L05: PCA & t-SNE

Dimensionality Reduction for Visualization and Preprocessing

Methods and Algorithms – MSc Data Science

Learning Objectives

By the end of this lecture, you will be able to:

1. Apply PCA for dimensionality reduction and feature extraction
2. Interpret variance explained and choose number of components
3. Use t-SNE for visualization of high-dimensional data
4. Compare linear (PCA) vs non-linear (t-SNE) methods

Finance Application: Portfolio risk decomposition, asset clustering

From many features to meaningful low-dimensional representations

The Business Problem

Curse of Dimensionality

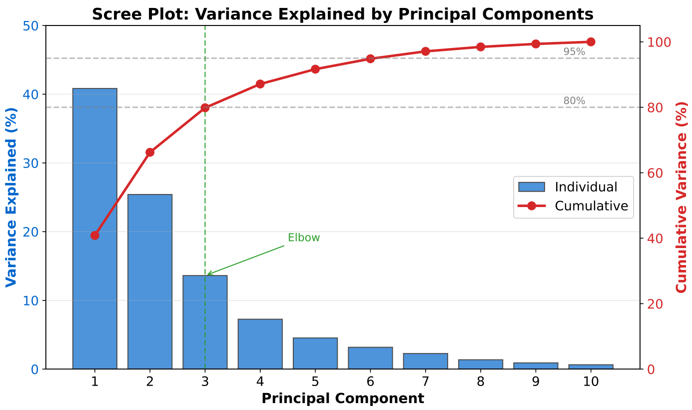
- Portfolio with 100+ assets: hard to visualize relationships
- Customer data with dozens of features: redundant information
- High dimensions cause sparsity and computational issues

Solutions

- **PCA**: Linear projection preserving maximum variance
- **t-SNE**: Non-linear embedding preserving local structure

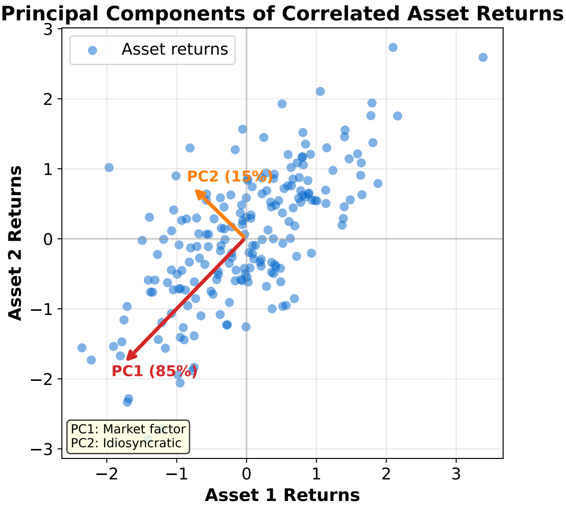
Reduce dimensions while preserving important information

Scree Plot: Choosing Components



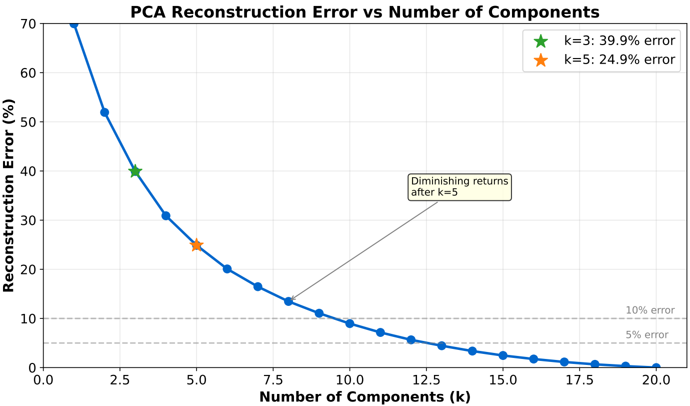
Choose k components capturing 80-95% of variance, or at the "elbow"

Principal Components



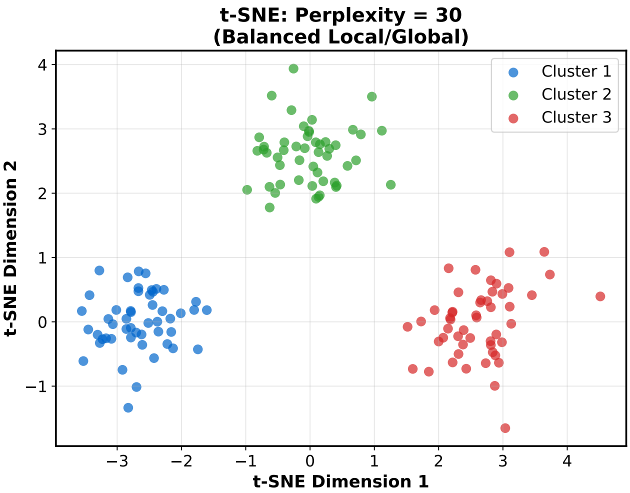
Principal components are orthogonal directions of maximum variance

Reconstruction Error



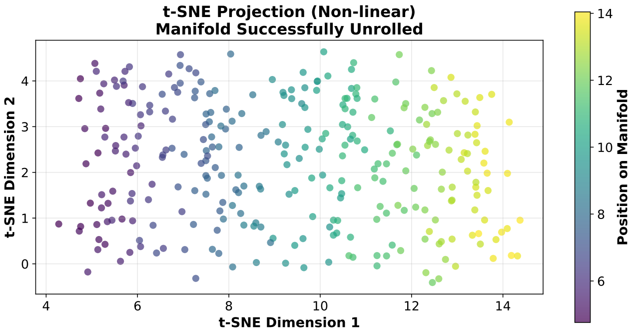
More components = lower error, but diminishing returns after elbow

t-SNE: Perplexity Effect



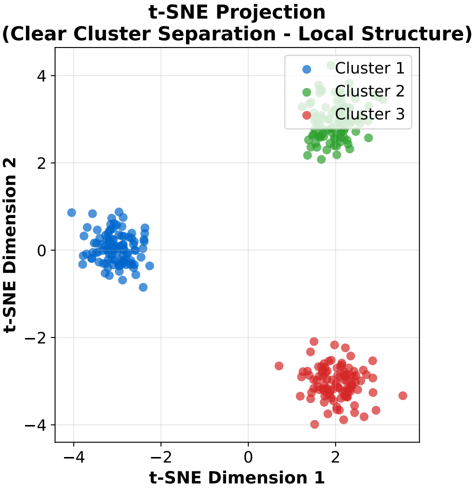
Perplexity controls local vs global structure preservation (try 5-50)

PCA vs t-SNE: Swiss Roll

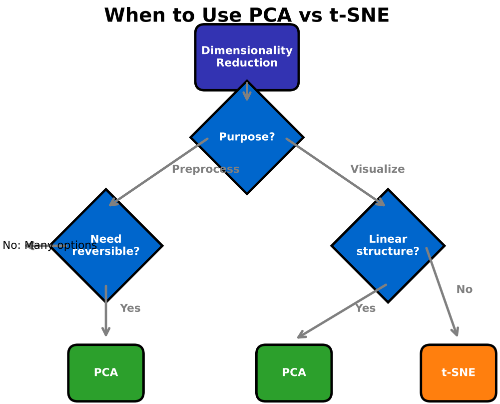


t-SNE unrolls non-linear manifolds that PCA cannot handle

Cluster Preservation



t-SNE better preserves cluster structure for visualization



PCA: Fast, linear, reversible, for preprocessing

t-SNE: Slow, non-linear, visualization only, preserves local structure

PCA for preprocessing/speed, t-SNE for visualization only