Mathematics of Image and Data Analysis
Math 5467

Lecture 4: Principal Component Analysis

Instructor: Jeff Calder
Email: jcalder@umn.edu

http://www-users.math.umn.edu/~jwcalder/5467S21

Last time

- Diagonalization and Vector Calculus
- Introduction to Numpy and reading/writing images in Python.

Today

- Principal Component analysis (PCA)

Recall

Let $v_1, \ldots, v_k$ be orthonormal vectors in $\mathbb{R}^n$ and set

$$L = \operatorname{span}\{v_1, v_2, \ldots, v_k\},$$

and

$$V = \begin{bmatrix} v_1 & v_2 & \ldots & v_k \end{bmatrix}.$$

Then we have

- $\operatorname{Proj}_L x = VV^T x$
- $\|\operatorname{Proj}_L x\|^2 = \sum_{i=1}^k (x^T v_i)^2$
- $\|x\|^2 = \|\operatorname{Proj}_L x\|^2 + \|x - \operatorname{Proj}_L x\|^2$

Given $x_0 \in \mathbb{R}^n$, projection onto an affine space $A = x_0 + L$ is given by

$$\operatorname{Proj}_A x = x_0 + \operatorname{Proj}_L (x - x_0).$$

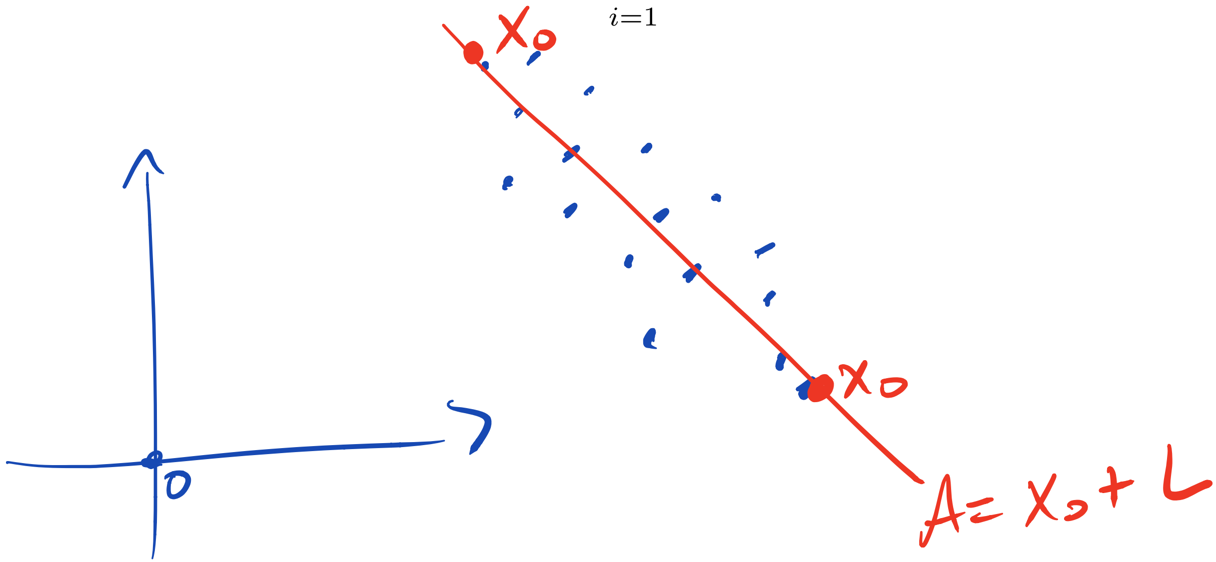
Also, for a symmetric matrix A

$$\nabla \|Ax\|^2 = 2A^2 x.$$

Principal Component Analysis (PCA)

Given points $x_1, x_2, \ldots, x_m$ in $\mathbb{R}^n$, find the k-dimensional linear or affine subspace that "best fits" the data in the mean-squared sense. That is, we seek an affine subspace $A = x_0 + L$ that minimizes the energy

$$E(x_0, L) = \sum_{i=1}^{m} \|x_i - \mathrm{Proj}_A x_i\|^2.$$



Optimizing over x_0

Claim: For any L, the function $x_0 \mapsto E(x_0, L)$ is minimized by the centroid

$$x_0 = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

Centroid

Proof:

$$E(x_0, L) = \sum_{i=1}^{m} \| x_i - \mathrm{proj}_A x_i \|^2$$

$A = x_0 + L$

$$= \sum_{i=1}^{m} \| x_i - (x_0 + \mathrm{proj}_L (x_i - x_0)) \|^2$$

$\mathrm{proj}_L x = VV^T x$

$$= \sum_{i=1}^{m} \| x_i - x_0 - \mathrm{proj}_L (x_i - x_0) \|^2$$

$V = [v_1 \cdots v_k]$

$$= \sum_{i=1}^{m} \| x_i - x_0 - VV^T (x_i - x_0) \|^2$$

$$= \sum_{i=1}^{m} \| (I - vv^T)(x_i - x_0) \|^2$$

$R = I - vv^T$

$$= \sum_{i=1}^{m} \| R(x_0 - x_i) \|^2$$

Take gradient in x_0 (assuming x_0 min)

