Lecture 23

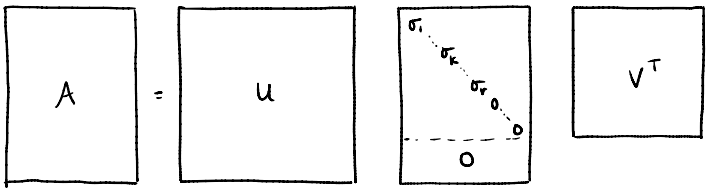
Q Let $A \in \mathbb{R}^{m \times n}$, $m \geq n$, rank r.
Let $k < r$. Recall the rank k approximation $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

What is the residual $\|A - A_k\|_2$? (Note: 2-norm!)

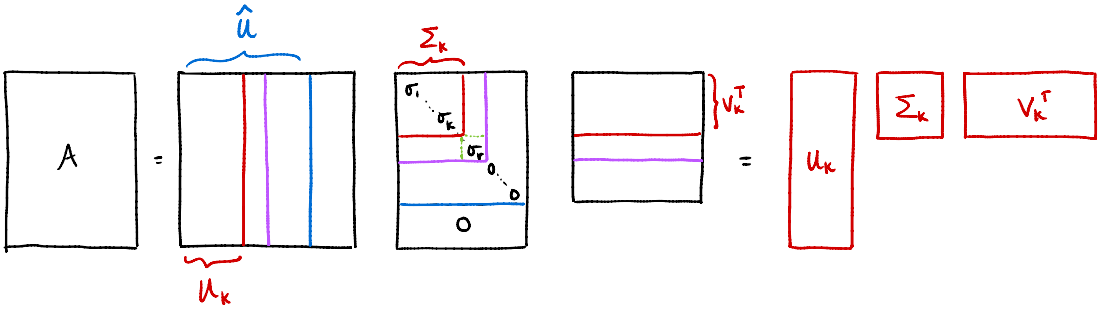
A. $\|A - A_k\|_2 =$ "biggest stretch factor of $A - A_k$" $= \sigma_{k+1}$. $\quad A - A_k = \sum_{i=k+1}^{r} \sigma_i u_i v_i^T$

Recall $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \qquad$ rank $r \leq n \leq m$

Visualization of low-rank approximation:



So $A - A_k$ has singular values $\sigma_{k+1}, \ldots, \sigma_r$.



Aside Uniqueness of SVD? $\quad A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \qquad A \in \mathbb{R}^{m \times n}$, rank r

- Σ is determined by A, since we require $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r > 0$.
- For $1 \leq i \leq r$, we may replace $u_i v_i^T$ with $(-u_i)(-v_i)^T$ $\quad (-1)^2 = 1$.
  - we can't scale them b/c they're supposed to be length 1!
  - $u_i v_i^T$ are linearly independent in a space of matrices since their sum is an exactly rank r matrix A.
- For $j \geq r+1$, we have more freedom
  - permutation of basis vectors u_j, v_j for $j \geq r+1$.
  - also must both u_j, v_j by (-1), as above

Condition # and choosing the rank k for approximation:

Recall $\kappa(A) = \sigma_1/\sigma_r$, $\kappa(A) \approx 10^k \Rightarrow$ lose k digits of precision

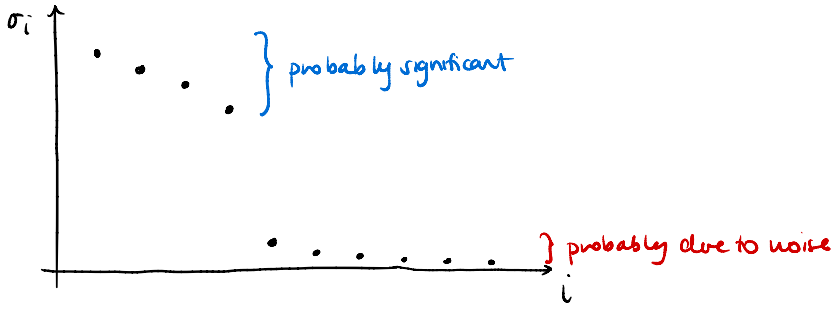
$k = \log_{10} \kappa(A)$

Imagine we are given a matrix $A \in \mathbb{R}^{10\times 10}$ that really should be lower rank but has noise + so is actually nonsingular

↳ nonsingular is "generic": $\det(A) = 0$ is very more special than $\det(A) \neq 0$.

