Eckart-Young

Wednesday, June 24, 2020 20:02

This lecture is about PCA (Principal Component Analysis)
A major tool in understanding a matrix of data.

From SVD, we know that any matrix A can be broken into r rank-1 pieces as:

$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \dots + \sigma_r u_r v_r^T$$

rank of A

u's are orthonormal and so are v's

Let's say that the important facts about a matrix are in its largest k singular values.

$$A_k = U_k \Sigma_k V_k^T = \sigma_1 u_1 v_1^T + \dots + \sigma_k u_k v_k^T$$

<u>Claim</u>: A_k is the best approximation of A with rank k.
This tells us the importance of SVD!

In other words, if rank$(B) = k$ for some matrix B, then

$\boxed{\|A-B\| \geq \|A-A_k\|}$ → This is true for the 3 norms below.

This is the Eckart-Young theorem.

We should talk about norms first.

→ Some possible norms for matrices that can be computed by the singular values:

1. $\|A\|_2 = \sigma_1$ (L2-norm, the largest singular value)

For a vector v:

$\|v\|_2 = \ell^2$ norm $= \sqrt{v_1^2 + \dots + v_n^2}$

$\|v\|_1 = \ell^1$ norm $= |v_1| + \dots + |v_n|$

$\|v\|_\infty = \ell^\infty$ norm $= \max |v_i|$

Properties of any vector norm:

$\|v\|_\infty$ ...

Properties of any vector norm:

1) $\|cv\| = |c|\,\|v\|$ — the norm is always +ve

2) $\|v+w\| \leq \|v\| + \|w\|$

For a matrix, this means $\sigma_1(A+B) \leq \sigma_1(A) + \sigma_1(B)$.

→ Let $B = A^TA$

Let $\lambda_1, \dots, \lambda_n$ be eigenvalues of B and $\{e_1, \dots, e_n\}$ be an orthonormal basis. We can do this since B is real & symmetric.

Let $x = a_1e_1 + \dots + a_ne_n$

$\|x\| = \sqrt{\langle \sum a_ie_i, \sum a_ie_i\rangle} = \sqrt{\sum a_i^2}$ since e_i are orthonormal.

$$Bx = B\left(\sum a_ie_i\right) = \sum a_i Be_i = \sum a_i \lambda_i e_i$$

$$\|Ax\| = \sqrt{\langle Ax, Ax\rangle} = \sqrt{\langle x, A^TAx\rangle} = \sqrt{\langle x, Bx\rangle}$$

$$= \sqrt{\langle \sum a_ie_i, \sum \lambda_i a_i e_i\rangle} = \sqrt{\sum a_i^2\lambda_i}$$

$\because \|x\| = \sqrt{\sum a_i^2}$, we have:

$$\|Ax\| \leq \sqrt{\lambda_1}\,\|x\|$$

Largest eigenvalue of B

Now take $x = e_1 \Rightarrow \|Ae_1\| = \sqrt{\lambda_1}$

So e_1 maximizes $\|Ax\|$

We've seen before that eigenvalues of A^TA are squares of the singular values of A. So we have $\|Ax\| = \sigma_1(A)$.

So, $\|(A+B)x\| = \|Ax + Bx\|$

Since Ax, Bx are vectors, we have

$$\|Ax + Bx\| \leq \|Ax\| + \|Bx\|$$

This also shows that: $\sigma_1(A+B) \leq \sigma_1(A) + \sigma_1(B)$.

2. $\|A\|_F$ (Frobenius) $= \sqrt{a_{ij}^2}$

3. $\|A\|_{nuclear} = \sigma_1 + \dots + \sigma_r$ (sum of singular values)

3. $\|A\|_{\text{nuclear}} = \sigma_1 + \cdots + \sigma_r$ (sum of singular values)

↳ This won the Netflix competition (partly about psychology) — To fill in missing data in the matrix.

Now used for MRI

→ e.g. $\Sigma = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\qquad \Sigma_2 = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

closest rank-2 approximation to Σ.

For some $A = U\Sigma V^T$, the singular values would be that of Σ, so the problem is essentially orthogonally invariant.

