Dimensionality Reduction and Unsupervised Clustering of the Million Song Dataset: A Comparative Analysis of PCA, UMAP, K-Means and DBSCAN

Fernando Ayala
Department of Electrical Engineering
Universidad de los Andes
Email: fsayala@miuandes.cl

***Abstract*—This work presents an unsupervised learning analysis applied to a subset of the Million Song Dataset (MSD), evaluating the effectiveness of dimensionality reduction and clustering methods based on timbre, pitch, loudness, rhythm and global audio descriptors extracted from `.h5` metadata. We compare PCA, UMAP, K-Means, and DBSCAN in terms of cluster coherence, geometric separability and interpretability. Results show that UMAP captures local structure more effectively than PCA, while clustering methods struggle to produce coherent groupings, reflecting the weak intrinsic cluster structure of the MSD. K-Means produces moderately stable partitions, while DBSCAN is highly sensitive to density parameters and performs poorly in reduced spaces. The study highlights the limitations of classic clustering techniques when applied to high-dimensional music representations.**

I. Introduction

Modern music analysis often requires navigating high-dimensional acoustic descriptors. The Million Song Dataset (MSD) provides large-scale metadata that encode timbre, harmonic structure, loudness envelopes and rhythmic markers. Dimensionality reduction techniques such as PCA and UMAP help reveal latent structure, while clustering methods such as K-Means and DBSCAN uncover song groupings without labels.

This study systematically applies these methods to a curated MSD subset, aiming to evaluate:

1) How PCA and UMAP differ in capturing song-space geometry.
2) How clustering behaves in reduced spaces of different dimensionalities.
3) Which combinations yield coherent and interpretable musical clusters.

II. Methods

A. Dataset and Feature Extraction

Each track is stored as a `.h5` file containing audio descriptors. From each file we extract:

- 12-D mean timbre vector,
- 12-D mean pitch vector,
- mean segment loudness,
- global tempo,
- duration,
- key and mode.

This results in a 28-dimensional representation per song.

TABLE I: Dataset size and feature summary.

Item	Value
Total .h5 files found	1 000
Valid songs processed	1 000
Features per song	28
Timbre coefficients	12
Pitch coefficients	12
Loudness (mean)	1
Tempo	1
Duration	1
Key, Mode	2

B. Preprocessing

All features are standardized using *StandardScaler*. Exploratory Data Analysis (EDA) includes histograms, correlation matrices and summary statistics.

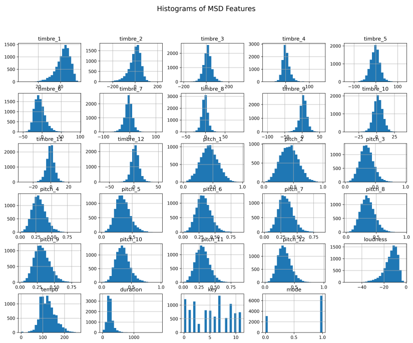

Fig. 1: Histograms of selected MSD features.

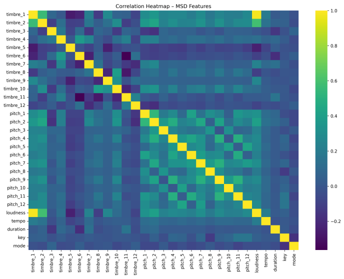

Fig. 2: Correlation heatmap of the 28 extracted features.

III. Dimensionality Reduction

A. *Principal Component Analysis (PCA)*

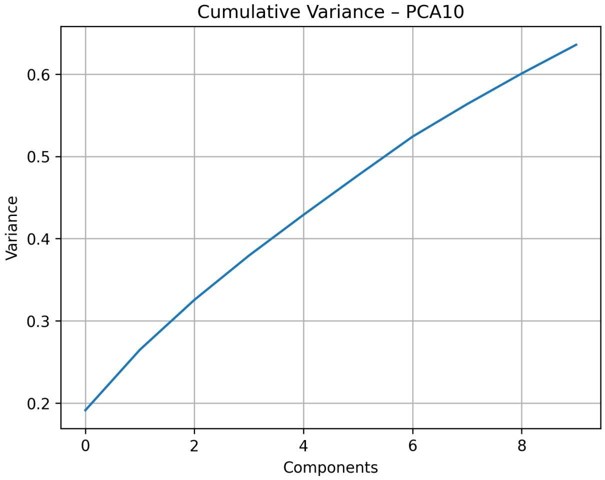

Fig. 3: Cumulative explained variance for the first PCA components.

