析的任务。

4 结论

本研究提出了 PCA-QS 框架，这是一种混合方法，旨在高效缩减数据，同时保留

原始数据集的基本分布特征。与传统的降维或随机抽样方法不同，PCA-QS 利用主成分分析揭示的结构来指导分层分位数抽样，确保保留的子集具有全局数据代表性。

针对合成高斯混合模型和各种基准数据集开展的大量数值实验表明，在保持分布相似性方面，PCA-QS始终优于简单随机抽样、杠杆分数抽样和核心集抽样等标准抽样技术。能量距离、马氏距离、Kullback–Leibler散度和最大均值差异等指标证实，与其他方法相比，PCA-QS能够使缩减后的数据集与原始数据集更加接近，同时保持合理的计算成本。根据多种分布相似性指标的测量结果，PCA-QS使保留集与原始数据集之间的失真度较低。

此外，整合A-最优、D-最优、G-最优和基于不确定性的选择等自适应抽样策略，表明PCA-QS在各种实验设计目标下依然稳健且灵活。这些扩展进一步提升了其生成信息丰富、具有代表性的子集的能力，并可根据不同分析任务进行定制；这些任务不仅包括回归，还包括聚类和探索性分析。

尽管PCA-QS在数据缩减与结构保真度之间实现了实用的平衡，但仍存在一些局限性。该方法依赖线性主成分分析，因此在处理高度非线性流形或具有强局部依赖关系的数据时，效果可能受限。此外，随着保留的主成分增多，分位数分层会导致组别组合数呈指数级增长，这可能增加超高维数据集的计算开销。未来研究可通过探索主成分分析步骤的非线性扩展、结合多个分层基准的混合抽样技术，或适用于海量数据集的可扩展分位数分配方法来应对这些挑战。

PCA-QS 提供了一个稳健且易于解释的框架，可生成能够忠实模拟原始数据分布的保留子集。因此，它是大规模数据缩减的宝贵工具，尤其是在后续统计建模、可视化或机器学习工作流中必须保持分布相似性的场景下。其自适应扩展进一步拓宽了适用范围，为未来开发高效且具有代表性的抽样方法提供了灵活的基础。

附录A 关于 PCA-QS、杠杆分数抽样、核心集抽样和简单随机抽样用法的更多细节

我们在下表中总结了 PCA-QS、简单随机抽样、杠杆分数抽样和核心集抽样的特征，并就其用法提供了一些实践建议。

A.1 PCA-QS

PCA-QS 展现出全面而出色的性能，在保留整体分布和确保模型误差较低之间实现了良好平衡。它的以下指标均处于最低水平之一：

能量距离、马氏距离、KL散度和最大均值差异，这表明它选出的子集在均值和方差结构上都与原始数据分布非常接近。它的均方误差略高于杠杆和核心集方法，但这一取舍带来了显著更好的分布保真度。由于包含主成分分析分解和分层抽样，其运行时间适中。

适用场景：当平衡保留全局数据结构与可解释性至关重要时，推荐使用此方法，例如用于探索性分析、可视化或通用数据缩减。

A.2 杠杆分数抽样

杠杆分数抽样在核心集之后取得了最低的均方误差，并通过高效的随机化奇异值分解保持了较低的计算成本。然而，它在分布相似性指标上的表现较差：该方法严重扭曲了数据分布（能量距离、马氏距离、KL散度和最大均值差异均较高）。这是因为它优先抽取高影响力的行，导致样本偏向极端点。

适用场景：最适合目标是维持线性模型的预测准确性，而非复现完整数据分布的情况。对于大数据集上的快速回归任务，这是一种务实的选择。

A.3 核心集抽样

核心集抽样通过K均值簇中心进行近似，在所有重复实验中均取得最低的均方误差且方差为零，凸显了其在概括簇均值方面的稳定性。然而，它会丢失簇内的局部变化，这反映在较高的能量距离、马氏距离、KL散度以及尤其高的最大均值差异上。由于采用迭代式聚类过程，其计算成本最高。

适用场景：当主要目标是近似表示数据，以用于聚类或基于质心的下游任务，且保留细粒度分布不那么重要时，这是理想选择。当准确性比精细的数据结构更重要，且运行时间可以接受时，此方法很有用。