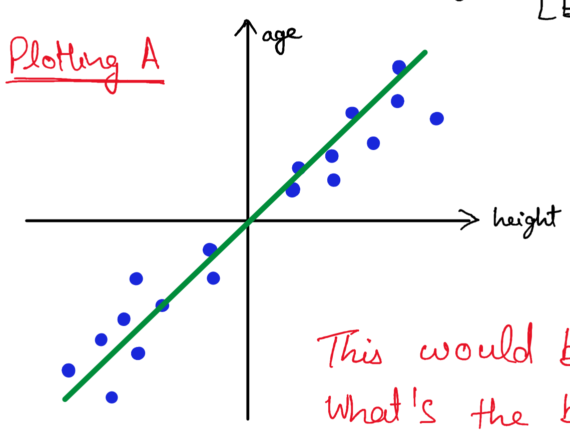
→ Let's say we have some data points in the plane:

$$A_0 = \begin{bmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_N \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{matrix} \leftarrow \text{heights} \\ \leftarrow \text{ages} \end{matrix}$$

So, we have N points in 2D.

The first thing a statistician does is make mean 0. So, we work with $A = A_0 - \begin{bmatrix} \mathbb{E}h & \cdots & \mathbb{E}h \\ \mathbb{E}a & \cdots & \mathbb{E}a \end{bmatrix}$, so each row of A adds to 0.

Plotting A



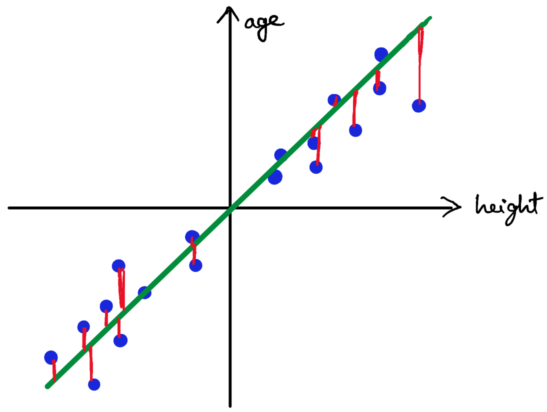
We're looking for the best line that describes the relat^{n}ship b/w height & age.

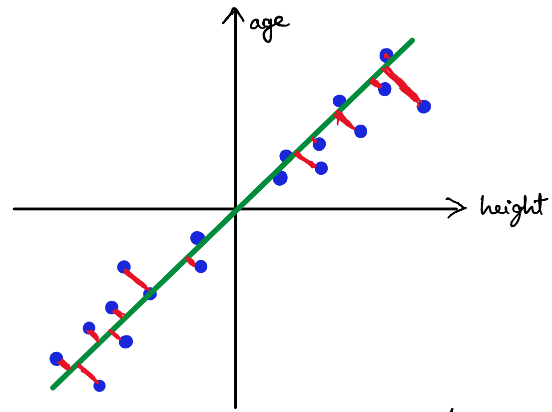
This would be an example problem in PCA: What's the best linear relation?

It looks like a thing least^2 would do, but it's not!

the best line of least^2 would be the one minimizing the (error)^2 → $\min \|b - Ax\|^2$

The (error) $\rightarrow$ $\min \| b - Ax \|$





Least ^{2} minimizes sum of sq of distances parallel to axis.

$\Downarrow$

Solve $A^T A \hat{x} = A^T b$

called normal equations,

"Regression" in statistics language.

PCA minimizes sum of sq of distances perpendicular to the line.

$\Downarrow$

Involves SVD, singular values.

To see the regression approach:

let heights $= h_1, \dots, h_n$

ages $= a_1, \dots, a_n$

Best line : $y = mx + c \rightarrow = 0$ as mean was made 0

Now we minimize:

$$\sum_{i=1}^{n} (a_i - m h_i)^2 = \Bigg\| \underbrace{\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}}_{b} - \underbrace{m \overbrace{\begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}}^{x}}_{Ax} \Bigg\|^2$$

We'll see later how $A^T A \hat{x} = A^T b$ helps here.

The second thing a statistician will look at the variance, in fact in this example the covariance matrix (2×2)

$\hookrightarrow$ sample covariance, empirical

$$= \frac{A A^T}{N-1}$$