Mathematics of Image and Data Analysis
Math 5467

Principal Component Analysis

Instructor: Jeff Calder
Email: jcalder@umn.edu

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

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

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

and

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

Then we have

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

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

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

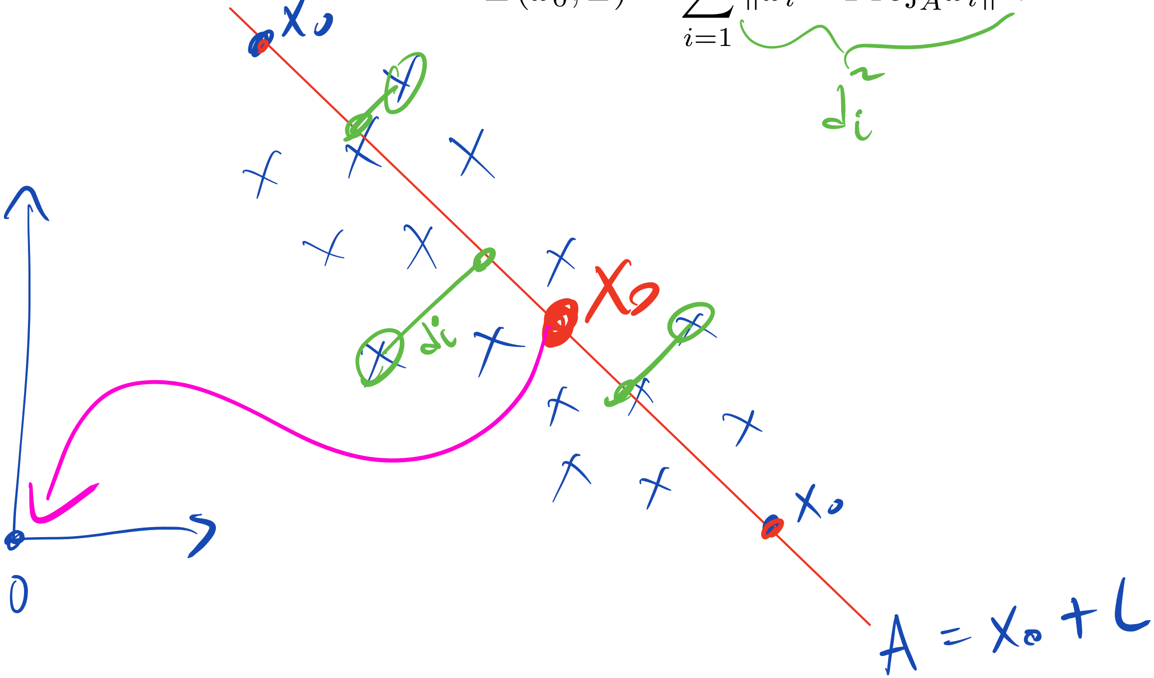
Also, for a symmetric matrix A

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

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

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



Optimizing over x_0

$\text{proj}_A x = x_0 + \text{proj}_L(x - x_0)$

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

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

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

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

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

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

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

residual operator

$$R = I - vv^T$$

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

$$0 = \nabla_{x_0} E(x_0, L) = \sum_{i=1}^{m} \nabla \| R(x_i - x_0) \|^2$$

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

$$R^2 = R$$

$$(I - vv^T)^2 = I - vv^T$$

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

$$Ry = 0 \ , \quad y = \sum_{i=1}^{m}(x_i - x_0)$$

$$(I - vv^T)y = 0 \quad \text{iff} \quad y \in L = \text{span}(v)$$

$$\downarrow$$

$$y = vv^T y$$

Choice $y = 0$

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

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

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

If $y \in L$, $y \neq 0$, then

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

$$x_0 = \underbrace{\frac{1}{m}\sum_{i=1}^{m} x_i}_{\text{centroid}} - y$$

$+ L$

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

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

Define $y_i = x_i - x_0$ (centering data).

