On K-means and PCA

Chloé Habasque [1] Colin Coërchon [1] Maëlle Fontaine [1]

[1]Master MVA, ENS Paris-Saclay

Principal Component Analysis (PCA) and K-means clustering are closely related unsupervised learning techniques. PCA reduces dimensionality by identifying principal components, while K-means clusters data points, and their synergy can enable innovative applications like improved algorithm initialization.

PCA as a Continuous Solution to K-means Clustering

In [1], a theoretical link between K-means and PCA is established, showing that **principal components provide a continuous solution for cluster membership indicators in K-means**.

> The foundation of the proof relies on rewriting the within-cluster inertia J_K, which K-means aims to minimize:
>
> $$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - m_k\|^2$$
>
> For $K = 2$: $J_K = n\bar{y}^2 - \frac{J_D}{2}$ For $K > 2$: $J_K = \mathrm{Tr}(X^T X) - \mathrm{Tr}(H_K^T X^T X H_K)$
>
> This reformulation reveals the link between PCA and K-means. To illustrate, using the eigenvectors from PCA to determine clusters is straightforward for $K = 2$. The clusters C_1, C_2 are given by:
>
> $$C_1 = \{i \mid v_1(i) \leq 0\} \quad \text{and} \quad C_2 = \{i \mid v_1(i) > 0\}$$

This insight can enhance K-means results, as PCA provides a relaxed solution to the K-means clustering problem. It enables **spectral methods inspired by PCA to approximate clusters before applying K-means**.

PCA-guided search for K-means

In K-means algorithm, random initialization carries the risk of converging to a local minimum that is far from the optimal solution. PCA-guided search aims to address this issue. It follows three steps:

1. **Dimensionality Reduction**: Reduce the data to a number of dimensions equal to the number of clusters using PCA.
2. **Clustering in Reduced Space**: Perform K-means in the reduced space, where the risk of falling into a local minima is lower.
3. **Centroid Projection**: Project the centroids back to the original space to serve as the initial centroids for a final K-means.

This technique accelerates execution and ensures the centroids start closer to the optimal solution in the original space.

Limitations

K-means has significant limitations due to its rigid assumptions: each cluster shares an identical covariance which reduces its flexibility, especially in lower-dimensional spaces.

While PCA-guided search improves clustering performance, it remains constrained by K-means random initialization.

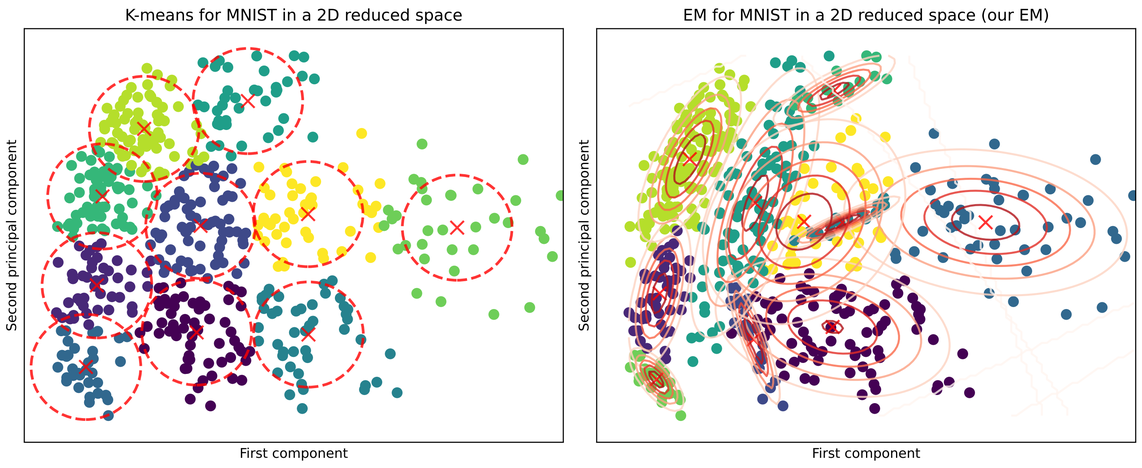
GMM versus K-Means Clustering

A proposed alternative involves replacing this step with a Gaussian Mixture Model (GMM) using the Expectation-Maximization (EM) algorithm. Unlike K-means, GMM allows for cluster-specific covariance structures and probabilistic point assignments, providing greater flexibility and a better alignment with the underlying data distribution.

