Compressive K-Means

Geometric Data Analysis

Ben Kabongo

December 15, 2023

Master MVA, ENS Paris-Saclay

Table of contents

Clustering and K-Means

K-Means

Clustering : partition a data set into homogeneous and disjoint subsets called *clusters*

K-Means [4] : a clustering algorithm that calculates cluster centers with the aim of minimizing the *sum of squared errors* (SSE)

SSE : Sum of squared errors

$$\mathrm{SSE}(X, C) = \sum_{i=1}^{n} \min_{k} \|x_i - c_k\|^2 \tag{1}$$

K-Means Algorithm

Algorithm 1 K-Means Algorithm

Input: Data set X, number of clusters k
Output: Cluster centroids $C_1, C_2, \ldots, C_k$
Initialize cluster centroids $C_1, C_2, \ldots, C_k$ randomly from X
repeat
    for each data point $x_i \in X$ **do**
        Assign x_i to the nearest centroid: $c_i = \arg\min_j \|x_i - C_j\|^2$
    end for
    for each cluster C_j **do**
        Update centroid C_j: $C_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$
    end for
until Convergence

Examples : Datasets

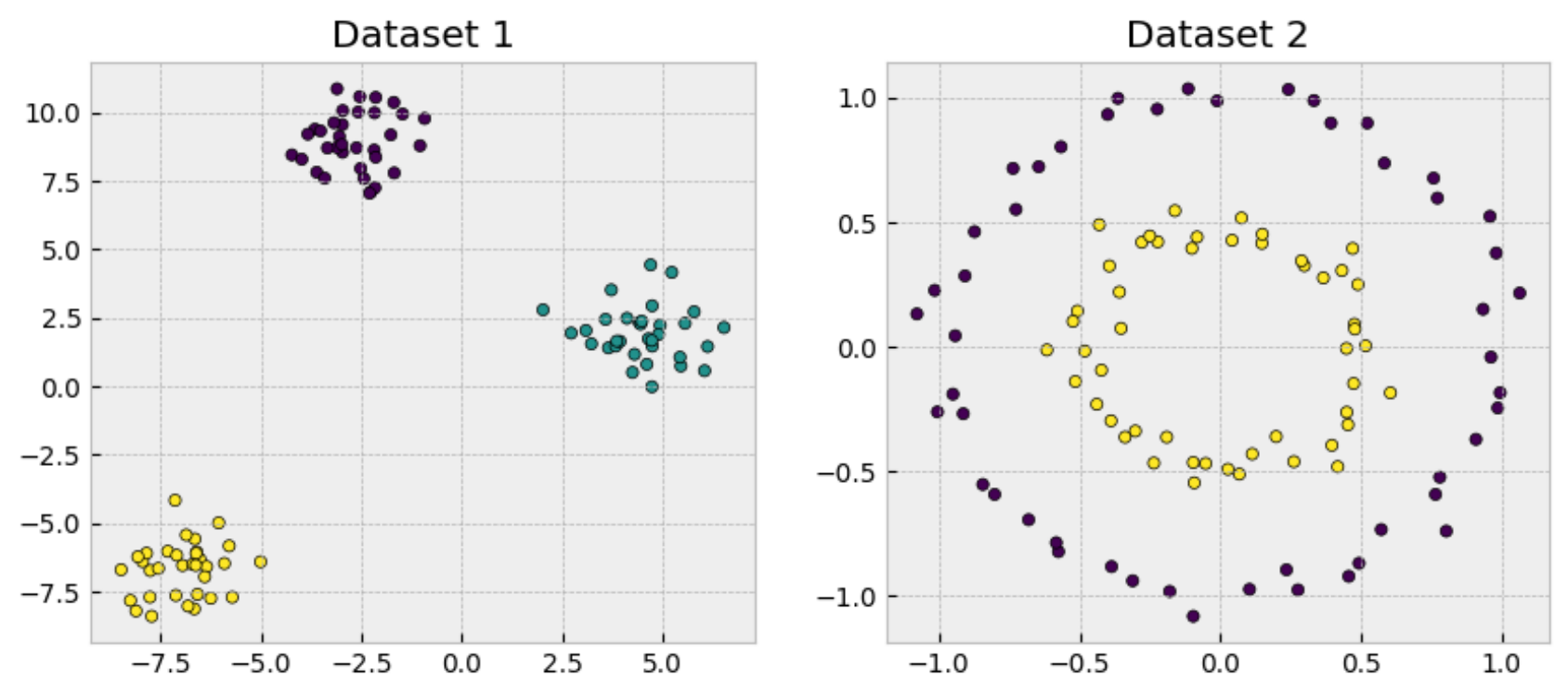

Figure 1: Toy datasets. The dataset on the left has three correctly disjointed clusters, and the one on the right has two with one nested within the other.