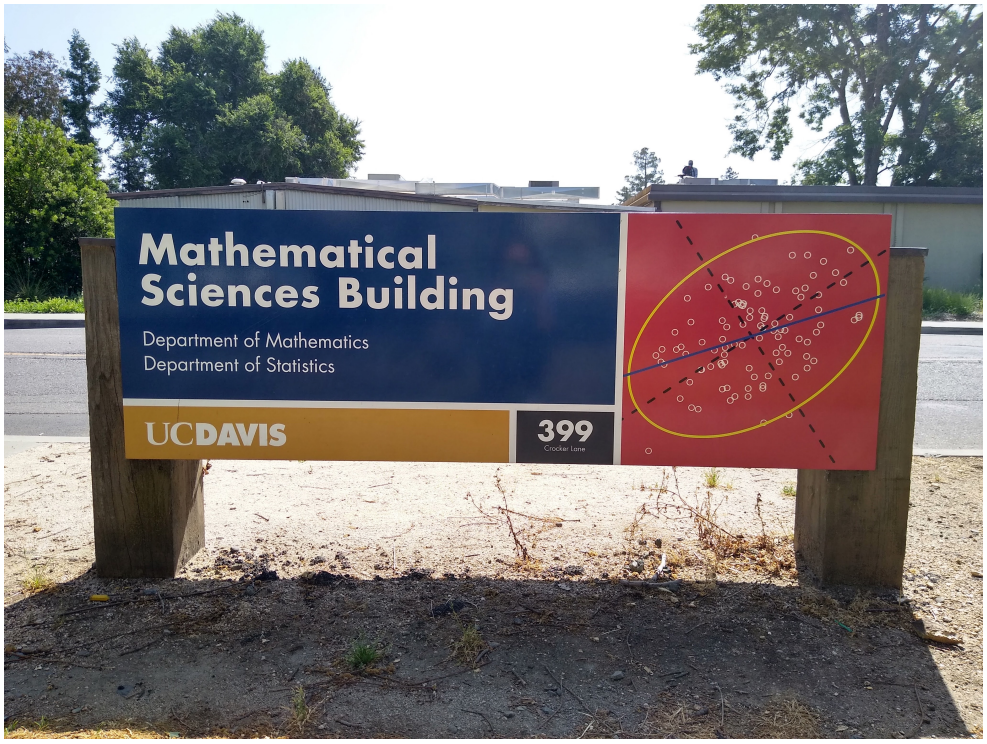
We want to capture the essence of the data in lower rank.

Plot the singular values:



Compare: $\kappa(A) = \sigma_1/\sigma_{10}$ large $\kappa(A_4)$ is much smaller.

So we might choose to focus on the approximation A_4.



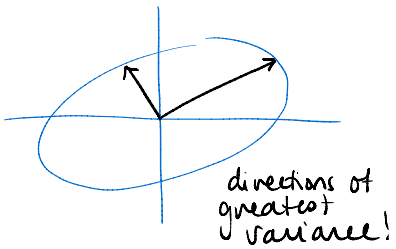
PCA = Principal Component Analysis is the same concept as SVD

PCA is a data analysis technique that uses orthogonal transformations to convert a set of observations of maybe correlated variables into a set of linearly uncorrelated variables called principal components.

"important directions", ie better linear combo of your original variables to study.

eg. the sign outside MSB (cloud of data points roughly in the shape of an ellipse in 2D)

high-dimensional data presumably normally distributed $\xrightarrow{\text{PCA}}$ fit a 2D ellipsoid (ie. ellipse) to data



$A = [a_1\ a_2\ \cdots\ a_{100}] \xrightarrow{\text{SVD}} A_2 = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$

Key idea: Obtain lower-dimensional visualization of a high-dimensional set of data, while preserving as much of the variance as possible.

eg. 2000 students, 8 homework scores b/w 0 and 100.

