Dimensionality Reduction and Clustering

Summer Marwany
Scientific Computing
Southern Methodist University

December 14, 2025

Project Description

This project applies dimensionality reduction (PCA, t-SNE, Isomap) and clustering (k-means, DBSCAN) to a synthetic manifold dataset and the Wine Quality dataset.

Part I: Synthetic Manifold

Data Generation

- Generated five disjoint patches in $(u, v) \in R^2$
- Applied a nonlinear map into R^{10}
- Added Gaussian noise to simulate real data

Embeddings and Observations

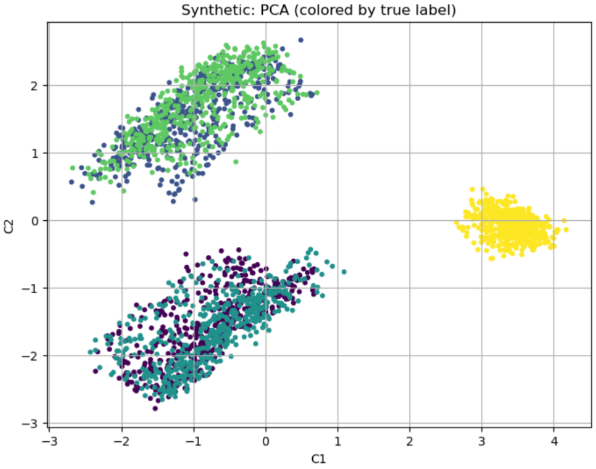

Figure 1: PCA embedding colored by true labels

- PCA partially separates clusters

- Linear projection cannot fully capture nonlinear structure

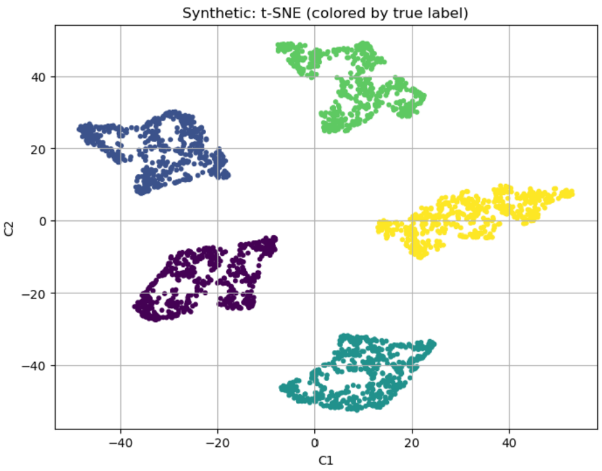

Figure 2: t-SNE embedding colored by true labels

- t-SNE clearly separates all five patches
- Preserves local neighborhoods, improving cluster visibility

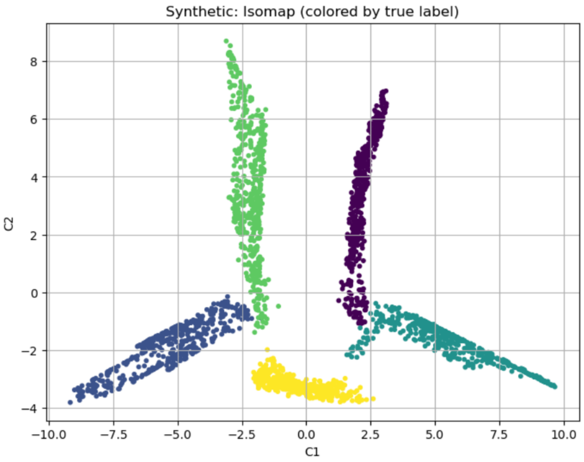

Figure 3: Isomap embedding colored by true labels

- Isomap captures manifold geometry
- Some distortion occurs due to graph approximation