Examples : Clustering with K-Means

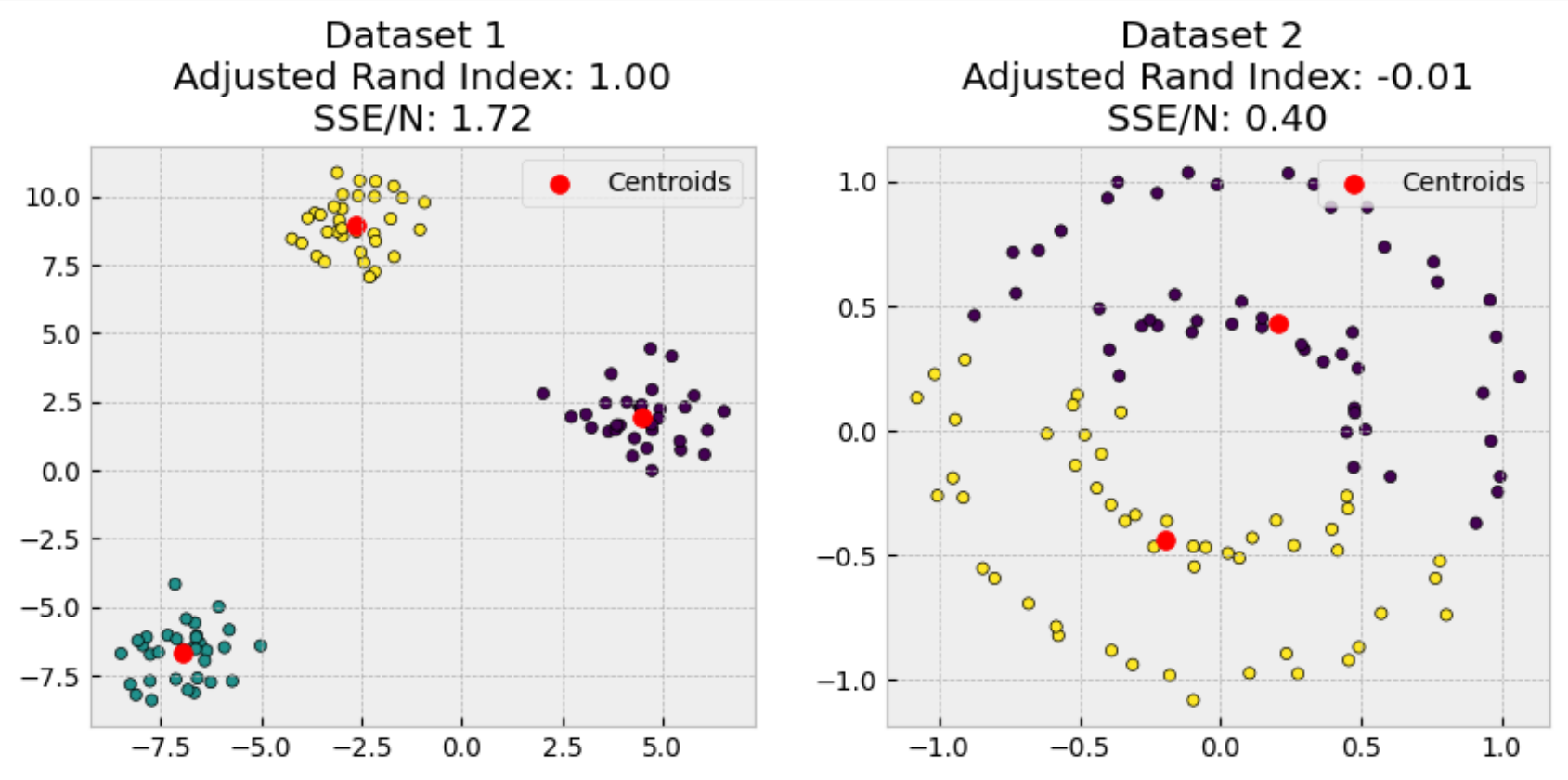

Figure 2: K-Means on toy datasets. The colors of the examples correspond to their K-Means cluster. Cluster centers are shown in red.

