Lecture 15: Principal Component Analysis

CS 169 / CS 268
Introduction to Optimization

Ioannis Panageas

November 24, 2026

Representing data with fewer numbers

Each observation is a vector $\boldsymbol{x}_i \in \mathbb{R}^d$. We want to describe it using only $k < d$ numbers.

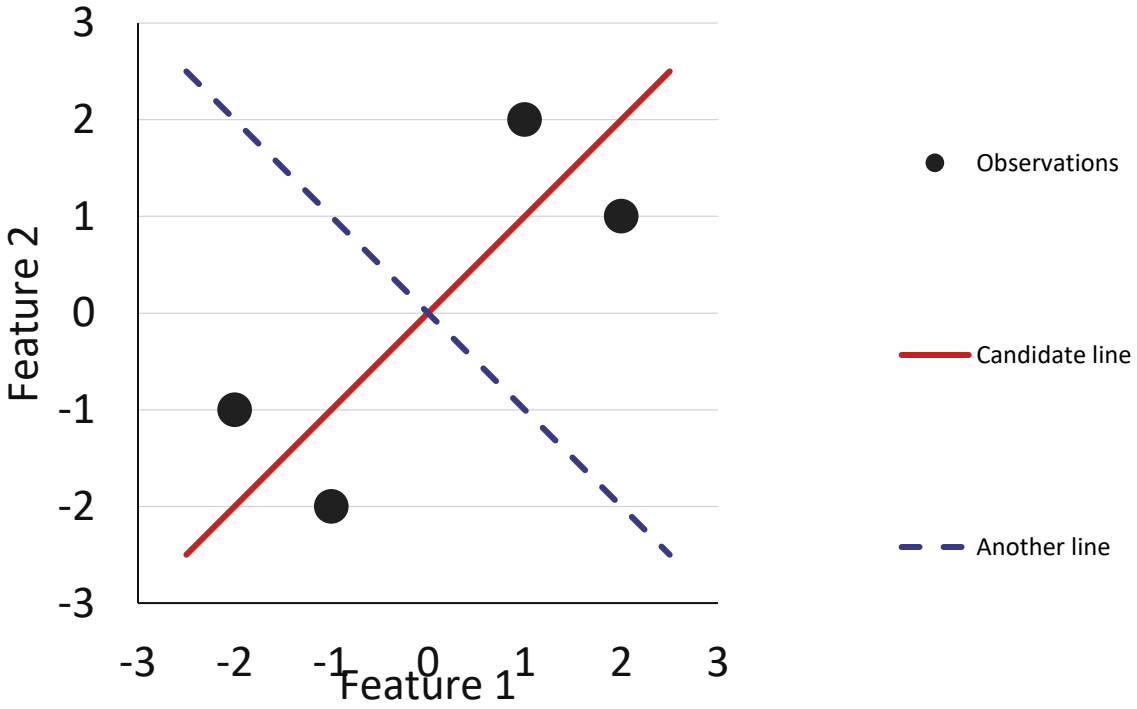
Dimensionality reduction replaces the original features by a smaller set of coordinates.

Principal component analysis (PCA) chooses a linear subspace that minimizes squared reconstruction error.

We use only the input vectors. There are no target labels y_i in this optimization problem.

A geometric example

Which line gives the best one-dimensional representation?



The error will be the perpendicular distance from each point to the line.

Centering the observations

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

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

Below, x_i denotes a centered observation. We omit the tilde to keep the notation simple.

After reconstruction, add the mean back to return to the original coordinates.

A line in arbitrary dimension

$$\{t\boldsymbol{u} : t \in \mathbb{R}\}, \qquad \boldsymbol{u} \in \mathbb{R}^d, \quad \|\boldsymbol{u}\| = 1.$$

A unit vector has Euclidean norm one. Fixing its length lets t describe distance along the line.

One coordinate: $z_i \in \mathbb{R}$. Reconstruction: $\widehat{\boldsymbol{x}}_i = z_i \boldsymbol{u}$.

First fix the line and find its best coordinate z_i. Then choose the best line.

Deriving the projection

For one vector $\boldsymbol{x}$ and a fixed unit direction $\boldsymbol{u}$, minimize the squared distance to $t\boldsymbol{u}$.