GMM models observed data X_i as Gaussian distributions conditioned on cluster membership $(X_i \mid Z_{ik} = 1) \sim \mathcal{N}(\mu_k, \Sigma_k)$. Here, the EM algorithm optimizes the parameters $\theta = \{\alpha_j, \mu_j, \Sigma_j \mid j \in [\![1, m]\!]\}$. At each step of the algorithm, the parameters θ are updated as follows:

$$\theta^{(q+1)} = \arg\max_{\theta} Q(\theta, \theta^{(q)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^{(q)} \log\left(\alpha_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)\right)$$

Where $w_{ij}^{(q)}$ are the **responsibility coefficients**, representing the posterior probability that data point x_i belongs to cluster j at iteration q.



The Importance of Initialization Methods for Clustering

We experiment with different initialization methods. Two of them directly initialize K-means in the full space (K-means++ and Random). The other methods utilize the PCA-guided search approach and are based on different initialization strategies applied in the reduced space:

- **Random**: K points are chosen randomly as centroids.
- **K-Means++**: The first centroid is chosen randomly, and subsequent centroids are selected iteratively with a probability proportional to their distance from the nearest centroid.
- **KKZ**: The first centroid is the point with the maximum norm, and the others are selected iteratively as the furthest points from the existing centroids.
- **GMM**: Points are assigned to clusters based on the assignment probabilities computed by the Expectation-Maximization (EM) algorithm.

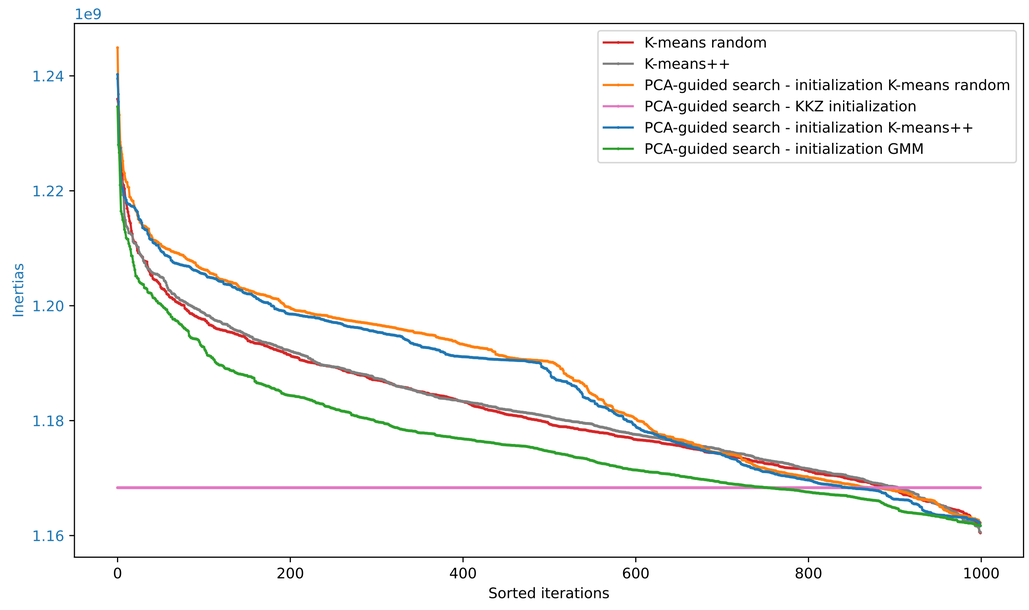


Figure 1. Ordered Inertias Across 1000 Runs for Various K-means Initialization Methods.

The GMM-based approach outperforms on average the alternatives in terms of inertia, even if when focusing on the best result across all iterations, all random-based algorithms achieve a similar minimal inertia. This can be attributed to the relatively small size of the dataset: over 1000 runs, randomness allows convergence to the same optimal solution at least once.