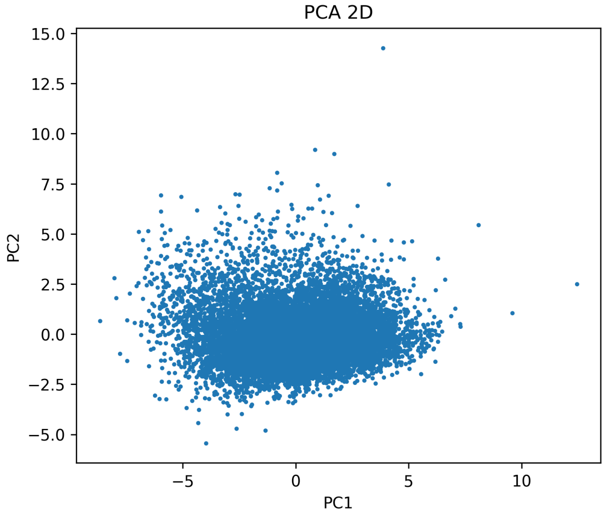

Fig. 4: PCA 2D embedding of the Million Song Dataset.

B. *UMAP*

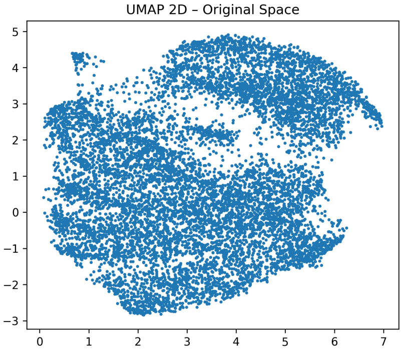

Fig. 5: UMAP 2D embedding using original standardized features.

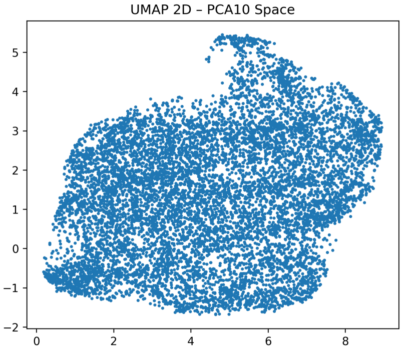

Fig. 6: UMAP 2D embedding computed from PCA-10.

IV. Clustering

A. *K-Means*

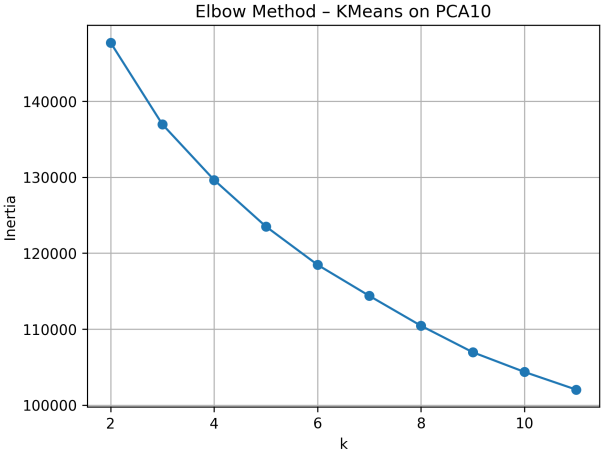

Fig. 7: Elbow plot for determining optimal k.

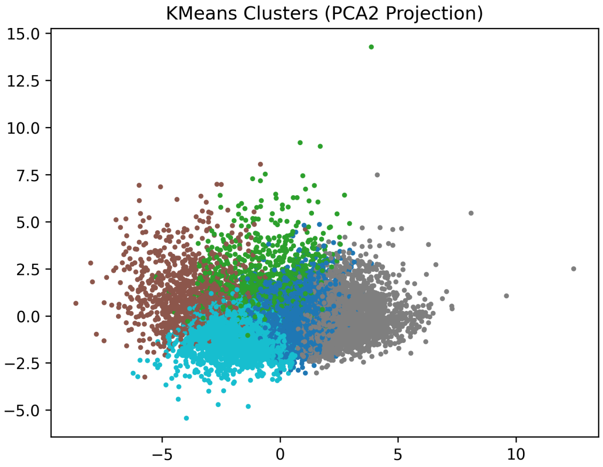

Fig. 8: K-Means clusters projected onto PCA 2D.