Clustering Results

Method	Space	Clusters	Silhouette	ARI
k-means	Raw	5	0.4876	1.0000
DBSCAN	Raw	6	0.3529	0.9808
k-means	PCA	5	0.5842	0.5031
DBSCAN	PCA	3	0.7300	0.6150
k-means	t-SNE	5	0.6644	1.0000
DBSCAN	t-SNE	2	-0.3428	0.0001
k-means	Isomap	5	0.5744	0.8384
DBSCAN	Isomap	3	0.3659	0.6150

Table 1: Clustering performance on the synthetic manifold dataset

- k-means performs best due to compact, well-separated clusters
- k-means achieves perfect recovery (ARI = 1.0) in raw and t-SNE spaces
- PCA reduces separability due to linear projection
- DBSCAN is sensitive to density and fails on t-SNE

Part I Summary

- The synthetic data were generated from a low-dimensional nonlinear manifold embedded in R^{10}.
- Linear PCA partially separates the patches but cannot fully capture nonlinear structure.
- Nonlinear methods (t-SNE and Isomap) better reveal the underlying manifold geometry.
- k-means achieves the best clustering performance, with perfect recovery (ARI = 1.0) in raw and t-SNE spaces.
- DBSCAN is sensitive to density variations and performs poorly in distorted embedding spaces.

Part II: Wine Quality Data

Data Preparation

- Combined red and white wine datasets (6497 samples)
- Added `type_white` label
- Standardized physicochemical features

Embeddings and Observations

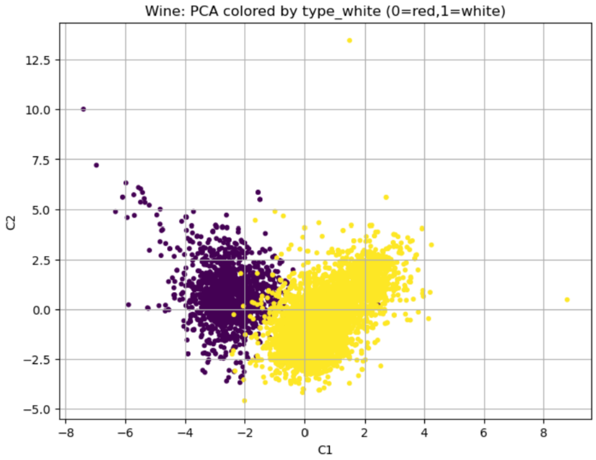

Figure 4: PCA colored by wine type

- Clear separation between red and white wines
- Indicates strong chemical differences

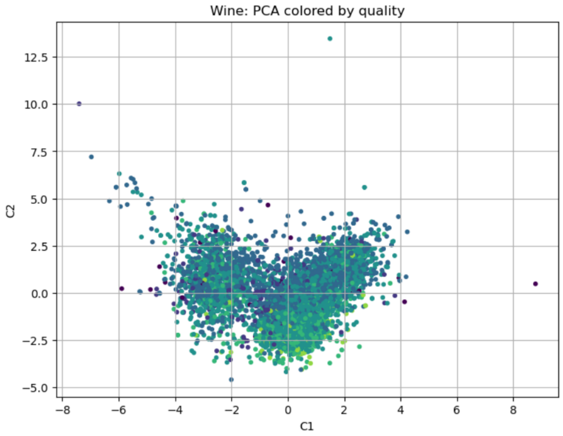

Figure 5: PCA colored by wine quality

- Significant overlap across quality levels
- Quality is not linearly separable

t-SNE Visualization

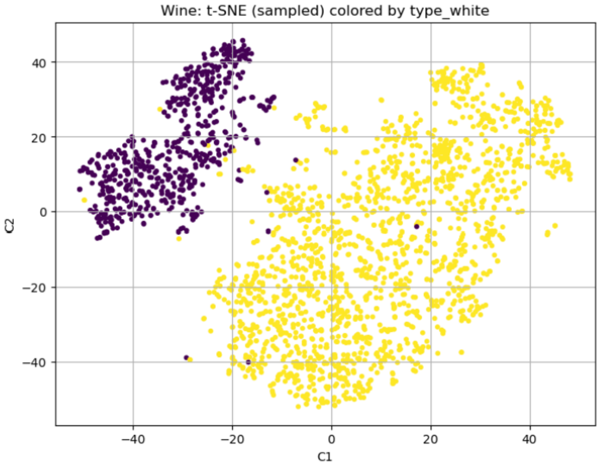

Figure 6: t-SNE (sampled) embedding colored by wine type

- Red and white wines form two clearly separated groups
- Confirms strong nonlinear separability of wine type
- Supports high ARI values from k-means clustering