A.4 简单随机抽样（SRS）

简单随机抽样（SRS）速度快且无偏，可作为基准方法，在所有指标上均能取得合理表现。其均方误差（MSE）略高于杠杆抽样和核心集抽样，但在能量距离、马哈拉诺比斯距离和 KL 散度方面表现尚可，最大均值差异（ MMD ）则处于中等水平。由于只需随机选择索引，其计算时间最短。适用场景：对于小型到中型数据集，或无法获取模型结构信息时，SRS 是一种实用的基准方法。当计算简便性至关重要时，它是一个稳妥的选择。

A.5 抽样方法比较：核心集、杠杆分数和 PCA-QS

核心集抽样。
核心集抽样旨在构造一个较小的加权数据子集，以保留原始数据集的几何性质和统计性质Feldman 和 Langberg(2011)；Bachem 等人(2018）。一种常见策略是将数据聚类为 k个簇，并从每个簇中选取代表性样本。这种方法对样本量较大的情况尤其有效（n），因为其计算复杂度随 n 和特征维度 d线性增长，运行时间约为 $\mathcal{O}(n \cdot d \cdot k)$。然而，在维数极高的空间中，维数灾难会使距离的概念变得不那么有意义，因此核心集可能难以发挥作用Aggarwal 等人(2001）。此外，当样本量 n 较小时，构建有意义的簇可能会很困难，因为每个簇可能仅包含一个或极少数数据点。由于聚类能够抑制孤立点的影响，核心集抽样通常对离群值具有较强的鲁棒性。

杠杆分数抽样。
杠杆分数抽样是一种概率技术，根据数据矩阵各行对线性模型的影响进行抽样（Drineas 等人，2012；Mahoney，2011b）。数据点的杠杆分数衡量其对最小二乘拟合的贡献，计算时依据数据矩阵顶端左奇异向量各行的平方范数。具体而言，对于一个奇异值分解为 $X = U\Sigma V^\top$ 的数据矩阵 $X \in \mathbb{R}^{n\times d}$，第 i 行的杠杆分数为 $\|\mathbf{u}_i\|^2$，其中 $\mathbf{u}_i$ 表示 U 的第 i 行。杠杆采样在理论上对线性回归和低秩近似是最优的（Drineas 等人，2006）。即使在高维情况下，它也能表现良好，因为它能够捕捉主要方差方向，而非单纯依赖距离。不过，计算杠杆分数的开销可能很大：精确奇异值分解需要 $\mathcal{O}(n \cdot d^2)$ 时间，随机化方法则需要 $\mathcal{O}(n \cdot d \cdot k)$（Halko 等人，2011）。此外，杠杆采样可能对离群值较为敏感，因为离群点的杠杆分数往往会高得不成比例。

PCA-QS 抽样。
PCA-QS 是一种混合方法，将降维与沿主成分进行的分层抽样相结合。首先，主成分分析将数据从 d 维降至 k 个主成分，同时保留大部分方差 Jolliffe (2002)。随后，计算每个选定主成分上的分位数，将低维空间划分为多个区域。接着从每个分位区域中选取点进行抽样，确保广泛覆盖数据变异性。PCA-QS 擅长保持主要方差方向上的多样性，在纯几何抽样方法与纯概率抽样方法之间实现平衡。其运行时间主要由主成分分析计算决定；对于随机化算法，计算成本约为 $\mathcal{O}(n \cdot d \cdot k)$Halko 等人 (2011)。然而，PCA-QS 可能对所选主成分数量和分位数阈值较为敏感。与杠杆采样类似，当样本量非常小时，它也可能表现不佳，因为数据不足可能导致主成分分析不稳定。

样本量和维数的影响。
这些方法的性能会随样本量 n 和维数 d而显著变化。核心集抽样在 n 较大时扩展性良好，但在高d情况下表现不佳。杠杆采样在高维空间中仍然有效，但计算成本很高。PCA-QS 则提供了一种折中方案，既能有效覆盖数据分布，又能减轻完整集杠杆分数计算带来的部分计算负担。当 n较小时，杠杆采样和 PCA-QS 仍可能适用，而核心集可能无法形成有意义的划分。

在实际应用中，核心集更适用于纯几何或聚类任务；杠杆采样最适合回归任务；而 PCA-QS 则是一种颇具吸引力的折中方案，既能在数据表示中保持多样性，又能控制计算成本。表 A1 总结了它们各自的优势与局限。