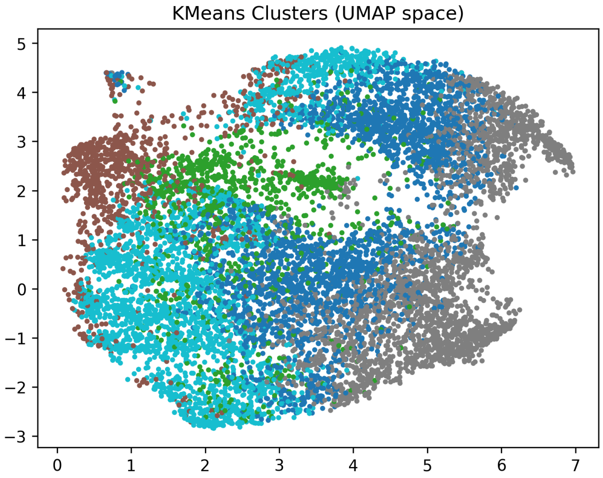

Fig. 9: K-Means clusters projected onto UMAP 2D.

B. DBSCAN

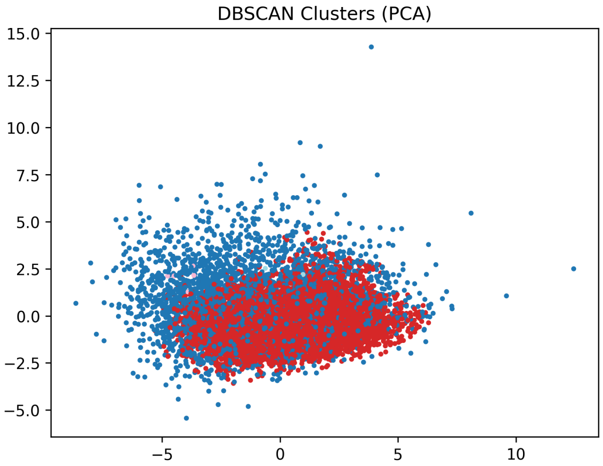

Fig. 10: DBSCAN clusters in PCA 10D projected to PCA2.

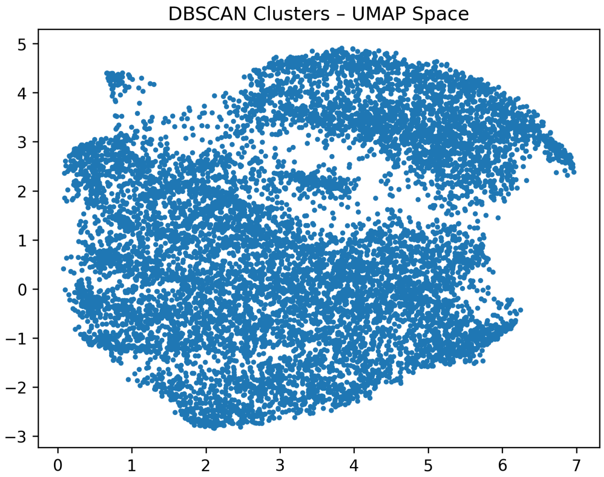

Fig. 11: DBSCAN clusters in UMAP 2D space.

V. Results and Discussion

A quantitative comparison was performed using the Silhouette Score, Rand Index and number of detected clusters.

TABLE II: Clustering metrics obtained from PCA10 and UMAP embeddings.

Method	Silhouette	RandIndex	Clusters
K-Means	0.1005	0.5014	5
DBSCAN	-0.2325	0.5014	3

A. Interpretation

The results indicate that the intrinsic cluster structure of the Million Song Dataset is weak. K-Means achieves a low but positive Silhouette score (0.10), suggesting mild but unstable cluster separation. DBSCAN obtains a negative Silhouette score (-0.23), indicating that density-based clustering fails to identify coherent groups in this feature space.

Both methods produce a Rand Index of 0.50, close to random agreement, showing that the two clustering techniques disagree substantially and do not capture consistent partitions.

Overall, the visualizations and metrics confirm that neither PCA nor UMAP reveals strongly separable groups with the set of 28 acoustic features extracted.

VI. Conclusions

This study shows that the Million Song Dataset exhibits low clusterability when represented through basic timbre, pitch, loudness and temporal descriptors. PCA provides global linear structure, and UMAP improves local organization, but neither transformation yields clearly separable clusters.

K-Means performs moderately better than DBSCAN, but both methods show low silhouette values and inconsistent partitions. These limitations highlight the need for richer feature representations—such as MFCC-based embeddings, temporal statistics, or deep audio encoders—to uncover meaningful structure in music similarity spaces.