Principal Component Analysis

Math 628

November 14, 2025

Motivation: High-Dimensional Data

- We observe data vectors
$$x_1, x_2, \ldots, x_n \in \mathbb{R}^d.$$
- d can be large (many features, sensors, pixels, etc.).
- We would like to:
  - **Compress** each x_i into a lower-dimensional representation.
  - **Reconstruct** an approximation $\hat{x}_i$ of the original x_i.
  - Lose as little information as possible.
- Principal Component Analysis (PCA) finds the best linear low-dimensional representation in the sense of **least squared reconstruction error**.

Data Matrix and Centering

- Stack the data into a matrix

$$X = \begin{bmatrix} - \\ x_1^\top \\ - \\ \vdots \\ - \\ x_n^\top \\ - \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

- We **center** the data:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \tilde{x}_i = x_i - \bar{x}.$$

- After centering, we have

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i = 0.$$

Goal: One-Dimensional Linear Compression

- Start with the simplest case: compressing to **one dimension**.
- Choose a direction (unit vector)

$$w \in \mathbb{R}^d, \quad \|w\| = 1.$$

- We will:
  - **Project** each x_i onto the line spanned by w.
  - Store only a scalar coordinate z_i.
  - **Reconstruct** an approximation $\hat{x}_i$ from that coordinate.
- Question: *Which direction w gives the smallest reconstruction error on average?*

Projection and Reconstruction on a Line

- Projection of $x \in \mathbb{R}^d$ onto direction w:

$$z = w^\top x \in \mathbb{R}.$$

- Geometrically, z is the coordinate of x along w.
- Reconstruction from z:

$$\hat{x} = z\, w = (w^\top x)\, w.$$

- For each data point x_i:

$$z_i = w^\top x_i, \qquad \hat{x}_i = (w^\top x_i)\, w.$$

- Compression: $x_i \mapsto z_i$ (from $\mathbb{R}^d$ to $\mathbb{R}$).
  Decompression: $z_i \mapsto \hat{x}_i$ (from $\mathbb{R}$ back to $\mathbb{R}^d$).

Reconstruction Error (Single Point)

- For a single point x, the reconstruction error is

$$\|x - \hat{x}\|^2 = \|x - (w^\top x)\, w\|^2.$$

- Let

$$e = x - (w^\top x)w.$$

  Then

$$\|x - \hat{x}\|^2 = \|e\|^2 = e^\top e.$$

- Expand:

$$e = x - \alpha w, \quad \alpha = w^\top x.$$

$$\|e\|^2 = (x - \alpha w)^\top (x - \alpha w) = x^\top x - \alpha x^\top w - \alpha w^\top x + \alpha^2 w^\top w.$$

Reconstruction Error Expansion (cont.)

- Recall $\alpha = w^\top x$ and $x^\top w = w^\top x = \alpha$.
- Also, $\|w\|^2 = w^\top w = 1$.
- Substitute:
$$\|x - (w^\top x)w\|^2 = x^\top x - \alpha^2 - \alpha^2 + \alpha^2 \cdot 1 = x^\top x - \alpha^2.$$
- Since $\alpha = w^\top x$,
$$\|x - (w^\top x)w\|^2 = x^\top x - (w^\top x)^2 = \|x\|^2 - (w^\top x)^2.$$
- Therefore, for a single point:
$$\boxed{\|x - \hat{x}\|^2 = \|x\|^2 - (w^\top x)^2}.$$

Total Reconstruction Error

- For all centered data points $x_1, \ldots, x_n$:

$$E(w) = \sum_{i=1}^{n} \|x_i - (w^\top x_i)w\|^2.$$

- Using the previous result:

$$\|x_i - (w^\top x_i)w\|^2 = \|x_i\|^2 - (w^\top x_i)^2.$$

- Hence

$$E(w) = \sum_{i=1}^{n} \left(\|x_i\|^2 - (w^\top x_i)^2\right) = \sum_{i=1}^{n} \|x_i\|^2 - \sum_{i=1}^{n} (w^\top x_i)^2.$$

- Our optimization problem is:

$$\min_w E(w) \quad \text{subject to} \quad \|w\| = 1.$$

From Minimizing Error to Maximizing Variance

- The term
$$\sum_{i=1}^{n} \|x_i\|^2$$
does not depend on w.
- Therefore, minimizing $E(w)$ is equivalent to *maximizing*
$$\sum_{i=1}^{n} (w^\top x_i)^2.$$
- So we can rewrite the problem as
$$\boxed{\max_{w} \sum_{i=1}^{n} (w^\top x_i)^2 \quad \text{subject to} \quad \|w\| = 1.}$$
- Intuition:
  - We seek the direction w along which the **projected data** have the largest squared norm (largest variance).