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

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

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

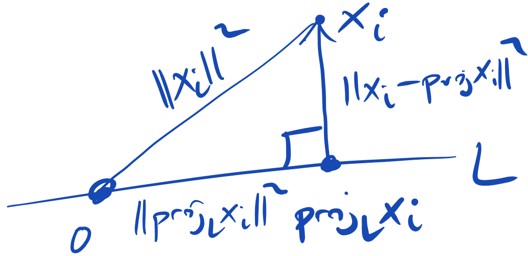
and treat the problem as optimizing over the orthonormal basis $v_1, v_2, \ldots, v_k$ of L.

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:

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

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

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

<u>Note</u>: $\mathrm{trace}(xx^T) = \mathrm{trace}\left(\begin{bmatrix} x(1)^2 & x(1)x(2) & \cdots & x(1)x(n) \\ x(1)x(2) & x(2)^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ x(n)x(1) & \cdots & \cdots & x(n)^2 \end{bmatrix}\right)$

$$= \|x\|^2$$

<u>First term</u>

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

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

<u>Second term</u>

$$
\begin{aligned}
\sum_{i=1}^{m} \|\mathrm{proj}_L x_i\|^2 &= \sum_{i=1}^{m} \sum_{j=1}^{k} (x_i^T v_j)^2 \\
&= \sum_{j=1}^{k} \sum_{i=1}^{m} (x_i^T v_j)(v_j^T x_i) \\
&= \sum_{j=1}^{k} \sum_{i=1}^{m} v_j^T (x_i x_i^T) v_j \\
&= \sum_{j=1}^{k} v_j^T \left(\sum_{i=1}^{m} x_i x_i^T \right) v_j \\
&= \sum_{j=1}^{k} v_j^T M v_j \qquad \blacksquare
\end{aligned}
$$

Covariance Matrix

$M^T = (X^T X)^T = X^T X$

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

$$v^T M v = \sum_{i=1}^{m} v^T x_i x_i^T v = \sum_{i=1}^{m} (x_i^T v)^2 \geq 0,$$

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

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

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

$$v^T M v = v^T \lambda v = \lambda v^T v = \lambda \underbrace{\|v\|^2}_{=1}$$

Covariance Matrix

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

$$M = \underline{PDP^T}$$

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

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

amount of variation missed.

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: We can consider maximizing

$$A = \sum_{j=1}^{k} v_j^T M v_j \quad \text{over } v_1, v_2, \ldots, v_k$$

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

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

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

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

$$D^{1/2} P^T v_j = \begin{bmatrix} \lambda_1^{1/2} & & 0 \\ & \lambda_2^{1/2} & \\ 0 & \ddots & \lambda_n^{1/2} \end{bmatrix} \begin{bmatrix} P_1^T \\ P_2^T \\ \vdots \\ P_n^T \end{bmatrix} v_j$$

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

$$= \left[\lambda_1^{1/2} p_1^T v_j \;\middle|\; \cdots \;\middle|\; \lambda_n^{1/2} p_n^T v_j \right]^T$$

$$\| D^{1/2} P^T v_j \|^2 = \sum_{i=1}^{n} \left(\lambda_i^{1/2} p_i^T v_j \right)^2 = \sum_{i=1}^{n} \lambda_i (p_i^T v_j)^2$$

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

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

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

$$0 \le a_i \le 1$$

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \sum_{j=1}^{k} (p_i^T v_j)^2$$

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

HW1 #6

$$\sum_{i=1}^{n} a_i \lambda_i \le \sum_{i=1}^{k} \lambda_i$$

Choice of $v_1 = p_1, v_2 = p_2, \ldots, v_k = p_k$

gives $a_i = \|\mathrm{proj}_L\, p_i\|^2 = \|p_i\|^2 = 1$

for $i \le k$.

If $i > k$ then $p_i \perp p_j$, $j \le k$

So $a_i = 0$ for $i > k$.