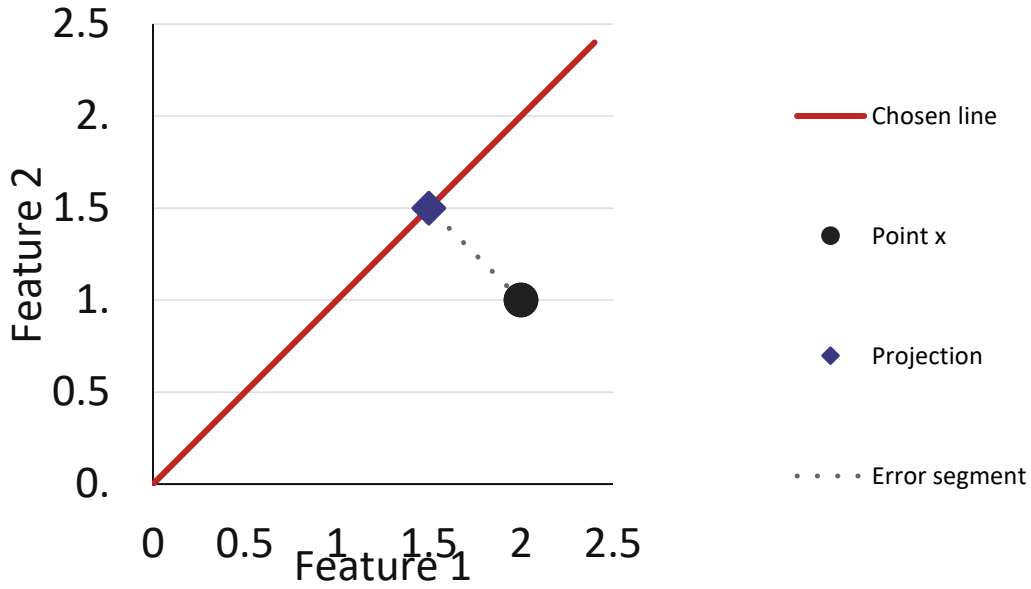
$$\|\boldsymbol{x} - t\boldsymbol{u}\|^2 = \|\boldsymbol{x}\|^2 - 2t\boldsymbol{u}^T\boldsymbol{x} + t^2.$$

$$\frac{d}{dt}\|\boldsymbol{x} - t\boldsymbol{u}\|^2 = -2\boldsymbol{u}^T\boldsymbol{x} + 2t = 0.$$

$$\boxed{z = \boldsymbol{u}^T\boldsymbol{x}, \qquad \widehat{\boldsymbol{x}} = (\boldsymbol{u}^T\boldsymbol{x})\boldsymbol{u}.}$$

Projection in the plane

$$\boldsymbol{x} = (2,1)^T, \quad \boldsymbol{u} = \frac{1}{\sqrt{2}}(1,1)^T, \quad \widehat{\boldsymbol{x}} = (3/2, 3/2)^T.$$



The coordinate is $z = 3/\sqrt{2}$. The squared error is $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|^2 = 1/2$.

The reconstruction objective

For centered observations, define the average squared reconstruction error:

$$J(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - (\boldsymbol{u}^T\boldsymbol{x}_i)\boldsymbol{u}\|^2, \qquad \|\boldsymbol{u}\| = 1.$$

The data vectors are fixed. The optimization variable is the direction $\boldsymbol{u}$.

PCA with one component chooses a line that minimizes $J(\boldsymbol{u})$.

The covariance matrix

$$C = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = \frac{1}{n} X^T X \in \mathbb{R}^{d\times d}.$$

Row i of $X \in \mathbb{R}^{n\times d}$ is $\boldsymbol{x}_i^T$. Entry C_{ab} averages the products $x_{ia}x_{ib}$.

$$C^T = C, \qquad \boldsymbol{v}^T C \boldsymbol{v} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{v}^T \boldsymbol{x}_i)^2 \geq 0.$$

Thus C is symmetric and positive semidefinite.

Variance along a direction

$$z_i = \boldsymbol{u}^T \boldsymbol{x}_i, \qquad \frac{1}{n}\sum_{i=1}^{n} z_i = 0.$$

For zero-mean scalar data, variance is the average squared coordinate.

$$\begin{aligned} V(\boldsymbol{u}) &= \frac{1}{n}\sum_{i=1}^{n} z_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u} = \boxed{\boldsymbol{u}^T C \boldsymbol{u}}. \end{aligned}$$

Larger variance means greater spread along the line.