Compressive K-Means (CKM)

Overview

- **CKM** [3] is an algorithm offering a compressive approach to K-Means.
- The idea is to build a compressive representation of the data, called a **sketch**, independent of the number of examples and the size of the data.
- The sketch is then used to find the cluster centers using a *Compressive Learning Orthogonal Matching Pursuit* (CLOMPR) algorithm [2], adapted for K-Means.

Compressive sensing

- Initially developed and successfully applied to sparse signals of finite dimension
- Allows the recovery of such signals from a number of measurements significantly smaller than the total dimension of the space
- Extended to more general classes of signals: *low-rank matrices* and *functions*
- **Traditional compressive sensing**:
  Dataset : $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, with $\mathbf{x}_i \in \mathbb{R}^n$
  Projection matrice : $\mathbf{M} \in \mathbb{R}^{m \times n}$

$$\mathbf{y} \approx \mathbf{M}\mathbf{x} \tag{2}$$

Sketching operator

Let be :

- $\mathbf{Y} = \{\mathbf{y}_l\}_{l=1}^{L}$, a dataset with $\mathbf{y}_l \in \mathbb{R}^n$
- $\beta = \{\beta_l\}_{l=1}^{L}$: L weights in $\mathbb{R}$
- $\mathbf{\Omega} = \{\mathbf{w}_j\}_{j=1}^{m}$: m frequencies in $\mathbb{R}^n$, learned on a portion of the dataset as indicated in another paper by the authors [2].

The sketching operator for the CKM algorithm is given by :

$$\mathrm{Sk}(\mathbf{Y}, \beta) = \left[\sum_{l=1}^{L} \beta_l e^{-i\mathbf{w}_j^T \mathbf{y}_l} \right]_{j=1}^{m} \in \mathbb{C}^m \tag{3}$$

Sketching operator

The sketching operator produces a sketch of size m, independent of the dataset size.

This sketching operator can be expressed as a linear $\mathbf{A}$ operator with respect to probability distributions.

$$\mathbf{A}p_{\mathbf{Y},\beta} = \left[\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{Y},\beta}} e^{-i\mathbf{w}_j^T\mathbf{y}}\right]_{j=1}^{m} \in \mathbb{C}^m \tag{4}$$

We therefore note that $p_{\mathbf{Y},\beta} = \sum_{l=1}^{L} \beta_l \delta_{y_l}$ and that when the weight vector is only indicated it is a uniform weight distribution $p_{\mathbf{Y}} = \frac{1}{L}\sum_{l=1}^{L} \delta_{y_l}$.

Compressive K-Means

Let be a dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, with $\mathbf{x}_i \in \mathbb{R}^n$. We want to find the set $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K}$ of K centroids, with $\mathbf{c}_k \in \mathbb{R}^n$.

CKM proposes a new formulation of K-Means given by :

$$\text{argmin}_{\mathbf{C},\alpha} \|\text{Sk}(\mathbf{X}, \frac{\mathbf{1}}{N}) - \text{Sk}(\mathbf{C}, \alpha)\|_2^2 \tag{5}$$

with $\alpha \geq 0$ and $\sum_{k=1}^{K} \alpha_k = 1$. The problem can also be formulated as :

$$\text{argmin}_{\mathbf{C},\alpha} \|\mathbf{z} - \mathbf{A}p_{\mathbf{C},\alpha}\|_2^2 \tag{6}$$

with $\mathbf{z} = \mathbf{A}p_{\mathbf{X}}$.

