Report

Comparison of PCA and t-SNE for Dimensionality Reduction

1. Trade-offs Between PCA and t-SNE

Principal Component Analysis (PCA)

- PCA is a **linear transformation technique** that maintains **global variance** in high-dimensional data.
- It aids in **feature selection and dimensionality reduction** while keeping interpretability intact.
- The explained variance ratio provides a way to **quantify the amount of retained information** after reduction.
- However, PCA **does not perform well for nonlinear relationships** and may fail to distinctly separate clusters.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE is a **nonlinear technique** that is effective in **clustering similar data points together**.
- It is particularly useful for **preserving local structures**, uncovering patterns that PCA might overlook.
- Unlike PCA, **t-SNE does not maintain original feature relationships**, making interpretation challenging.
- It is computationally expensive and may **yield different results on multiple runs** due to random initialization.

Key Trade-offs

Feature	PCA	t-SNE
Type of Transformation	Linear	Nonlinear
Interpretability	High	Low
Structure Preserved	Global	Local

Cluster Separation	Weak	Strong
Computational Cost	Low	High
Usage	Feature selection, preprocessing	Data visualization, clustering

---

2. Key Observations from the Visualizations

PCA Visualization (2D Projection)

- The PCA scatter plot **did not show distinct clusters**, as data was spread based on variance.
- It preserved the **global structure** but **overlapped different wine quality scores**.
- Since PCA is a linear transformation, it did not capture **nonlinear patterns** within the dataset.

t-SNE Visualization (2D Projection)

- The t-SNE scatter plot **revealed more distinct groupings**, suggesting potential clusters in the dataset.
- Similar wine quality scores appeared **more closely packed**, demonstrating **better local structure preservation**.
- However, t-SNE's axes **lack direct interpretability**, making it less suitable for understanding feature relationships.

Dimensionality Reduction vs. Information Loss

- From the **explained variance ratio** in PCA, **PC1 to PC6** accounted for **85% of the variance**, offering a balance between **dimensionality reduction and data retention**.
- t-SNE does not provide a metric like explained variance but excels at **clustering and uncovering hidden patterns**.

---

3. Conclusion

- **PCA is ideal for feature selection and understanding data variance**, making it useful in predictive modeling.

- **t-SNE is best suited for visualizing underlying structures** in high-dimensional datasets but is computationally demanding.
- The choice between PCA and t-SNE depends on the **purpose of the analysis**:
  - Use **PCA when reducing dimensionality for machine learning models**.
  - Use **t-SNE for clustering insights and exploratory data analysis**.