表 A1：核心集抽样、杠杆分数抽样与 PCA-QS 的比较

	核心集	**杠杆分数**	**PCA-QS**
大型 n	可高效扩展，效果良好 聚类	成本高，但效果显著	可随 n良好扩展
小型 n	簇变得不可靠 靠	仍然 有效 适用于 回归	如果样本过少则不稳定 few 样本
高 d	聚类 不可靠 由于维度灾难而—— 维度	仍然有效，但 代价高昂	有效地 降低 维度
低 d	表现良好	计算快速	快速且稳定
离群值敏感性-ity	总体稳健	敏感 to 高 杠杆离群值	敏感 if 离群值 占主导地位 主 主成分

参考文献

AlKarawi, A., AlJanabi, K.：数据缩减技术：一项比较研究。库法数学与计算机杂志 **9**(2), 1–17 (2022)https://doi.org/10.31642/JoKMC/2018/090201

Ester, M.,Kriegel,H.P.,Sander, J.,Xu, X.：一种基于密度的算法，用于在大型空间数据库中发现带有噪声的簇，页码 226–231 (1996)。AAAI出版社

Ai, R., Boyle, D.L., Wang, W.：类风湿关节炎外周血和滑膜组织T细胞具有不同的DNA甲基化模式。ACR开放风湿病学**3**，127–132（2021）https://doi.org/10.1002/acr2.11231

Arbelaitz, O.,Gurrutxaga, I., Muguerza, J., P´erez, J.M., Perona, I.：聚类有效性指标的广泛比较研究。《模式识别》**46**（1），243–256（2013）Aggarwal, C.C., Hinneburg, A., Keim, D.A.：高维空间中距离指标的惊人表现。载于：第8届国际数据库理论会议（ICDT）论文集，第420–434页。施普林格，???（2001）Abdi, H., Williams, L.J.：主成分分析。《威利跨学科评论：计算统计学》**2**（4），433–459（2010）Abdi, H., Williams, L.J.：主成分分析。《威利跨学科评论：计算统计学》**2**（4），433–459（2010）Belkin, M.，等：用于嵌入和聚类的拉普拉斯特征映射与谱技术。载于：《神经信息处理系统进展》（NeurIPS）（2001）Baglama, J., Ch´avez-Casillas, J.A., Perovi´c, V.：一种满足给定阈值的部分奇异值分解混合算法。《数值算法》（2024）https://doi.org/10.1007/s11075-024-01906-9Baik, J., Deift, P., Johansson, K.：随机排列最长递增子序列长度的分布。《美国数学学会杂志》**12**（4），1119–1178（1999）https://doi.org/10.1090/S0894-0347-99-00307-0Bachem, O., Lucic, M., Krause, A.：通过轻量级核心集实现可扩展的分布式聚类。载于：第24届ACM SIGKDD国际知识发现与数据挖掘会议（KDD）论文集，第1119–1127页。ACM，???（2018）Baglama, J., Lewis, M., Reichel, L.：《irlba》软件包。美国罗德岛大学（URI）和肯特州立大学，美国，（2019）。美国罗德岛大学（URI）和肯特州立大学，美国。取自uni.wroc.pl。http://r.meteo.uni.wroc.pl/web/packages/irlba/irlba.pdfBecht, E., McInnes, L., Healy, J., Dutertre, C.A.：使用UMAP对单细胞数据进行可视化降维。《自然·生物技术》（2019）https://doi.org/10.1038/nbt.4314Baglama, J., Reichel, L.：增广隐式重启兰乔斯双对角化方法。《SIAM科学计算杂志》**27**（1），19–42（2005）Craig, A.T., Harold, W.：能级间距分布与艾里核。《数学物理学通讯》**159**（1），151–174（1994）https://doi.org/10.1007/BF02100489