Since $\sum_{i=1}^{n} a_i \lambda_i = \sum_{i=1}^{k} \lambda_i$, this

choice is optimal. $\blacksquare$

PCA for dimension reduction

The steps for dimension reduction to $\mathbb{R}^k$ are outlined below. We assume we are given an $m \times n$ data matrix X

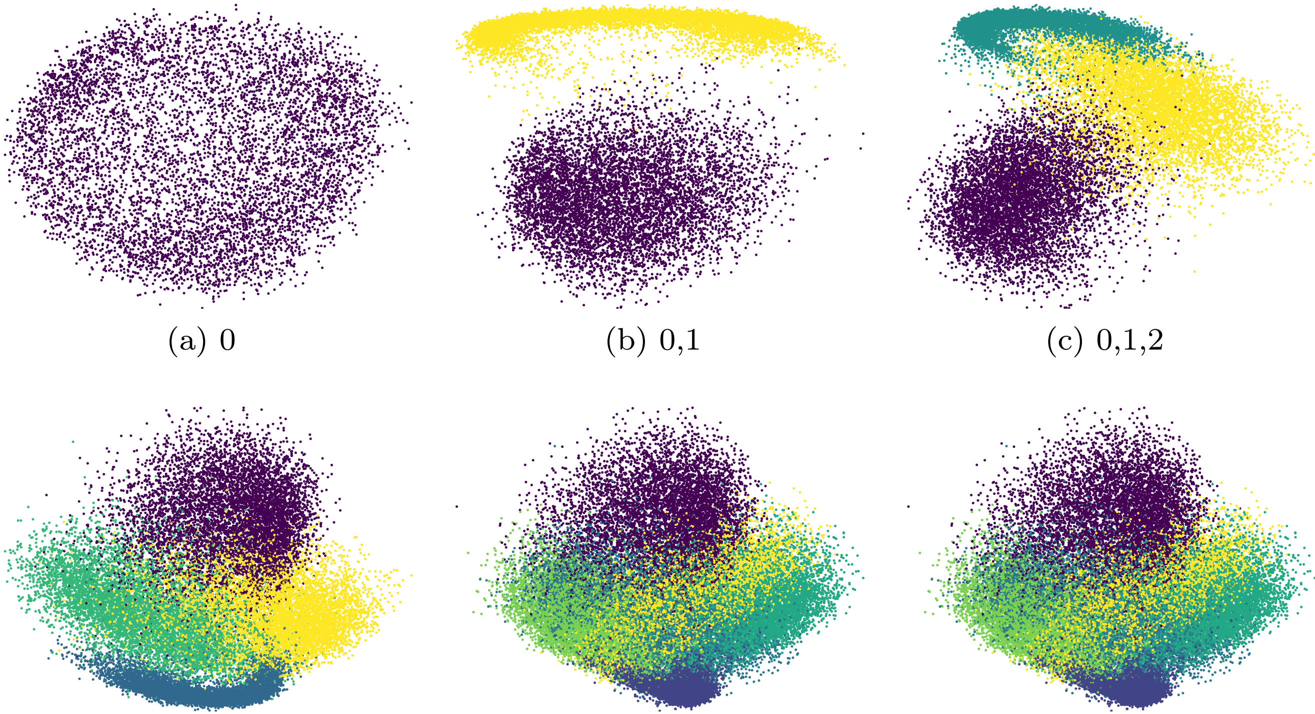
1. Compute the PCA covariance matrix $M = X^T X$, with the option of centering X first.

2. Compute the top k eigenvectors of M, and store them in a matrix P of size $n \times k$.

3. Compute the PCA dimension reduced dataset $B = XP$. $\in \mathbb{R}^{m \times k}$

$$B = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} \begin{bmatrix} p_1 & p_2 & \cdots & p_k \end{bmatrix}$$

Example on MNIST



(a) 0 (b) 0,1 (c) 0,1,2

(d) 0,1,2,3 (e) 0,1,2,3,4 (f) 0,1,2,3,4,5

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

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

Intro to PCA Notebook: (.ipynb)