Matrix Form and Covariance Matrix

- Recall the centered data matrix

$$X = \begin{bmatrix} - \\ x_1^\top \\ - \\ \vdots \\ - \\ x_n^\top \\ - \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

- Then the vector of projections is

$$Xw = \begin{bmatrix} w^\top x_1 \\ \vdots \\ w^\top x_n \end{bmatrix} \in \mathbb{R}^n.$$

- The sum of squared projections can be written as

$$\sum_{i=1}^{n}$$

Covariance Matrix

- Define the sample covariance matrix

$$S = \frac{1}{n-1} X^\top X \in \mathbb{R}^{d \times d}.$$

- Then

$$X^\top X = (n-1)S.$$

- Consequently,

$$\sum_{i=1}^{n} (w^\top x_i)^2 = w^\top X^\top X w = (n-1)\, w^\top S w.$$

- The factor $(n-1)$ does not depend on w, so maximizing $\sum_{i=1}^{n} (w^\top x_i)^2$ is equivalent to maximizing $w^\top S w$.
- Our optimization becomes

$$\boxed{\max_{w} w^\top S w \quad \text{subject to} \quad \|w\| = 1.}$$

Solving the Optimization: Eigenvalue Problem

- We must solve
$$\max_{w} w^\top S w \quad \text{subject to} \quad w^\top w = 1.$$
- This is a constrained optimization problem; use a Lagrange multiplier λ:
$$\mathcal{L}(w, \lambda) = w^\top S w - \lambda(w^\top w - 1).$$
- Take derivative with respect to w and set to zero:
$$\nabla_w \mathcal{L}(w, \lambda) = 2Sw - 2\lambda w = 0.$$

Hence
$$Sw = \lambda w.$$

Rayleigh Quotient and First Principal Component

- The condition

$$Sw = \lambda w$$

means (λ, w) is an **eigenpair** of S.

- The value of the objective is

$$w^\top Sw = w^\top(\lambda w) = \lambda(w^\top w) = \lambda,$$

since $w^\top w = 1$.

- Therefore:
  - Maximizing $w^\top Sw$ over unit vectors w is equivalent to choosing the eigenvector associated with the **largest eigenvalue** of S.
  - This eigenvector is the direction that:
    - Maximizes the variance of the projected data.
    - Minimizes the total reconstruction error.
- This direction is called the **first principal component**.

Geometric Interpretation

- Each point x_i can be decomposed as

$$x_i = (w^\top x_i)w + r_i,$$

where

$$r_i = x_i - (w^\top x_i)w$$

is orthogonal to w:

$$w^\top r_i = 0.$$

- The reconstruction keeps only the component along w:

$$\hat{x}_i = (w^\top x_i)w.$$

- The reconstruction error is the squared length of the orthogonal residual:

$$\|x_i - \hat{x}_i\|^2 = \|r_i\|^2.$$

- PCA chooses w such that the sum of these squared orthogonal distances is minimized.

Generalization to k-Dimensional Subspace

- Instead of a single direction w, choose k orthonormal directions:

$$W = [w_1, w_2, \ldots, w_k] \in \mathbb{R}^{d \times k}, \qquad W^\top W = I_k.$$

- Projection of x_i onto this k-dimensional subspace:

$$z_i = W^\top x_i \in \mathbb{R}^k.$$

- Reconstruction from z_i:

$$\hat{x}_i = W z_i = W W^\top x_i.$$

- Total reconstruction error:

$$E(W) = \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 = \sum_{i=1}^{n} \|x_i - W W^\top x_i\|^2.$$

- Goal:

$$\min_W E(W) \quad \text{subject to} \quad W^\top W = I_k.$$

Solution for k Components (Sketch)

- It can be shown (using similar arguments and properties of eigenvalues) that:
  - The matrix W that minimizes $E(W)$ is obtained by taking the eigenvectors of S corresponding to the k **largest eigenvalues** $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.
  - These are the first k **principal components**.
- The total variance of the data (trace of S) is

$$\text{tr}(S) = \sum_{j=1}^{d} \lambda_j.$$

- The variance captured by the first k components is

$$\sum_{j=1}^{k} \lambda_j.$$

Computing PCA: Naïve vs Practical

- In theory, we defined PCA using the covariance matrix

$$S = \frac{1}{n-1} X^\top X \in \mathbb{R}^{d \times d},$$

  and its eigen-decomposition:

$$S w_j = \lambda_j w_j.$$

- **Naïve approach:**
  1. Form $S = \frac{1}{n-1} X^\top X$.
  2. Compute eigenvalues/eigenvectors of S.
- Potential problems:
  - Explicitly forming S costs $\mathcal{O}(nd^2)$ operations.
  - Storing S costs $\mathcal{O}(d^2)$ memory.
  - For very large d, this becomes expensive or impossible.
- In practice, we usually compute PCA via **Singular Value Decomposition (SVD)**.

Summary

- PCA can be derived as:
  - Choosing a low-dimensional linear subspace.
  - Projecting data onto that subspace and reconstructing back.
  - **Minimizing** the total squared reconstruction error.
- For 1D:
  - Find unit vector w minimizing $\sum_i \|x_i - (w^\top x_i)w\|^2$.
  - Equivalent to maximizing $w^\top S w$.
  - Solution: eigenvector of S with largest eigenvalue.
- For kD:
  - Choose W with k orthonormal columns to minimize reconstruction error.
  - Solution: first k eigenvectors of S (principal components).
- Thus, **maximum variance** and **minimum reconstruction loss** are two equivalent views of PCA.