Craig, A.T., Harold, W.：关于正交矩阵系综与辛矩阵系综。数学物理通讯**177**(3), 727–754 (1996) https://doi.org/10.1007/BF02099545Chang, Y.-c.I., Lee, Y.-J., Pao, H.-K., Lee, M.-H., Huang, S.-Y.：通过核机器进行数据可视化。载于：Chen, C.-h., H¨ardle, W., Unwin, A.（编）《计算统计学手册》（第三卷）——数据可视化。施普林格，???（2008）。发表于2006年数据与信息可视化研讨会，柏林 Cox, D.R., Reid, N.：实验规划。《皇家统计学会杂志》B辑**20**(1), 1–32 (1958) Chaloner,K., Verdinelli, I.：贝叶斯实验设计：综述。《统计科学》**10**(3), 273–304 (1995) Davies, D.L., Bouldin, D.W.：一种聚类分离度量。《IEEE模式分析与机器智能汇刊》**1**(2), 224–227 (1979) Ding, C., He, X.：通过主成分分析进行K均值聚类。载于：第21届国际机器学习会议（ICML）论文集，第29页（2004）Drineas, P., Kannan, R., Mahoney, M.W.：矩阵快速蒙特卡洛算法 II：计算矩阵的低秩近似。《SIAM计算杂志》**36**(1), 158–183 (2006) Drineas, P., Magdon-Ismail, M., Mahoney, M.W., Woodruff, D.P.：矩阵相干性和统计杠杆值的快速近似。《机器学习研究杂志》**13**, 3475–3506 (2012) Drineas, P., Mahoney, M.W., Muthukrishnan, S.：L2回归的采样算法及其应用。载于：第十七届ACM-SIAM离散算法年度研讨会论文集，第1127–1136页（2006）。工业与应用数学学会 EasyBA.co：数据缩减：数据分析解读。EasyBA.co（2023）。访问日期：2024-11-28 Erokhin, S.D., Borisenko, B., Martishin, I.D., Fadeev, A.：输入数据维数缩减现有方法分析。T-comm（2022）https://doi.org/10.36724/2072-8735-2022-16-1-30-37Erichson, N.B., Voronin, S., Brunton, S.L.：使用 R 进行随机矩阵分解。arXiv预印本（2016）arXiv:1608.02148。可在 arXiv 获取 Fedorov, V.V.：最优实验理论。《统计科学》**10**(3), 1–400 (1972) Feldman, D., Langberg, M.：用于数据近似与聚类的统一框架。载于：第四十三届ACM理论年度研讨会论文集

计算，569–578 页（2011）。ACM

Forrester, P.J.：《对数气体与随机矩阵》。伦敦数学学会专著，第34卷。普林斯顿大学出版社，美国新泽西州普林斯顿（2010）。https://doi.org/10.1515/9781400835416

Galton, F.：百分位数与分位数概念。《自然》**9**，207（1874）Grosset, P., Biwer, C.M., Pulido, J., Mohan, A.T., Biswas, A., Patchett, J., Turton, T.L., Rogers, D.H., Livescu, D., Ahrens, J.：Foresight：对数据缩减至关重要的分析。载于：SC20：高性能计算、网络、存储与分析国际会议，第1–15页（2020）。https://doi.org/10.1109/SC41405.2020.00087Hamerly, G., Elkan, C.：学习 k 均值中的 k。载于：第17届神经信息处理系统年会论文集（NIPS），第281–288页（2003）Hastie, R. T. Tibshirani, Friedman, J.：《统计学习要素》。施普林格，纽约，纽约（2009）Halko, N., Martinsson, P.-G., Tropp, J.A.：利用随机性发现结构：用于构造近似矩阵分解的概率算法。《SIAM评论》**53**(2)，217–288（2011）Hotelling, H.：将一组统计变量分解为主成分。《教育心理学杂志》**24**(6)，417–441（1933）Har-Peled, S., Mazumdar, S.：关于 k 均值与 k 中位数聚类的核心集。载于：第三十六届 ACM计算理论研讨会年会论文集，第291–300页（2004）。ACM Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.：《稳健统计：基于影响函数的方法》。威利，纽约（1986）Jackson, J.E.：《主成分用户指南》。Wiley-Interscience，纽约（1991）Jain, A.K.：数据聚类：超越 k 均值的50年。模式识别快报**31**(8)，651–666（2010）Johnstone, I.M.：主成分分析中最大特征值的分布。《统计年鉴》**29**(2)，295–327（2001）https://doi.org/10.1214/aos/1009210544Jolliffe, I.T.：《主成分分析》，第2版。施普林格，纽约（2002）Koenker, R., Bassett, G.：回归分位数。《计量经济学：计量经济学会期刊》**46**(1)，33–50（1978）

