Lesson 31: PCA Dimensionality Reduction

Data Science with Python – BSc Course

45 Minutes

Learning Objectives

The Problem: We have 100 features but many are correlated. How do we reduce dimensions while preserving the most important information?

After this lesson, you will be able to:

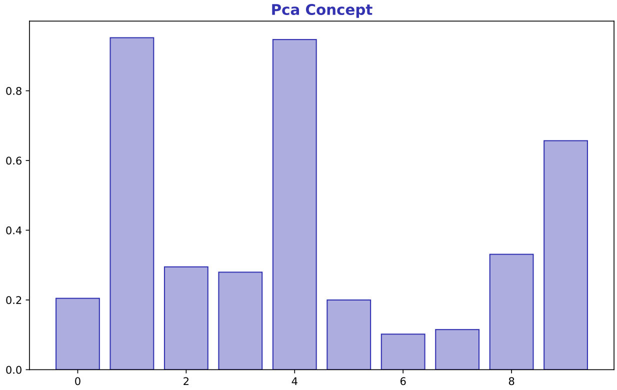
- Understand principal components as new axes
- Apply PCA with sklearn
- Interpret explained variance ratio
- Reduce feature dimensions for visualization and modeling

Finance Application: Factor discovery, risk decomposition, noise reduction

PCA Concept

Finding New Axes

- PC1: direction of maximum variance
- PC2: orthogonal to PC1, captures next most variance

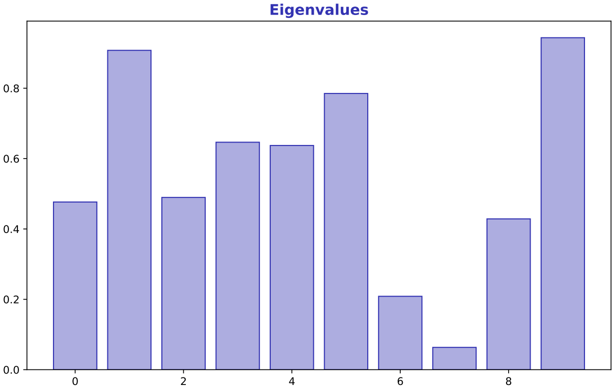


PCA rotates coordinate system to align with data's natural directions

Eigenvalues and Eigenvectors

The Math Behind PCA

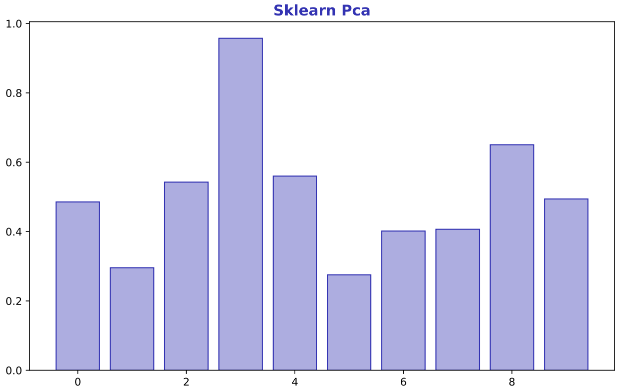
- Eigenvectors of covariance matrix = principal component directions
- Eigenvalues = variance explained by each component



Larger eigenvalue = more important component. Sum of eigenvalues = total variance.

Implementation in Python

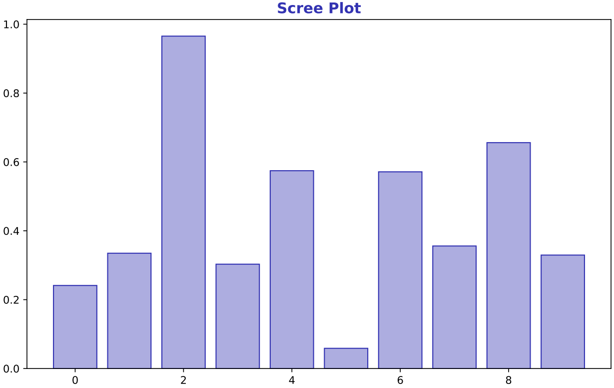
- `from sklearn.decomposition import PCA`
- `pca = PCA(n_components=2).fit_transform(X)`



Always standardize features first! PCA is sensitive to scale.

How Many Components to Keep?