Compressive K-Means Algorithm

Algorithm 2 CKM (CLOMPR for K-Means) Algorithm

Input: Sketch $\mathbf{z}$, frequencies $\mathbf{\Omega}$, number of clusters k, centroids bounds $\mathbf{l}$, $\mathbf{u}$
Output: Centroids $\mathbf{C}$, weights α
$\mathbf{r} = \mathbf{z}$; $\mathbf{C} = \emptyset$
for $t \leftarrow 1$ to $2K$ **do**
    Step 1 : Find a new centroid $\mathbf{c}$
    Step 2 : Expand support $\mathbf{C} = \mathbf{C} \cup \{\mathbf{c}\}$
    Step 3 : Enforce sparsity by Hard Thresholding if $t > K$
    Step 4 : Project to find α
    Step 5 : Global gradient descent $\mathbf{C}, \alpha$
    Update residual: $\mathbf{r} \leftarrow \mathbf{z} - \sum_{k=1}^{|\mathbf{C}|} \alpha_k \mathbf{A}\delta_{\mathbf{c_k}}$
end for

Compressive K-Means Algorithm

Step 1 : Find a new centroid $\mathbf{c} \leftarrow \text{maximize}_{\mathbf{c}} \left(\text{Re}\langle \frac{\mathbf{A}\delta_{\mathbf{c}}}{\|\mathbf{A}\delta_{\mathbf{c}}\|}, \mathbf{r} \rangle, \mathbf{l}, \mathbf{u} \right)$

Step 2 : Expand support $\mathbf{C} = \mathbf{C} \cup \{\mathbf{c}\}$

Step 3 : Enforce sparsity by Hard Thresholding if $t > K$

- $\beta \leftarrow \text{argmin}_{\beta \geq 0} \left\| \mathbf{z} - \sum_{k=1}^{|\mathbf{C}|} \beta_k \frac{\mathbf{A}\delta_{\mathbf{c_k}}}{\|\mathbf{A}\delta_{\mathbf{c_k}}\|} \right\|$
- Select K largest entries $\beta_1, \beta_2 \ldots \beta_K$
- Reduce the support $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K}$

Step 4 : Project to find α

$$\alpha \leftarrow \text{argmin}_{\alpha \geq 0} \left\| \mathbf{z} - \sum_{k=1}^{|\mathbf{C}|} \alpha_k \mathbf{A}\delta_{\mathbf{c_k}} \right\|$$

Step 5 : Global gradient descent

$$\mathbf{C}, \alpha \leftarrow \text{minimize}_{\mathbf{C},\alpha} \left(\left\| \mathbf{z} - \sum_{k=1}^{|\mathbf{C}|} \alpha_k \mathbf{A}\delta_{\mathbf{c_k}} \right\|, \mathbf{l}, \mathbf{u} \right)$$

Evaluation metrics

Evaluation metrics

ARI (Adjusted Rand Index)

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{N}{2}}}$$

AMI (Adjusted Mutual Information)

$$\text{AMI} = \frac{I(Y;Z) - \mathbb{E}[I(Y;Z)]}{\max\{H(Y), H(Z)\} - \mathbb{E}[I(Y;Z)]}$$

Evaluation metrics

Homogeneity

$$\text{Homogeneity} = 1 - \frac{H(C|K)}{H(C)}$$

Completeness

$$\text{Completeness} = 1 - \frac{H(K|C)}{H(K)}$$

V-Measure

$$\text{V-Measure} = 2 \times \frac{\text{Homogeneity} \times \text{Completeness}}{\text{Homogeneity} + \text{Completeness}}$$