Lewis, B.W., Baglama, J.：Irlba 软件包。（2021）。可从 CRAN 获取。
https://cran.uib.no/web/packages/irlba/vignettes/irlba.pdfLittle, R.J.A., Rubin, D.B.：《缺失数据的统计分析》。约翰·威立父子出版公司，纽约，纽约（2002）Linderman, J., Steinerberger, S.：t-SNE 嵌入的清晰性：成对吸引力与排斥力。技术报告，耶鲁大学（2017）Li, Y., Yue, S., Zhang, Y.：用于高维数据特征提取的核主成分分析。《机器学习应用期刊》**12**(4)，23–34（2009）Liu, J., Zhang, Z., Metaxas, D.N.：并行与分布式主成分分析算法：综述。《IEEE 神经网络与学习系统汇刊》**29**(1)，215–232（2017）Mika, S., 等：特征空间中的核主成分分析与去噪。载于：《神经信息处理系统进展》（NeurIPS）（1999）MacKay, D.J.：用于主动数据选择的基于信息的目标函数。《神经计算》**4**(4)，590–604（1992）Mahoney, M.W.：用于矩阵与数据的随机化算法。《机器学习基础与趋势》®《机器学习基础与趋势》**3**(2)，123–224（2011）Mahoney, M.W.：用于矩阵与数据的随机化算法。《机器学习基础与趋势》**3**(2)，123–224（2011）Mehta, M.L.：《随机矩阵》，第3版。爱思唯尔，阿姆斯特丹（2004）。
https://doi.org/10.1016/B978-012088409-4/50001-2McInnes, L., Healy, J., Melville, J.：UMAP：用于降维的统一流形近似与投影。技术报告，Tutte 数学与计算研究所（TIMC），加拿大（2018）Marchenko, V.A., Pastur, L.A.：若干随机矩阵集合的特征值分布。《苏联数学汇刊》**1**(4)，457–483（1967）
https://doi.org/10.1070/SM1967v001n04ABEH001994Martinsson, P.-G., Tropp, J.A.：随机数值线性代数：基础与算法。《数值数学纪事》**29**，403–572（2020）Provost, F., Fawcett,T., Kohavi, R.：反对使用准确率估计比较归纳算法。载于：第十五届国际机器学习会议论文集（1998）Pukelsheim, F.：《试验的最优设计》第50卷。美国工业与应用数学学会（工业与应用数学学会

与应用数学），???（2006）

Ringn´er，M.：什么是主成分分析？《自然·生物技术》**26**（3），303–304（2008）Rousseeuw，P.J.、Leroy，A.M.：《稳健回归与异常值检测》。约翰·威立父子出版公司，纽约，纽约（2005）Rousseeuw，P.J.：轮廓系数：用于解释和验证聚类分析的图形辅助工具。《计算与应用数学杂志》**20**，53–65（1987）Shao，J.：抽样调查中的重抽样方法。《统计科学》**18**（2），206–231（2003）Shlens，J.：主成分分析教程。技术报告，印度理工学院卡拉格普尔分校（IIT Kharagpur），印度（2014）Sarwar，B.、Karypis，G.、Konstan，J.、Riedl，J.：降维在推荐系统中的应用：案例研究。载于：ACM WebKDD研讨会论文集。ACM，???（2000）Sch¨olkopf，B.、Smola，A.、M¨uller，K.-R.：作为核特征值问题的非线性成分分析。《神经计算》**10**（5），1299–1319（1998）Turk，M.、Pentland，A.：用于识别的特征脸。《认知神经科学杂志》**3**（1），71–86（1991）Taguchi，Y.、Turki，T.：单细胞多组学数据分析中的基于张量分解的无监督特征提取。《基因》**12**（9），1442（2021）https://doi.org/10.3390/genes12091442Tibshirani，R.、Walther，G.、Hastie，T.：通过间隙统计量估计数据集中的聚类数。《皇家统计学会杂志：B辑（统计方法学）》**63**（2），411–423（2001）Vendramin，L.、Campello，R.J.、Hruschka，E.R.：相对聚类有效性准则：比较综述。《统计分析与数据挖掘：美国统计学会数据科学期刊》**3**（4），209–235（2010）Maaten，L.、Hinton，G.：使用t-SNE可视化数据。《机器学习研究杂志》**9**，2579–2605（2008）Vahdat，A.、Kautz，J.：NVAE：一种深层分层变分自编码器。载于：NeurIPS会议（2020）