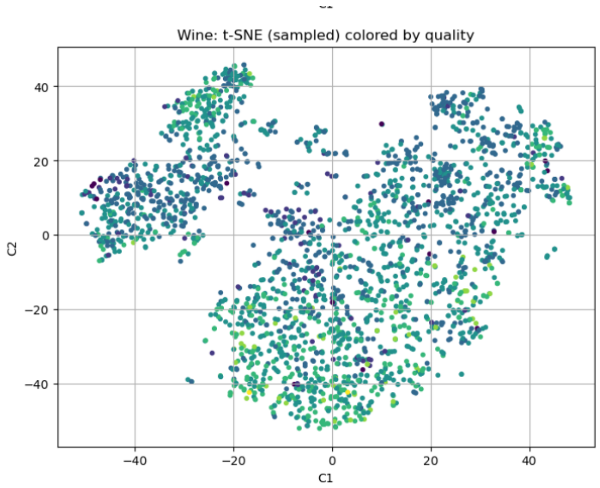

Figure 7: t-SNE (sampled) embedding colored by wine quality

- Quality labels are heavily mixed across the embedding
- No distinct clusters correspond to quality levels
- Explains poor clustering performance for quality

Summary

- PCA and t-SNE both clearly separate red and white wines, indicating strong chemical differences.
- Quality labels remain heavily mixed across all embeddings.
- Nonlinear embeddings confirm that quality does not align with low-dimensional structure.

Clustering Results

Method	Space	Clusters	Silhouette	ARI
k-means	Raw	2	0.2766	0.9417
DBSCAN	Raw	21	-0.4138	-0.0218
k-means	PCA	2	0.4632	0.9284
DBSCAN	PCA	2	0.5638	0.0131
k-means	t-SNE	2	0.4153	0.8678
DBSCAN	t-SNE	0	NaN	0.0000

Table 2: Clustering results using wine type (red vs white) as reference

- Wine type is well clustered using k-means
- k-means separates wine type well across all spaces
- PCA improves silhouette by removing noise
- DBSCAN fails due to varying density and overlap
- Wine quality shows poor clustering performance

Method	Space	Clusters	Silhouette	ARI
k-means	Raw	3	0.2351	0.0229
DBSCAN	Raw	21	-0.4138	0.0020
k-means	PCA	3	0.4931	0.0236
DBSCAN	PCA	2	0.5638	-0.0005
k-means	t-SNE	3	0.5091	0.0324
DBSCAN	t-SNE	0	NaN	0.0000

Table 3: Clustering results using regrouped wine quality as reference

- ARI values are near zero for all methods
- Quality labels do not align with cluster structure
- Confirms visual overlap observed in embeddings

Clustering Summary

- k-means consistently separates wine type with high ARI across raw and embedded spaces.
- PCA improves silhouette scores by reducing noise and redundancy.
- DBSCAN performs poorly due to varying density and overlap in wine data.
- Wine quality clustering yields ARI values near zero, even after regrouping.

Conclusion

- This project demonstrates how dimensionality reduction reveals structure that is not visible in high-dimensional data.
- Nonlinear embeddings such as t-SNE provide clearer visualization of manifold-based data than linear PCA.
- k-means performs best when clusters are compact and well-separated, while DBSCAN is sensitive to density variations.
- Wine quality does not form distinct clusters, indicating that quality is influenced by factors beyond the measured features.