8. Singular Value Decomposition (SVD)

Σ^T –

- Columns of V are orthogonal eigenvectors of A^TA $v = \sigma^2 v$
- $Av=\sigma u$ gives orthonormal eigenvectors u of AA^T U
- σ^2 = eigenvalue of A^TA = eigenvalue of $AA^T \neq 0$ > 0
- A = (rotation)(stretching)(rotation) $U\Sigma V^T$ for every A

- Why is the SVD so important?
  - It separates the matrix into rank one pieces like the other factorizations A=LU, A=QR, $S=Q\Lambda Q^T$ ✓ k
  - Those pieces come in order of importance $\sigma_1 > \sigma_2 > \cdots > \sigma_r$
  - First piece $\sigma_1 u_1 v_1^T$ is the closest rank one matrix to A
  - Sum of the first k pieces is best possible for rank k

$Ax = \lambda x$

$(xy^T)x = x(y^Tx) = \lambda x$

$|\lambda_1| = |y^Tx| \le \sigma_1 = \|y\|\|x\|$

Schwartz inequality

$Av_i = \sigma_i u_i$ $\qquad$ $2 \times n$

$u_i = \frac{Av_i}{\sigma_i}$ $\qquad$ AA^T $\qquad$ $A^TA = n \times n$, $AA^T = 2 \times 2$

- Example

Find the matrices $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}$ for $\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$.

$A^TA = \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}$

$\lambda_1 + \lambda_2 = 50$, $\lambda_1\lambda_2 = 225$ $\quad$ 45, 5

$$\mathbf{U} = \frac{1}{\sqrt{10}}\begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \sqrt{45} & \\ & \sqrt{5} \end{bmatrix}, \mathbf{V} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

$A =$ SVD

$$\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T = \mathbf{A}$$

$A = U\Sigma V^T = Q\Lambda Q^T$

$U = V = Q$, $\Sigma = \Lambda$

- If $S=Q\Lambda Q^T$ is symmetric positive definite, what is its SVD?
- If $S=Q\Lambda Q^T$ has a negative eigenvalue($Sx=-\alpha x$) $= \alpha(-x)$, what is the singular value and what are the vectors v and u? U or V $\quad$ $\sigma = +\sigma > 0$
- If A=Q is an orthogonal matrix, why does every singular value equal 1? $\quad$ $Q^TQ = I = A^TA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow \Sigma = I$
- Why are all eigenvalues of a square matrix A less than or equal to σ_1? $\quad$ $\|Ax\| = \|U\Sigma V^Tx\| = \|\Sigma V^Tx\| \le \sigma_1 \|V^Tx\| = \sigma_1\|x\|$ $\qquad$ $\|Ax\| = |\lambda|\|x\| \le \sigma_1\|x\|$
- If $A=xy^T$ has rank 1, what are u_1, v_1, σ_1? Check that $|\lambda_1| \le \sigma_1$ $\quad$ $|\lambda| \le \sigma_1$