Many names: Invented by Pearson, Hotelling (1901, 1930s)

- Hotelling transform
- Karhunen-Loève transform (KLT) in signal processing
- proper orthogonal decomposition (POD) in mechE
- eigenvalue decomp of the symm. matrix A^TA ...
- empirical orthogonal functions (EOF), spectral decomposition...

eg. (small example) Measure temperature (T), air pressure (P), humidity (H)

* I am not a meteorologist

$\Rightarrow$ clusters of dots in 3D. (say, 2000 data points in $\mathbb{R}^3$)

Change of variables: $(T, P, H) \rightsquigarrow (T-\bar{T}, P-\bar{P}, H-\bar{H})$ so that we are centered around $(0,0,0) \in \mathbb{R}^3$

$\tilde{X} = [\tilde{x}_1 \; \tilde{x}_2 \cdots \tilde{x}_{2000}] \in \mathbb{R}^{3\times 2000}$ (very likely rank 3) $d=3$, $n=2000$

From the (sample) covariance matrix ← will explain later

$$S := \frac{1}{n-1}\tilde{X}\tilde{X}^T = \Phi\Lambda\Phi^{-1}$$

(symmetric matrix $\Rightarrow$ SVD is eigenvalue decomp.)

$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ (decreasing)

note. λ_i here are scaled versions of $\sigma_i(\tilde{X})$: $\lambda_i = \sigma_i^2/n$

$\tilde{X} = USV^T \Rightarrow \tilde{X}\tilde{X}^T = USV^TVSU^T = U\cdot S^2 \cdot U^T$ (the basis Φ is very related to U)

$\Rightarrow \Phi = [\varphi_1 \; \varphi_2 \; \varphi_3] \in \mathbb{R}^d = \mathbb{R}^3$ an orthonormal basis of $\mathbb{R}^d$.

- $\varphi_i \perp \varphi_j$ when $i \neq j$ $\Rightarrow$ uncorrelated
- φ_1 would be (a normalized version of) some linear combo of T, P, H. (unnormalized lengths depend on the units used)

↪ "feature extraction": maybe $\varphi_1 \sim T + H$.

$\Rightarrow$ T, H vary together very strongly!

cf. Exercise 3 on HW07! Face: pixels near eyes should be related. in these, d could be $\gg n$.

Some background statistics (what is this S?)

Given: set of observations $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ each $\vec{x}_i \in \mathbb{R}^d$ eg. $d = 128 \times 128$ pixels!

(n data points)

Let $X := [x_1 \ \cdots \ x_n]$

Mean/average of data set: $\bar{X} = \frac{1}{n} \sum_{j=1}^{n} x_j$ simultaneously averaging all components, independently

Define the centered data matrix $\tilde{X} = [\underbrace{x_1 - \bar{x}_1}_{\tilde{x}_1} \quad x_2 - \bar{x}_2 \quad \cdots \quad x_n - \bar{x}_n]$

$= [\tilde{x}_1 \ \cdots \ \tilde{x}_n]$

Aside Covariance b/w variables V, W w/ data (a_i, b_i) $i \in \{1, \dots, N\}$ is

$$\text{cov}_{V,W} = \frac{\sum (a_i - \bar{a})(b_i - \bar{b})}{N-1}$$ ← typo in bound in class!

(Sample) covariance matrix $S = \frac{1}{n} \tilde{X} \tilde{X}^T \in \mathbb{R}^{d \times d}$

d >> n in eg.

$$S = \frac{1}{n} [\tilde{x}_1 \ \tilde{x}_2 \ \cdots \ \tilde{x}_n] \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T$$

n-1

$$\vec{V} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \quad \frac{1}{n} \vec{V} \vec{V}^T = \frac{1}{n} \begin{bmatrix} v_1^2 & v_1 v_2 & v_1 v_3 \\ v_2 v_1 & v_2^2 & v_2 v_3 \\ v_3 v_1 & v_3 v_2 & v_3^2 \end{bmatrix} \Rightarrow \text{eg. } S_{12} = \frac{1}{n} \sum_{i=1}^{n} x_{i1} x_{i2}$$

the average $v_1 \cdot v_2$ value among all the vectors V.

~ covariance is mutual correlation b/w 1st and 2nd entries of each vector