Model	Min Inertia (10e9)	Execution time (s)
PCA-guided Search with Random	1.1622	1.59
PCA-guided Search with KKZ	1.1683	2.27
PCA-guided Search with K-Means++	1.1910	1.21
PCA-guided Search with GMM	**1.1617**	2.35

Table 1. Performance Comparison: Inertia and Execution Time for Clustering on the MNIST Dataset

Choosing the Number of Clusters

As an extension, the MNIST dataset was first analyzed with $K = 10$ clusters to match its 10 classes, with the PCA-guided search with GMM approach. However, some digits spanned multiple clusters while others lacked clear assignments, and a few clusters acted as catch-all groups. Allowing certain digits to occupy multiple clusters (by increasing K) helped "free up" space for new, distinct clusters, improving separation and addressing variations in writing styles.

To optimize K, the balance between distinct clusters and reasonable class sizes was considered. An empirical criteria based on Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC) was used to evaluate K values from 10 to 20. Finally, $K = 18$ was chosen. This improved clustering achieved better separation, ensuring each digit had at least one dedicated cluster with minimal overlap.

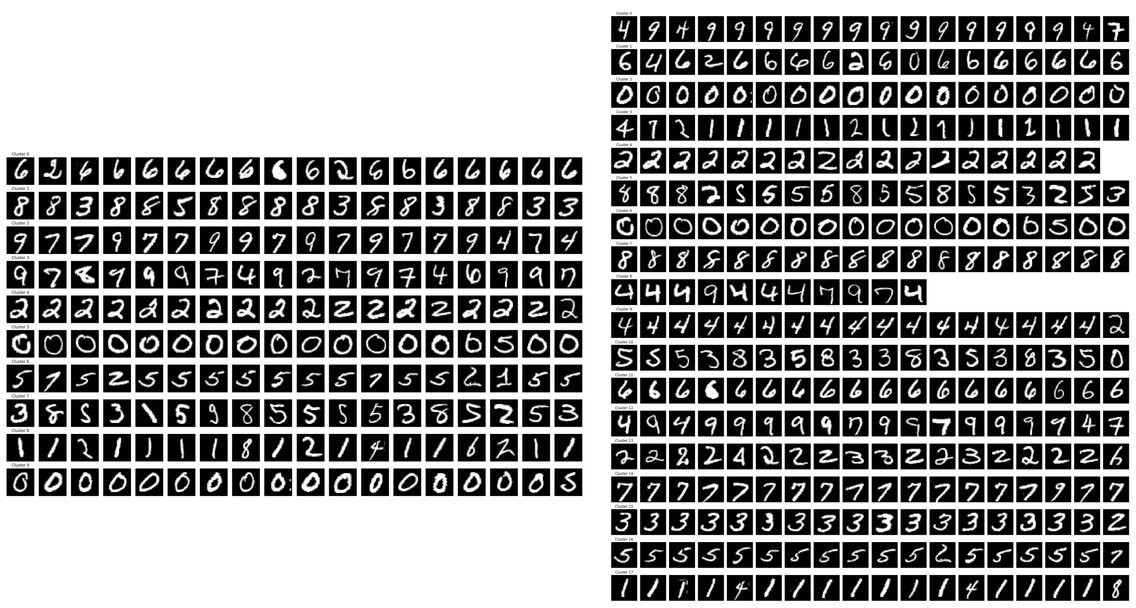


Figure 2. Random Samples from Each Clusters $K = 10$ (Left) vs $K = 18$ (Right).

References

[1] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. InTwenty-first international conference on Machine learning - ICML '04, NewYork, NewYork, USA, 2004. ACM Press.

[2] Ade Jamal, Annisa Handayani, Ali Akbar Septiandri, Endang Ripmiatin, andYunus Effendi. Dimensionality reduction using PCA and K-Means clustering for breast cancer prediction. Lontar Komputer Jurnal Ilmiah Teknologi Informasi, page 192, Dec 2018.

[3] Qin Xu, Chris Ding, Jinpei Liu, and Bin Luo. PCA-guided search for k-means. Pattern Recognit. Lett., 54:50–55, March 2015.