$xy^T = \frac{x}{\|x\|}(\|x\|\|y\|)\frac{y^T}{\|y\|} = u_1\sigma_1 v_1^T$

$Ax = \lambda x$

$(xy^T)x = x(y^Tx) = \lambda x$

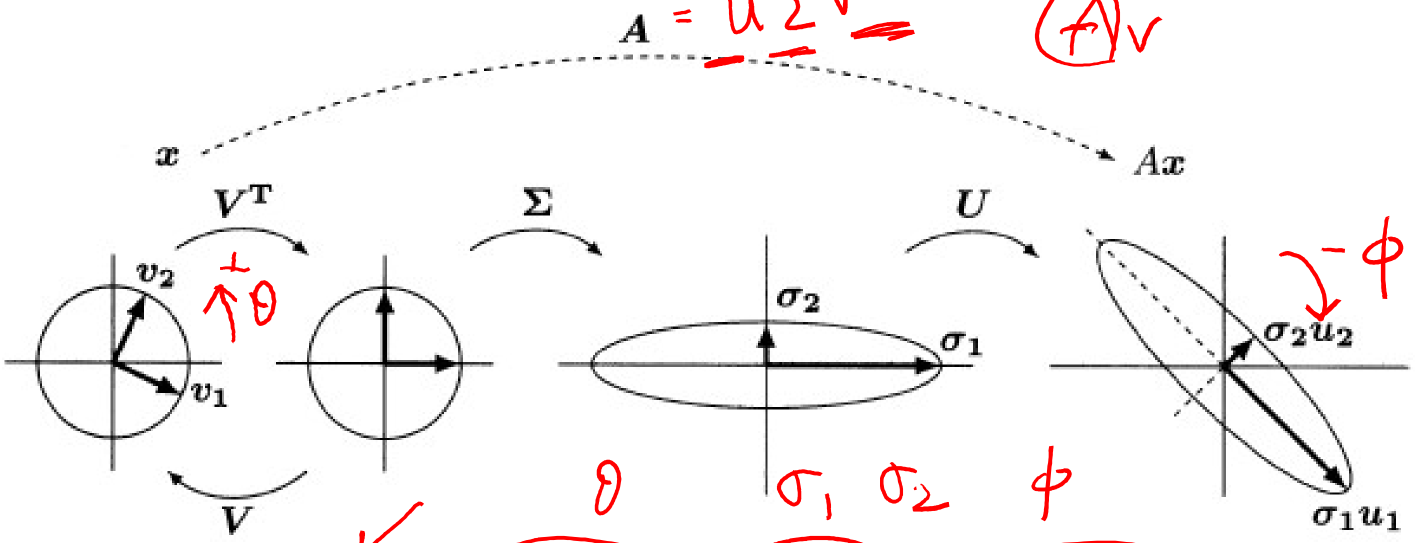
- If A is m by n and B is n by m, then AB and BA have the same nonzero eigenvalues

$AB \neq BA$

- Geometry of SVD

$A = U\Sigma V^T$

Av



θ σ_1 σ_2 ϕ

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \underbrace{\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix}\begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}}_{\text{4 parameters: two angles, two numbers}}$$

- First singular vector $\mathbf{v}_1$

$\|A\| =$ Maximize the ratio $\dfrac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \rightarrow$ The maximum is σ_1 at the vector $\mathbf{x} = \mathbf{v}_1$

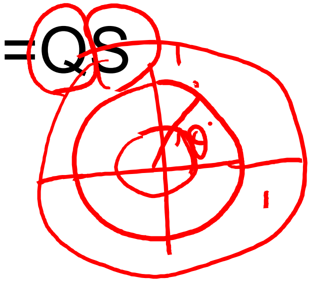
largest growth factor

$$\left(\frac{b}{a}\right)' = \frac{a'b - ab'}{a^2} = 0$$

$\Rightarrow$ Find the maximum value λ of $\dfrac{\|\mathbf{Ax}\|^2}{\|\mathbf{x}\|^2} = \dfrac{(\mathbf{Ax})^T \mathbf{Ax}}{\mathbf{x}^T \mathbf{x}} = \dfrac{\mathbf{x}^T \mathbf{Sx}}{\mathbf{x}^T \mathbf{x}} \rightarrow 2\mathbf{Sx} = 2\lambda \mathbf{x}$

Maximize $\dfrac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}$ under the conditon $\mathbf{v}_1{}^T \mathbf{x} = 0 \rightarrow$ The maximum is σ_2 at $\mathbf{x} = \mathbf{v}_2$

- Polar decomposition: A= UΣV^T= (UV^T)(VΣV^T)=QS



$$\underbrace{x+iy}_{\text{complex number}} = \underbrace{re^{i\theta}}_{\text{polar form}} \rightarrow \begin{cases} e^{i\theta} : \text{orthogonal matrix} \\ r \geq 0 : \text{positive semideinite matrix} \end{cases}$$

$$U\Sigma V^T = \quad \mathbf{A} = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} = \underbrace{\mathbf{Q}}_{\text{rotation}} \underbrace{\mathbf{S}}_{\text{stretch}} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} \sqrt{5} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