Error equals lost variance

For a unit direction, the projection expansion gives

$$\begin{aligned}\|\boldsymbol{x}_i - (\boldsymbol{u}^T\boldsymbol{x}_i)\boldsymbol{u}\|^2 &= \|\boldsymbol{x}_i\|^2 - 2(\boldsymbol{u}^T\boldsymbol{x}_i)^2 + (\boldsymbol{u}^T\boldsymbol{x}_i)^2 \\ &= \|\boldsymbol{x}_i\|^2 - (\boldsymbol{u}^T\boldsymbol{x}_i)^2.\end{aligned}$$

$$\boxed{J(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2 - \boldsymbol{u}^T C\boldsymbol{u}.}$$

The first term is fixed. Minimizing error is equivalent to maximizing retained variance.

PCA as constrained optimization

$$\max_{\boldsymbol{u} \in \mathbb{R}^d} \ \boldsymbol{u}^T C \boldsymbol{u} \qquad \text{subject to} \qquad \boldsymbol{u}^T \boldsymbol{u} = 1.$$

The unit sphere is the set of vectors with norm one. It fixes the scale of the direction.

The feasible set is not convex: the midpoint of $\boldsymbol{u}$ and $-\boldsymbol{u}$ is zero, which is not feasible.

We will use stationarity to identify candidates, then prove which candidates are global maximizers.

The Lagrange multiplier equation

$$\mathcal{L}(\boldsymbol{u}, \lambda) = \boldsymbol{u}^T C \boldsymbol{u} - \lambda(\boldsymbol{u}^T \boldsymbol{u} - 1).$$

$$\nabla_{\boldsymbol{u}} \mathcal{L} = 2C\boldsymbol{u} - 2\lambda \boldsymbol{u} = \mathbf{0}.$$

$$\boxed{C\boldsymbol{u} = \lambda \boldsymbol{u}, \qquad \|\boldsymbol{u}\| = 1.}$$

The constraint gradient is $2\boldsymbol{u} \neq \mathbf{0}$ on the sphere, so the equality-constraint stationarity condition applies.

Eigenvalues describe directional variance

A nonzero vector $\boldsymbol{v}$ is an eigenvector of C with eigenvalue λ when $C\boldsymbol{v} = \lambda\boldsymbol{v}$.

$$C\boldsymbol{u} = \lambda\boldsymbol{u},\ \|\boldsymbol{u}\| = 1 \quad \Longrightarrow \quad \boldsymbol{u}^T C\boldsymbol{u} = \lambda.$$

A real symmetric matrix has an orthonormal eigenbasis: unit eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ that are mutually perpendicular.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

Why the largest eigenvalue wins

Expand any unit direction in the orthonormal eigenbasis:

$$\boldsymbol{u} = \sum_{j=1}^{d} a_j \boldsymbol{v}_j, \qquad \sum_{j=1}^{d} a_j^2 = 1.$$

$$\boldsymbol{u}^T C \boldsymbol{u} = \sum_{j=1}^{d} \lambda_j a_j^2 \leq \lambda_1 \sum_{j=1}^{d} a_j^2 = \lambda_1.$$

Choosing $\boldsymbol{u} = \boldsymbol{v}_1$ attains the bound. A top eigenvector is therefore a global optimum.

Stationarity does not identify the best

$$C\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1, \qquad C\boldsymbol{v}_d = \lambda_d \boldsymbol{v}_d.$$

Both directions satisfy the Lagrange multiplier equations. Their retained variances can be very different.

$$V(\boldsymbol{v}_1) = \lambda_1, \qquad V(\boldsymbol{v}_d) = \lambda_d.$$

When $\lambda_1 > \lambda_d$, the bottom direction gives the largest reconstruction error among unit directions.