- Plot eigenvalues in decreasing order
- Look for “elbow” where values level off

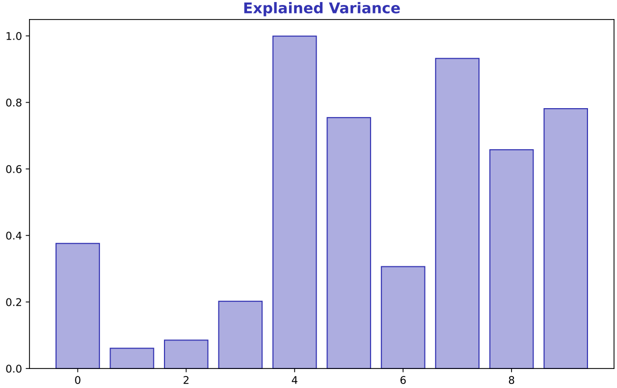


Alternative: keep components until cumulative variance $>$ 90%

Explained Variance

Cumulative Information Retained

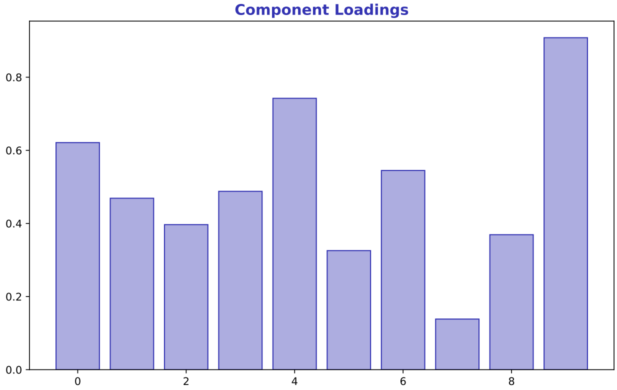
- `pca.explained_variance_ratio_` shows each component's share
- Cumulative sum tells total information retained



Example: 3 components explain 85% of variance – 97 features were mostly redundant

Interpreting What Components Mean

- Loadings = correlations between original features and components
- High loading = feature strongly influences that component

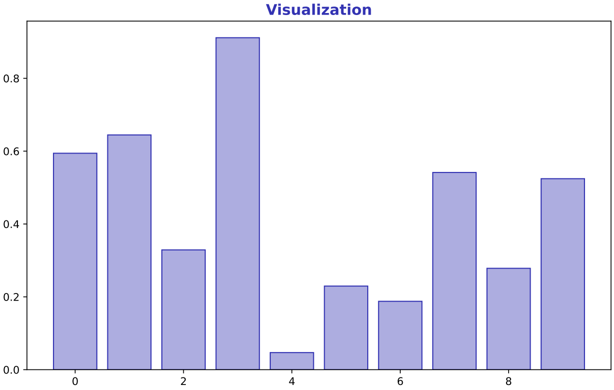


Finance: PC1 often = "market factor", PC2 = "size" or "value"

Visualization in 2D

Seeing High-Dimensional Data

- Project to PC1 vs PC2 for 2D scatter plot
- Reveals clusters and outliers invisible in original space

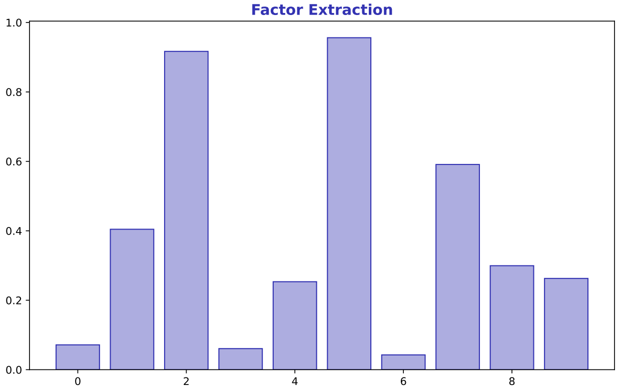


PCA visualization is exploratory – always check explained variance

Factor Extraction

Finance Application: Statistical Factors

- PCA on stock returns finds latent market factors
- First few PCs often correspond to market, sector, and style factors



Statistical PCA vs economic factors (Fama-French) – different but related

Task: Discover Factors in Stock Returns

1. Calculate daily returns for 30 stocks (1 year)
2. Standardize returns and fit PCA
3. Plot scree plot – how many components needed for 80% variance?
4. Examine PC1 loadings – what does it represent?
5. Project stocks to PC1 vs PC2 and color by sector

Deliverable: Scree plot + 2D projection with sector colors.

Extension: Compare PC1 to S&P 500 returns – are they correlated?

Lesson Summary

Problem Solved: We can now reduce high-dimensional data while preserving most information.

Key Takeaways:

- PCA finds orthogonal directions of maximum variance
- Scree plot and cumulative variance guide component selection
- Always standardize before PCA
- Finance: PCA extracts statistical factors from returns

Next Lesson: ML Pipeline (L32) – putting it all together

Memory: PCA = rotate to max variance axes. PC1 = most important direction.