11. Norms of Vectors and Matrices

– The norm of a nonzero vector v is a positive number ||v||

– That number measures the “length” of the vector

size

every norm for vectors or functions or matrice must share these two properties

s

of the absolute value $|c|$ of a number

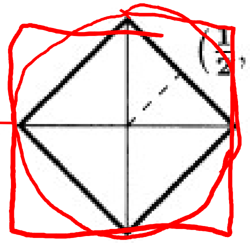
$$\text{All norms}\begin{cases}\text{multiply } \mathbf{v} \text{ by } c \text{ (rescaling)} \rightarrow \|c\mathbf{v}\| = |c|\|\mathbf{v}\| \\ \text{add } \mathbf{v} \text{ to } \mathbf{w} \text{ (Triangle inequality)} \rightarrow \|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\|\end{cases}$$

$$\begin{cases} l^2 \text{ norm} = \text{Euclidean norm}: \ \|\mathbf{v}\|_{l2} = \sqrt{|v_1|^2 + \cdots + |v_n|^2} \\ l^1 \text{ norm} = 1\text{-norm}: \ \|\mathbf{v}\|_{l1} = |v_1| + \cdots + |v_n| \\ l^\infty \text{ norm} = \text{max norm}: \ \|\mathbf{v}\|_{l\infty} = \text{maximum of } |v_1|, \ldots, |v_n| \end{cases}$$

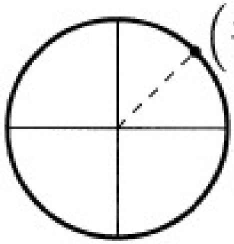
$$\|\mathbf{v}\|_p = \left(|v_1|^p + \cdots + |v_n|^p\right)^{1/p}$$

- Important vector norms and a failure

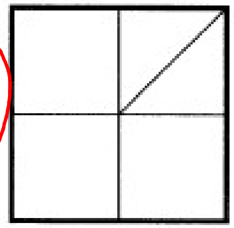
ℓ^1 norm
$|v_1| + |v_2| \le 1$
diamond



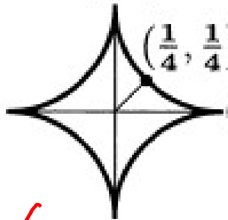
ℓ^2 norm
$v_1^2 + v_2^2 \le 1$
circle



ℓ^∞ norm
$|v_1| \le 1, |v_2| \le 1$
square



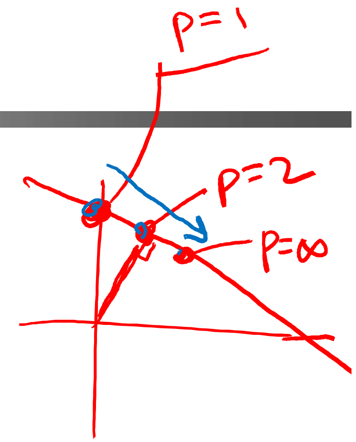
$\ell^{1/2}$ norm
$\sqrt{|v_1|} + \sqrt{|v_2|} \le 1$
not convex



$p = 1 \to \infty$ blow-up

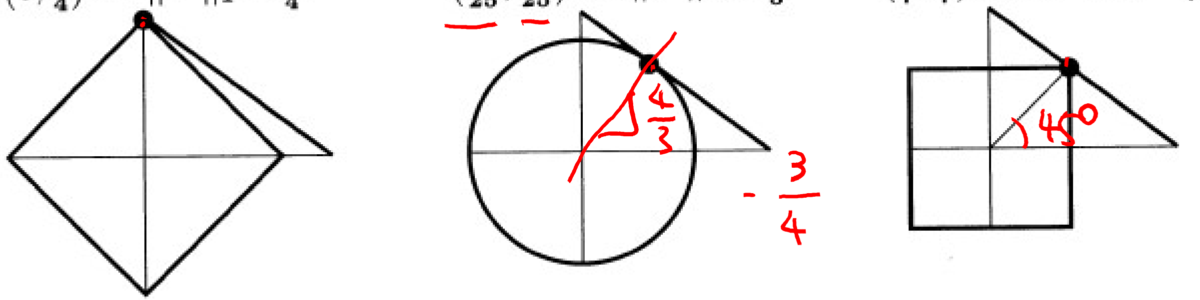
$p < 1$ X

- Minimum of $||v||_p$ on the line $a_1v_1 + a_2v_2 = 1$



Minimize $\|\|\boldsymbol{v}\|\|_p$ among vectors (v_1, v_2) on the line $3v_1 + 4v_2 = 1$

$(\mathbf{0}, \frac{1}{4})$ has $||\boldsymbol{v}^*||_1 = \frac{1}{4}$ $\quad$ $(\frac{3}{25}, \frac{4}{25})$ has $||\boldsymbol{v}^*||_2 = \frac{1}{5}$ $\quad$ $(\frac{1}{7}, \frac{1}{7})$ has $||\boldsymbol{v}^*||_\infty = \frac{1}{7}$