Computing the example covariance

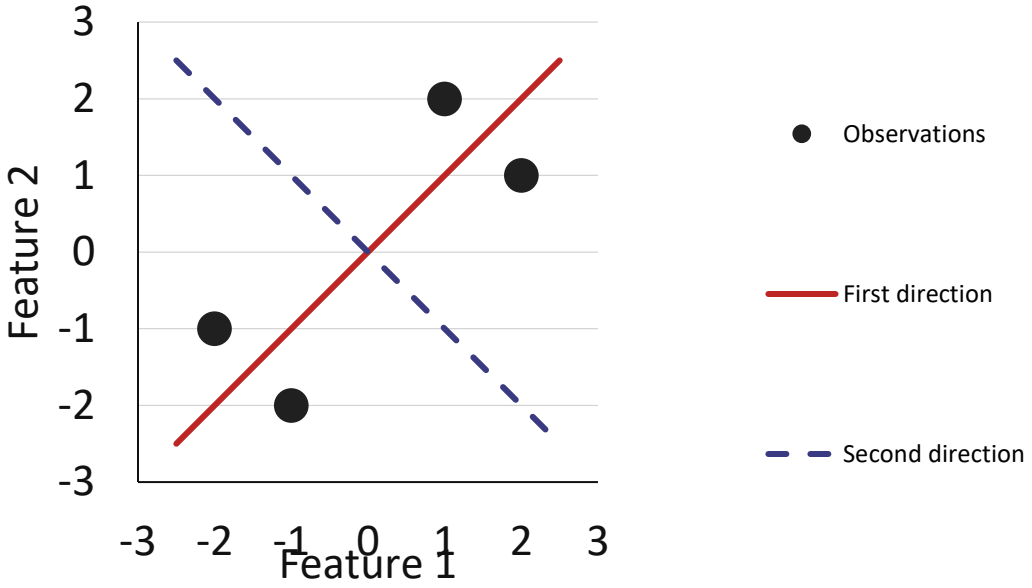
$$X = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -2 & -1 \\ -1 & -2 \end{pmatrix}, \qquad \bar{\boldsymbol{x}} = \mathbf{0}.$$

$$C = \frac{1}{4} X^T X = \frac{1}{4} \begin{pmatrix} 10 & 8 \\ 8 & 10 \end{pmatrix} = \begin{pmatrix} 5/2 & 2 \\ 2 & 5/2 \end{pmatrix}.$$

Total variation is $C_{11} + C_{22} = 5$. We now choose which one-dimensional direction to retain.

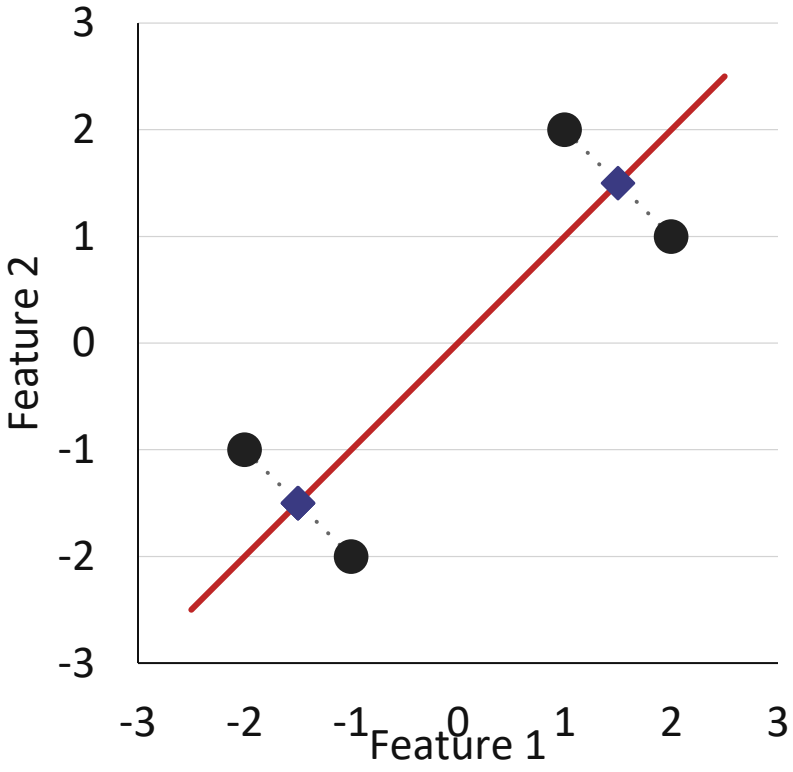
The example principal directions

$$v_1 = \frac{1}{\sqrt{2}}(1,1)^T,\ \lambda_1 = \frac{9}{2}; \qquad v_2 = \frac{1}{\sqrt{2}}(1,-1)^T,\ \lambda_2 = \frac{1}{2}.$$



The first component keeps variance $9/2$. Its line is $x_2 = x_1$.

The reconstructed observations



$$(2,1),(1,2) \mapsto (3/2,3/2), \qquad (-2,-1),(-1,-2) \mapsto (-3/2,-3/2).$$

Every squared error is $1/2$, so the average error is $J(v_1) = 1/2$.

When is the PCA direction unique?