Experimentations

Preliminary 2D tests

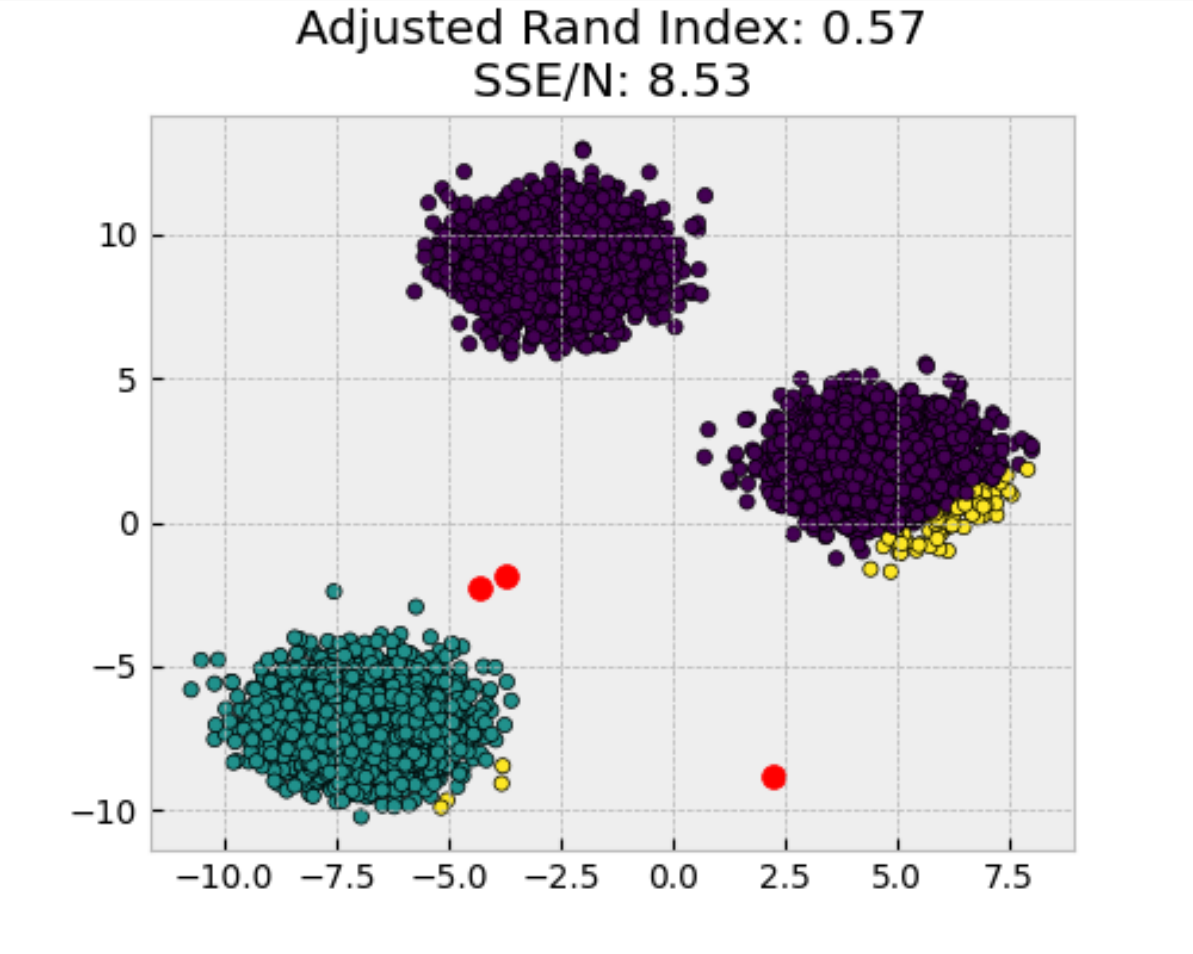

Figure 3: CKM clustering of 2D data, minimizing SSE.