- Inner products and S=norm

$$\text{Inner product} = \text{length squared}: \mathbf{v}\cdot\mathbf{v} = \mathbf{v}^T\mathbf{v} = \|\mathbf{v}\|^2$$
$$\text{Angle } \theta \text{ between vector } v \text{ and } w: \mathbf{v}^T\mathbf{w} = \|\mathbf{v}\|\|\mathbf{w}\|\cos\theta$$
$$\rightarrow \begin{cases} \text{Cauchy-Schwarz}: |\mathbf{v}^T\mathbf{w}| \le \|\mathbf{v}\|\|\mathbf{w}\| \\ \text{Triangle Inequality}: \|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\| \end{cases}$$

Choose any symmetric positive definite matrix $\mathbf{S}$

$\|\mathbf{v}\|_S^2 = \mathbf{v}^T\mathbf{S}\mathbf{v}$ gives a norm for $\mathbf{v}$ in $\Re^n$ (called the S-norm)

$(\mathbf{v},\mathbf{w})_S = \mathbf{v}^T\mathbf{S}\mathbf{w}$ gives the S-inner product for $\mathbf{v},\mathbf{w}$ in $\Re^n$

Norm of Matrices

- Frobenius Norm
- Matrix Norm ||A|| from vector norm ||v||
- Nuclear Norm

$\sigma_1, \cdots, \sigma_r$ (SVD)

$$
\begin{cases}
\|\mathbf{A}\|_F = \sqrt{|a_{11}|^2 + \cdots + |a_{1n}|^2 + \cdots + |a_{mn}|^2} \\
\|\mathbf{A}\|_F = \|\mathbf{U}\Sigma\Sigma^T\|_F = \|\Sigma\mathbf{V}^T\|_F = \|\Sigma\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} \\
\|\mathbf{A}\|_F^2 = \text{trace of } \mathbf{A}^T\mathbf{A} = \text{sum of eigenvalues} = \sigma_1^2 + \cdots + \sigma_r^2
\end{cases}
$$

$$
\begin{cases}
\|\mathbf{A}\|_2 = \max_{\mathbf{v}\neq 0} \dfrac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|} = \text{largest growth factor} = \sigma_1 \\
l^2 \text{ norm}: \|\mathbf{A}\|_2 = \text{largest singular value } \sigma_1 \text{ of } \mathbf{A} \\
l^1 \text{ norm}: \|\mathbf{A}\|_1 = \text{largest } l^1 \text{ norm of the columns of } \mathbf{A} \\
l^\infty \text{ norm}: \|\mathbf{A}\|_\infty = \text{largest } l^1 \text{ norm of the rows of } \mathbf{A}
\end{cases}
$$

$$\|\mathbf{A}\|_N = \sigma_1 + \cdots + \sigma_r = \text{trace norm}$$

9. Principal Components

- major tool in understanding a matrix of data
- Eckart-Young low rank approximation theorem
  - The norm of $A-A_k$ is below the norm of all other $A-B_k$
  - $A_k = \sigma_1 u_1 v_1^T + \ldots + \sigma_k u_k v_k^T$
- Frobenius norm squared = sum of squares of all entries

$$\text{Eckart-Young: If } \mathbf{B} \text{ has rank } k\text{, then } \underbrace{\|\mathbf{A}-\mathbf{B}\|}_{\text{error}} \geq \|\mathbf{A}-\mathbf{A}_k\|$$

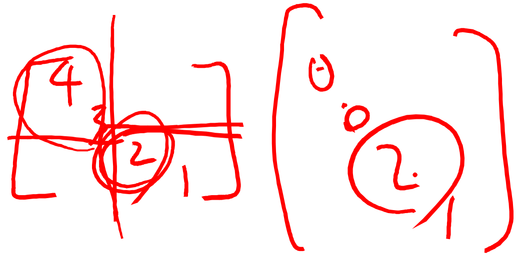
$$A \rightarrow \mathbf{A}_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^T \leftarrow \text{SVD}$$

$$\begin{cases} \text{Spectral norm: } \|\mathbf{A}\|_2 = \max\limits_{\mathbf{x}\neq 0} \dfrac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sigma_1 \\ \text{Frobenius norm: } \|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} \\ \text{Nuclear norm: } \|\mathbf{A}\|_N = \sigma_1 + \cdots + \sigma_r \end{cases}$$