The vectors v_1 and $-v_1$ describe the same line and give the same reconstructed points.

$$(-v_1)(-v_1)^T = v_1 v_1^T.$$

If $\lambda_1 > \lambda_2$, the optimal line is unique. Its unit direction is unique only up to sign.

If $\lambda_1 = \lambda_2$, different lines in the top eigenspace can attain the same optimum.

Keeping several components

To keep k coordinates, choose k orthonormal directions $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$.

$$U = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_k] \in \mathbb{R}^{d \times k}, \qquad U^T U = I_k.$$

The chosen subspace consists of the vectors $U\boldsymbol{z}$, where $\boldsymbol{z} \in \mathbb{R}^k$.

We assume $1 \leq k < d$. Orthogonality prevents two components from representing the same direction.

Coordinates and projection onto a subspace

$$\|\boldsymbol{x} - U\boldsymbol{z}\|^2 = \|\boldsymbol{x}\|^2 - 2\boldsymbol{z}^T U^T \boldsymbol{x} + \|\boldsymbol{z}\|^2,$$
$$\nabla_z \|\boldsymbol{x} - U\boldsymbol{z}\|^2 = -2U^T \boldsymbol{x} + 2\boldsymbol{z}.$$

$$\boxed{\boldsymbol{z} = U^T \boldsymbol{x}, \qquad \widehat{\boldsymbol{x}} = UU^T \boldsymbol{x}.}$$

The matrix $P = UU^T$ is the orthogonal projector onto the selected subspace.

Retained variance with k components

$$\|U^T \boldsymbol{x}_i\|^2 = \sum_{\ell=1}^{k} (\boldsymbol{u}_\ell^T \boldsymbol{x}_i)^2.$$

$$V(U) = \frac{1}{n} \sum_{i=1}^{n} \|U^T \boldsymbol{x}_i\|^2 = \sum_{\ell=1}^{k} \boldsymbol{u}_\ell^T C \boldsymbol{u}_\ell.$$

$$\max_{U^T U = I_k} \sum_{\ell=1}^{k} \boldsymbol{u}_\ell^T C \boldsymbol{u}_\ell.$$

Projection again splits total squared length into retained length and reconstruction error.

Weights in the eigenbasis

For orthonormal columns of U, define a weight for each covariance eigenvector $\boldsymbol{v}_j$.

$$q_j = \sum_{\ell=1}^{k} (\boldsymbol{v}_j^T \boldsymbol{u}_\ell)^2 = \|U^T \boldsymbol{v}_j\|^2.$$

Projection cannot increase length, so $0 \leq q_j \leq \|\boldsymbol{v}_j\|^2 = 1$.

$$\sum_{j=1}^{d} q_j = \sum_{\ell=1}^{k} \sum_{j=1}^{d} (\boldsymbol{v}_j^T \boldsymbol{u}_\ell)^2 = \sum_{\ell=1}^{k} \|\boldsymbol{u}_\ell\|^2 = k.$$

Why the top k eigenvectors win

Use $0 \le q_j \le 1$ and $\sum_j q_j = k$ to bound the retained variance.

$$V(U) = \sum_{j=1}^{d} \lambda_j q_j \le \sum_{j=1}^{k} \lambda_j q_j + \lambda_k \sum_{j>k} q_j.$$

$$V(U) \le k\lambda_k + \sum_{j=1}^{k} (\lambda_j - \lambda_k) q_j$$

$$\le k\lambda_k + \sum_{j=1}^{k} (\lambda_j - \lambda_k) = \sum_{j=1}^{k} \lambda_j.$$

$U = [v_1 \ \cdots \ v_k]$ attains this bound.

The optimal reconstruction error

$$\mathrm{tr}(C) = \sum_{a=1}^{d} C_{aa} = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 = \sum_{j=1}^{d} \lambda_j.$$

The trace of a square matrix is the sum of its diagonal entries.

$$J_k^* = \sum_{j=k+1}^{d} \lambda_j, \qquad V_k^* = \sum_{j=1}^{k} \lambda_j.$$

If the centered data lie in a subspace of dimension at most k, reconstruction can be exact.

The PCA procedure

1. Compute the data mean and center the observations.

2. Form $C = X^T X / n$ and find orthonormal eigenvectors with the k largest eigenvalues.

3. Store those directions as the columns of U. For a new observation, use

$$z = U^T(x - \bar{x}), \qquad \hat{x} = \bar{x} + Uz.$$