$$\nabla_{x_0} E(x_0, L) = \sum_{i=1}^{m} \nabla_{x_0} \| R(x_0 - x_i) \|^2$$

Gradient in x_0

$$= \sum_{i=1}^{m} \cancel{2} R^2 (x_0 - x_i) = 0$$

Last time $R^2 = R$ $\left[(I - vv^T)^2 = (I - vv^T) \right]$

Since $V^T V = I$

$$\sum_{i=1}^{m} (I - VV^T)(x_0 - x_i) = 0$$

$$(I - VV^T) \underbrace{\sum_{i=1}^{m} (x_0 - x_i)}_{y} = 0 \Rightarrow y \in L.$$

Hence $$\sum_{i=1}^{m} (x_0 - x_i) = y \in L$$

$$m x_0 - \sum_{i=1}^{m} x_i = y$$

$$\Rightarrow \quad x_0 = \underbrace{\frac{1}{m}\sum_{i=1}^{m} x_i}_{\text{centroid.}} + \underbrace{\frac{1}{m} y}_{\in L}$$

Choose $y = 0$ to complete proof

$$E(x_0, L) = \sum_{i=1}^{m} \|\underbrace{x_i - x_0}_{y_i} - \text{proj}_L \underbrace{(x_i - x_0)}_{y_i}\|^2$$

<u>Center Data</u>: Replace x_i with $x_i - x_0$

$$E(L) = \sum_{i=1}^{m} \| y_i - \mathrm{proj}_L\, y_i \|^2$$

Reduction to fitting a linear subspace

Since the centroid is optimal, we can center the data (replace x_i by $x_i - x_0$), and reduce to the problem of finding the optimal linear subspace L. Thus, we can consider the problem

$$\min_L E(L) = \sum_{i=1}^{m} \|x_i - \mathrm{Proj}_L x_i\|^2,$$

where the $\min_L$ is over k-dimensional linear subspaces L. We can write

$$L = \mathrm{span}\{v_1, v_2, \ldots, v_k\},$$

and treat the problem as optimizing over the orthonormal basis $v_1, v_2, \ldots, v_k$ of L.

The covariance matrix

Lemma 1. *The energy $E(L)$ can be expressed as*

$$E(L) = \mathrm{Trace}(M) - \sum_{j=1}^{k} v_j^T M v_j, \tag{1}$$

where M is the covariance matrix *of the data, given by*

$$M = \sum_{i=1}^{m} x_i x_i^T. \tag{2}$$

Note: We can write $M = X^T X$, where $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}^T$.

Proof: Let $x \in \mathbb{R}^n$, note

$$x x^T = \begin{bmatrix} x(1)^2 & x(1)x(2) & \cdots & x(1)x(n) \\ x(2)x(1) & x(2)^2 & \cdots & x(2)x(n) \\ \vdots & & \ddots & \vdots \\ x(n)x(1) & \cdots & & x(n)^2 \end{bmatrix}$$

$$\mathrm{Trace}(xx^T) = x(1)^2 + x(2)^2 + \cdots + x(n)^2$$
$$= \|x\|^2$$

Hence

$$\mathrm{Trace}(M) = \mathrm{Trace}\left(\sum_{i=1}^{m} x_i x_i^T\right)$$
$$= \sum_{i=1}^{m} \mathrm{Trace}(x_i x_i^T)$$
$$= \sum_{i=1}^{m} \|x_i\|^2$$

To prove theorem:

$$E(L) = \sum_{i=1}^{m} \|x_i - \mathrm{proj}_L x_i\|^2$$

Pythagorean Theorem

$$= \sum_{i=1}^{m} \left(\|x_i\|^2 - \|\mathrm{proj}_L x_i\|^2 \right)$$

$$= \sum_{i=1}^{m} \|x_i\|^2 - \sum_{i=1}^{m} \|\mathrm{proj}_L x_i\|^2$$

$$= \mathrm{Trace}(M) - \sum_{i=1}^{m} \sum_{j=1}^{K} (x_i^T v_j)^2$$

$$= \mathrm{Trace}(M) - \sum_{j=1}^{K} \sum_{i=1}^{m} v_j^T x_i x_i^T v_j$$

$$= \mathrm{Trace}(M) - \sum_{j=1}^{K} v_j^T \underbrace{\left(\sum_{i=1}^{m} x_i x_i^T \right)}_{M} v_j \qquad \square$$

Covariance Matrix

The covariance matrix

$$M = \sum_{i=1}^{m} x_i x_i^T = X^T X$$

is a positive semi-definite (i.e., $v^T M v \geq 0$) and symmetric matrix. Indeed, for a unit vector v we have

$$\underline{v^T M v} = \sum_{i=1}^{m} v^T x_i x_i^T v = \sum_{i=1}^{m} \underbrace{(x_i^T v)^2}_{\text{proj}_v x_i} \geq 0,$$

which is exactly the amount of *variation* in the data in the direction of v.