Eckart-Young Theorem

$k=2$
$\sigma_{k+1}=\sigma_3$

- Best approximation by A_k



1 Eckart - Young in L^2:

If $\operatorname{rank}(\mathbf{B}) \le k$, then $\|\mathbf{A}-\mathbf{B}\| = \max_{\mathbf{x}\neq 0} \dfrac{\|(\mathbf{A}-\mathbf{B})\mathbf{x}\|}{\|\mathbf{x}\|} \ge \sigma_{k+1} = \|A - A_k\|$

Eckart - Young in the Frobenius norm:

2 If $\mathbf{B}$ is closest to $\mathbf{A}$, then $\mathbf{U}^T\mathbf{B}\mathbf{V}$ is closest to $\mathbf{U}^T\mathbf{A}\mathbf{V}$

$$\mathbf{B} = \mathbf{U}\begin{bmatrix} \underbrace{\mathbf{D}}_{k\times k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{V}^T, \mathbf{A} = \begin{bmatrix} \mathbf{L}+\mathbf{E}+\mathbf{R} & \mathbf{F} \\ \mathbf{H} & \mathbf{G} \end{bmatrix}$$

The matrix $\mathbf{D}$ must be the same as $\mathbf{E} = diag(\sigma_1, \ldots, \sigma_k)$

The singular values of $\mathbf{H}$ must be the smallest $(n-k)$ singular values of $\mathbf{A}$

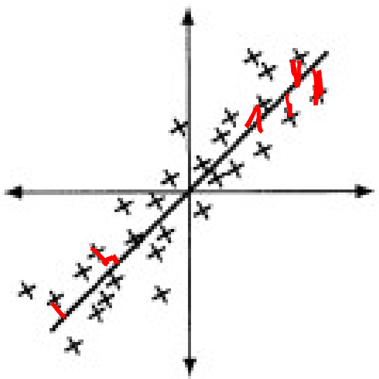
The smallest error $\|\mathbf{A}-\mathbf{B}\|_F$ must be $\|\mathbf{H}\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}$

Principal Component Analysis

↔ Least²

- Understand n sample points in m-dimensional space
- Data matrix A_0: n samples, m variables
  - Find the average (the sample mean) along each row of A_0
  - Subtract that mean from m entries in the row
  - Centered matrix $A=A_0$-(mean)
  - How will linear algebra find that closest line through (0,0)? It is in the direction of the first singular vector u_1 of A

age
height



A is $\mathbf{2} \times \boldsymbol{n}$ (large nullspace)

AA^{T} is $\mathbf{2} \times \mathbf{2}$ (small matrix)

$A^{\mathrm{T}}A$ is $\boldsymbol{n} \times \boldsymbol{n}$ (large matrix)

Two singular values $\boldsymbol{\sigma_1} > \boldsymbol{\sigma_2} > \mathbf{0}$

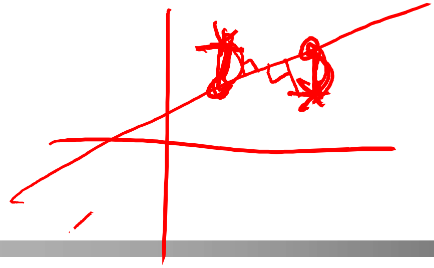
Least^2 $\longleftrightarrow$ PCA

$\min \|\underline{b} - Ax\|^2$

$\cdot\, Cx + D$

(age) = $\square$(height)

ratio : σ_1



- Statistics behind PCA
  - Variances: diagonal entries of the matrix AA^T
  - Covariances: off- diagonal entries of the matrix AA^T ✓
  - Sample covariance matrix: $S=AA^T/(n-1)$
- Geometry behind PCA
  - Sum of squared distances from the data points to the line is a minimum
- Linear algebra behind PCA
  - Singular values σ_i and singular vectors u_i of A
  - Total variance:

$$T = \frac{\|\mathbf{A}\|_F^2}{n-1} = \frac{\sigma_1^2 + \cdots + \sigma_r^2}{n-1}$$

$\|A\| = \|A_0 - \text{mean}\|$

separate each col a_j of A

$$\sum \|a_j\|^2 = \underbrace{\sum |a_j^T u_1|^2}_{\underset{\max}{u_1^T(A^TA)u_1}} + \underbrace{\sum |a_j^T u_2|^2}_{\min}$$