# MA477: Data Science

**Lesson 32 Outline — 13 April 2026**

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

---

# 1 Administrative

- Project 3 Pitch
- Project 3 speed dating
- Dim Reduction Lecture
- Coding Exercise

# 2 Transfer Lesson Objectives

- Understand the curse of dimensionality
- Know how to calculate principal components (score and loading vectors).
- Understand PCA as an application of SVD and eigen decomposition of the covariance matrix.
- Use functions in sci-kit learn to perform principal components analysis for dimension reduction and identifying outliers.
- Use PCA to reduce dimensionality and reconstruct images or data using principal components.

# 3 Dimensionality Reduction, PCA, and Manifold Learning

## 3.1 The Curse of Dimensionality

As the number of features $d$ increases:

- The volume of space grows exponentially
- Data becomes sparse, even with large sample sizes
- Distances between points become less informative
- Models require more data and are prone to overfitting

**Key intuition:** High-dimensional data often lies near a much lower-dimensional structure.

## 3.2 What is Dimensionality Reduction?

Dimensionality reduction seeks to:

- Represent data using fewer variables
- Preserve as much information (variance/structure) as possible

Two main approaches:

- Linear methods: projection (e.g., PCA)
- Nonlinear methods: manifold learning

## 3.3 Principal Component Analysis (PCA)

**Core Idea:**

- First component captures the greatest variance
- Each subsequent component captures the next largest variance
- All components are orthogonal (uncorrelated)

## 3.4 Computing Principal Components

**Step 1: Center the data**

$$X_{\text{centered}} = X - \bar{X}$$

**Step 2: Compute covariance matrix**

$$\Sigma = \frac{1}{n} X^T X$$

**Step 3: Eigen decomposition**

- Eigenvectors: principal directions (loadings)
- Eigenvalues: variance explained

## 3.5 PCA Decomposition: Intuition and Mechanics

Let $X$ be the data matrix:

- Rows represent observations
- Columns represent features

After centering, PCA identifies the most important directions in the data.

### 3.5.1 Eigenvectors and Eigenvalues

- Eigenvectors define directions in feature space
- Eigenvalues measure how much variance lies along each direction

**Interpretation:**

- Eigenvectors = directions
- Eigenvalues = importance

### 3.5.2 SVD Representation

$$X = U\Sigma V^T$$

- $V$: principal directions (loadings)
- $\Sigma$: magnitude of variance
- $U$: representation of observations in new basis

#### 3.5.3 Scores (Projection)

$$Z = XV$$

- Rows of $Z$: observations in new coordinate system
- Columns of $Z$: principal components

#### 3.5.4 Dimensionality Reduction

Keep only the top $k$ components:

$$Z_k = XV_k$$

- Reduces dimension from $d$ to $k$, where $k \ll d$
- First components capture signal; later components capture noise

#### 3.5.5 Reconstruction

$$\hat{X} = Z_k V_k^T$$

- $\hat{X}$ approximates the original data
- Reconstruction error corresponds to discarded variance

#### 3.5.6 Geometric Interpretation

- Rotate the data
- Align with directions of maximum variance
- Project onto a lower-dimensional subspace

### 3.6 Variance Explained

- Each eigenvalue represents variance captured by a component
- Choose $k$ such that most variance is retained (e.g., 90–95%)

### 3.7 Interpretation of PCA

- Identifies dominant structure in the data
- Transforms correlated variables into uncorrelated components

### 3.8 Limitations of PCA

- Captures only linear relationships
- Sensitive to feature scaling
- Components may be difficult to interpret

### 3.9 Manifold Learning

**Core Idea:** Data lies on a low-dimensional curved surface embedded in high-dimensional space.

#### 3.9.1 Common Methods

- Isomap: preserves global geometry
- Locally Linear Embedding (LLE): preserves local structure
- t-SNE: preserves local similarity (useful for visualization)

## 3.10 PCA vs Manifold Learning

| Aspect | PCA | Manifold Methods |
|---|---|---|
| Type | Linear | Nonlinear |
| Structure | Global variance | Local geometry |
| Interpretability | High | Low |
| Use | Compression, preprocessing | Visualization |

## 3.11 Key Takeaways

- High-dimensional data leads to sparsity and modeling challenges
- PCA decomposes data into orthogonal directions ranked by importance
- Dimensionality reduction is achieved by keeping top $k$ components
- Reconstruction measures information loss
- Manifold learning extends dimensionality reduction to nonlinear structures