If v is an eigenvector with eigenvalue λ, then $\underline{Mv = \lambda v}$ and

$\|v\| = 1$

$$\lambda = v^T M v = \text{Variation in direction } v.$$

$v^T M v = v^T \lambda v = \lambda$

Covariance Matrix

Since the covariance matrix M is symmetric, it can be diagonalized:

$$M = PDP^T$$

where $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and

$$P = \begin{bmatrix} p_1 & p_2 & \cdots & p_n \end{bmatrix}.$$

We choose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and note that $p_1, p_2, \ldots, p_n$ are orthonormal eigenvectors of M, so

$$Mp_i = \lambda_i p_i.$$

Principal Component Analysis (PCA)

Theorem 2. *The energy $E(L)$ is minimized over k-dimensional linear subspaces $L \subset \mathbb{R}^n$ by setting*

$$L = span\{p_1, p_2, \ldots, p_k\}$$

and the optimal energy is given by

$$E(L) = \sum_{i=k+1}^{n} \lambda_i.$$

Note: The p_i are called the *principal components* of the data, and the λ_i are the principal values. The prinipal components are the directions of highest variation in the data.

Proof: By lemma, we can just focus on maximizing $\sum_{j=1}^{K} v_j^T M v_j$

over othonormal vectors $v_1, v_2, \dots, v_k$.

$$\sum_{j=1}^{k} v_j^T M v_j = \sum_{j=1}^{k} v_j^T P D P^T v_j$$

$$= \sum_{j=1}^{k} \left(v_j^T P D^{1/2}\right)\left(D^{1/2} P^T v_j\right)$$

$$D^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & & & 0 \\ & \lambda_2^{1/2} & & \\ & & \ddots & \\ 0 & & & \lambda_n^{1/2} \end{bmatrix}$$

$$= \sum_{j=1}^{k} \left(D^{1/2} P^T v_j\right)^T \left(D^{1/2} P^T v_j\right)$$

$$= \sum_{j=1}^{k} \left\| D^{1/2} P^T v_j \right\|^2$$

$$(*) \quad = \sum_{j=1}^{k} \sum_{i=1}^{n} \lambda_i \left(p_i^T v_j\right)^2$$

$$D^{1/2} P^T v_j = \begin{bmatrix} \lambda_1^{1/2} & & 0 \\ & \lambda_2^{1/2} & \\ 0 & & \ddots \quad \lambda_n^{1/2} \end{bmatrix} \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_n^T \end{bmatrix} v_j$$

$$= \begin{bmatrix} \lambda_1^{1/2} & & 0 \\ & \lambda_2^{1/2} & \\ 0 & & \ddots \quad \lambda_n^{1/2} \end{bmatrix} \begin{bmatrix} p_1^T v_j \\ p_2^T v_j \\ \vdots \\ p_n^T v_j \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1^{1/2} p_1^T v_j \\ \lambda_2^{1/2} p_2^T v_j \\ \vdots \\ \lambda_n^{1/2} p_n^T v_j \end{bmatrix}$$

$$(*) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{K} (p_i^T v_j)^2$$

$$= \sum_{i=1}^{n} \lambda_i \|\mathrm{proj}_L p_i\|^2$$

Hence

$$\sum_{j=1}^{k} v_j^T M v_j = \sum_{i=1}^{n} \lambda_i \underbrace{\|\mathrm{proj}_L p_i\|^2}_{0 \le a_i \le 1}$$

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \|\mathrm{proj}_L p_i\|^2$$

$$= \sum_{j=1}^{k} \underbrace{\sum_{i=1}^{n} (p_i^T v_j)^2}_{\|v\|^2 = 1} = \sum_{j=1}^{k} 1 = k$$

By HW 1 #6 $\sum_{i=1}^{\hat{}} \lambda_i \|\text{proj}_L p_i\|^2 \le \sum_{i=1}^{k} \lambda_i$

and by choosing $v_1 = p_1, v_2 = p_2, \dots, v_k = p_k$

we set

$$\|\text{proj}_L p_i\|^2 = \begin{cases} 1, & 1 \le i \le k \\ 0, & i \ge k+1 \end{cases}$$

and

$$\sum_{i=1}^{\hat{}} \lambda_i \|\text{proj}_L p_i\|^2 = \sum_{i=1}^{k} \lambda_i$$

$\blacksquare$

How many principal directions?

If we wish to capture $\alpha \in [0,1]$ fraction of the total variation in the data, we can choose k so that

$$\sum_{i=1}^{k} \lambda_i \geq \alpha \, \mathrm{Trace}(M).$$

$$\frac{\sum_{i=1}^{k} \lambda_i}{\mathrm{Trace}(M) = \sum_{i=1}^{n} \lambda_i} \geq \alpha$$

Intro to PCA Notebook: (.ipynb)