Preliminary 2D tests

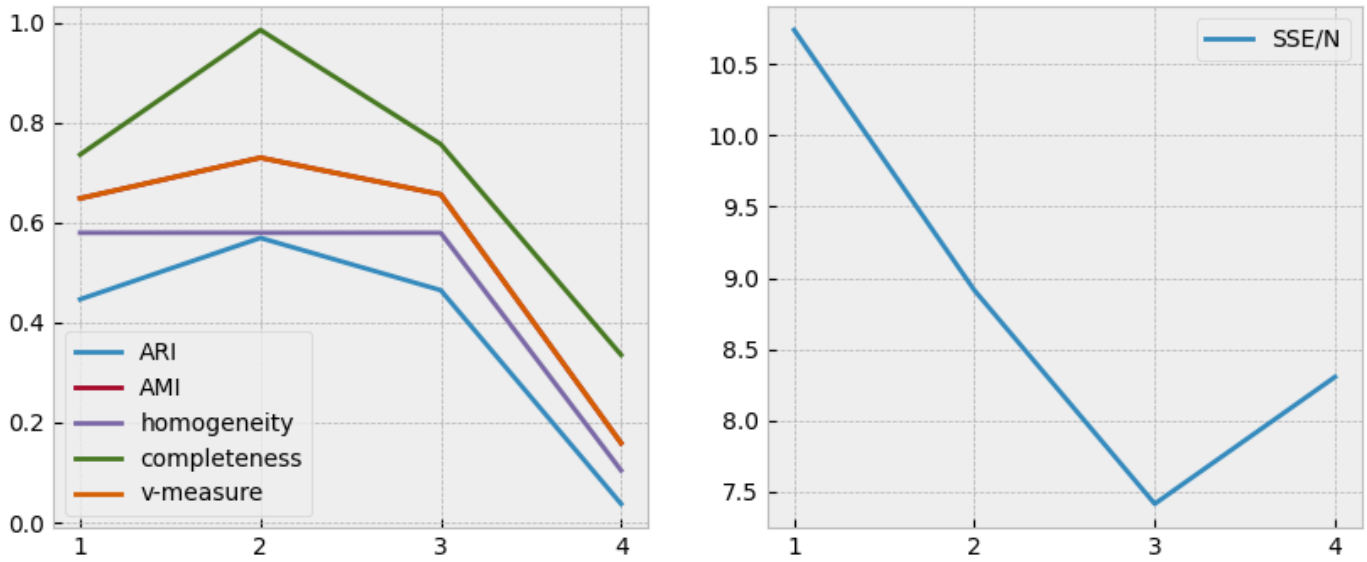

Figure 4: Evaluation of CKM on 2D data as a function of sketch size (KN).

Initalisation strategies

- **Random/Range**: random initialization of centroids. In the case of CKM, each centroid is drawn uniformly, taking into account the bounds passed to the algorithm.
- **Sample**: initialize centriodes by randomly selecting K examples from the dataset.
- **K-means++**: initialization given by K-Means ++ algorithm [1]. For CKM, select $c = x_i$ from the data with a probability inversely proportional to its distance to the current set of centroids.

Clustering of digits dataset

As part of this experiment, we recovered datasets of digit images. Our experiments were based on the digits dataset $N = 1797$, $K = 10$ clusters, $n = 8x8 = 64$ features.

In these experiments, we attempt to compare the K-Means and CKM algorithms in terms of different metrics and different types of initialization.

Results

Metric	K-Means	CKM
ARI		
RANGE	**0.605** (0.054)	0.084 (0.040)
SAMPLE	**0.609** (0.048)	0.298 (0.061)
KPP	**0.642** (0.045)	0.288 (0.058)
AMI		
RANGE	**0.718** (0.025)	0.194 (0.053)
SAMPLE	**0.720** (0.024)	0.442 (0.049)
KPP	**0.732** (0.021)	0.435 (0.050)

Table 1: Results of the ARI and AMI evaluation metrics for clustering with K-Means and CKM on the digits dataset. Values are presented with mean and standard deviation in parentheses.

Results

Metric	K-Means	CKM
Homogeneity		
RANGE	**0.709** (0.029)	0.160 (0.052)
SAMPLE	**0.711** (0.027)	0.434 (0.051)
KPP	**0.728** (0.024)	0.427 (0.051)
Completeness		
RANGE	**0.733** (0.021)	0.284 (0.050)
SAMPLE	**0.735** (0.021)	0.464 (0.046)
KPP	**0.743** (0.018)	0.456 (0.048)

Table 2: Results of the Homogeneity and Completeness evaluation metrics for clustering with K-Means and CKM on the digits dataset. Values are presented with mean and standard deviation in parentheses.

Results

Metric	K-Means	CKM
V-measure		
RANGE	**0.721** (0.025)	0.202 (0.053)
SAMPLE	**0.723** (0.024)	0.448 (0.048)
KPP	**0.735** (0.021)	0.441 (0.049)
SSE/N		
RANGE	**25.099** (0.214)	59.227 (1.494)
SAMPLE	**25.132** (0.238)	35.637 (0.938)
KPP	**25.044** (0.168)	35.664 (1.014)

Table 3: Results of the V-measure and SSE/N evaluation metrics for clustering with K-Means and CKM on the digits dataset. Values are presented with mean and standard deviation in parentheses.

Conclusion

Conclusion

CKM is an algorithm that attempts to solve the problem of the complexity of the K-Means algorithm for large data.

However, based on the experiments we have carried out, the K-Means algorithm presents the best results than the CKM algorithm.

The results of the CKM algorithm are not far from the K-Means algorithm. Perhaps an optimal calculation of the sketch would allow better results to be obtained.

References i


D. Arthur and S. Vassilvitskii.
k-means++: The advantages of careful seeding.
In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.


N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez.
Sketching for large-scale learning of mixture models.
arXiv preprint arXiv:1606.02838, pages 1–50, 2016.


N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval.
Compressive k-means.
arXiv preprint arXiv:1610.08738, 2016.


S. P. Lloyd.
Least squares quantization in pcm.
IEEE Transactions on Information Theory, 28(2):129–137, 1982.