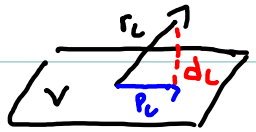
Lesson 24

Read chapter 8

PCA

Th: Given $A \in \mathbb{R}^{m \times n}$ of rank r the best fit k dimensional subspace of $\mathbb{R}^n$ to the rows of A for $k = 1 \dots r$ is span $(v_1 \dots v_k)$ where $v_1 \dots v_k$ are the first k right singular vectors of A.

Best fit subspace V is the subspace that minimizes $\sum_{i=1}^{m}$ (distance of row i of A from V)^2



$$\|d_i\|^2 = \|r_i\|^2 - \|p_i\|^2$$

Look for orthonormal set of vectors $b_1 \dots b_k$

s.t $\sum_{i=1}^{m} \|$Projection r_i on span $(b_1 \dots b_k)\|^2$ is max

$$p_i = r_i^T b_1 \, b_1 + r_i^T b_2 \, b_2 + \dots + r_i^T b_k \, b_k$$

$$\|p_i\|^2 = (r_i^T b_1)^2 + \dots + (r_i^T b_k)^2$$

want to maximize $(r_1^T b_1)^2 + \dots + (r_1^T b_k)^2 +$
$(r_2^T b_1)^2 + \dots + (r_2^T b_k)^2 +$
$\dots$
$(r_m^T b_1)^2 + \dots + (r_m^T b_k)^2$

$$\|Ab_1\|^2 + \|Ab_2\|^2 + \cdots + \|Ab_k\|^2$$

$$\|$$

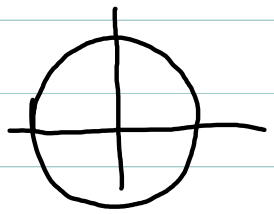
$$(r_1^T b_1)^2 + (r_1^T b_2)^2 + \cdots + (r_1^T b_k)^2 +$$
$$(r_2^T b_1)^2 + (r_2^T b_2)^2 + \cdots + (r_2^T b_k)^2 +$$
$$\cdots \quad \cdots \quad \cdots$$
$$(r_m^T b_1)^2 + (r_m^T b_2)^2 + \cdots + (r_m^T b_k)^2$$

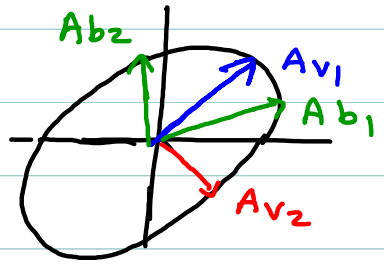
$\downarrow$ Taking $b_2 = v_2$ maximizes this column $\|Ab_2\|^2$

$\downarrow$

Taking $b_1 = v_1$ maximizes this column $\|Ab_1\|^2$

but how do I know I cannot choose b_1 and b_2 with $\|Ab_1\| < \|Av_1\|$ but $\|Ab_2\| > \|Av_2\|$ so that $\|Ab_1\|^2 + \|Ab_2\|^2 > \|Av_1\|^2 + \|Av_2\|^2$



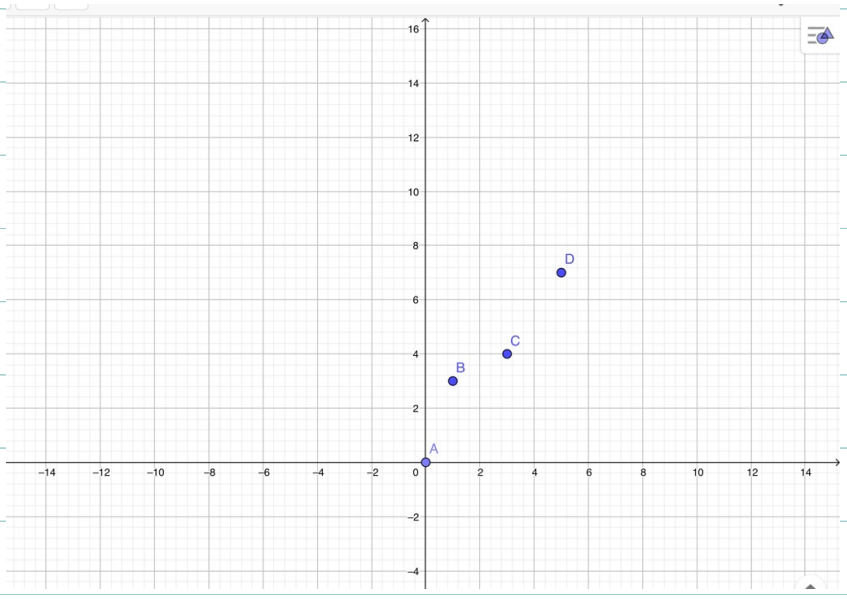


Th: Given $A \in \mathbb{R}^{m \times n}$ of rank r
the best fit k dimensional subspace of $\mathbb{R}^m$
to the columns of A for $k = 1 \dots r$ is span $(u_1 \dots u_k)$
where $u_1 \dots u_k$ are the first k left singular
vectors of A.

Ex: In hw 5 you found linear regression line to:

(0,0) (1,3) (3,4) (5,7) : $y = 0.68 + 1.25x$

What is the best fit line?



Note: maybe no line through origin is very good fit to data, if data is not centered around origin, so we may want to shift our data first.

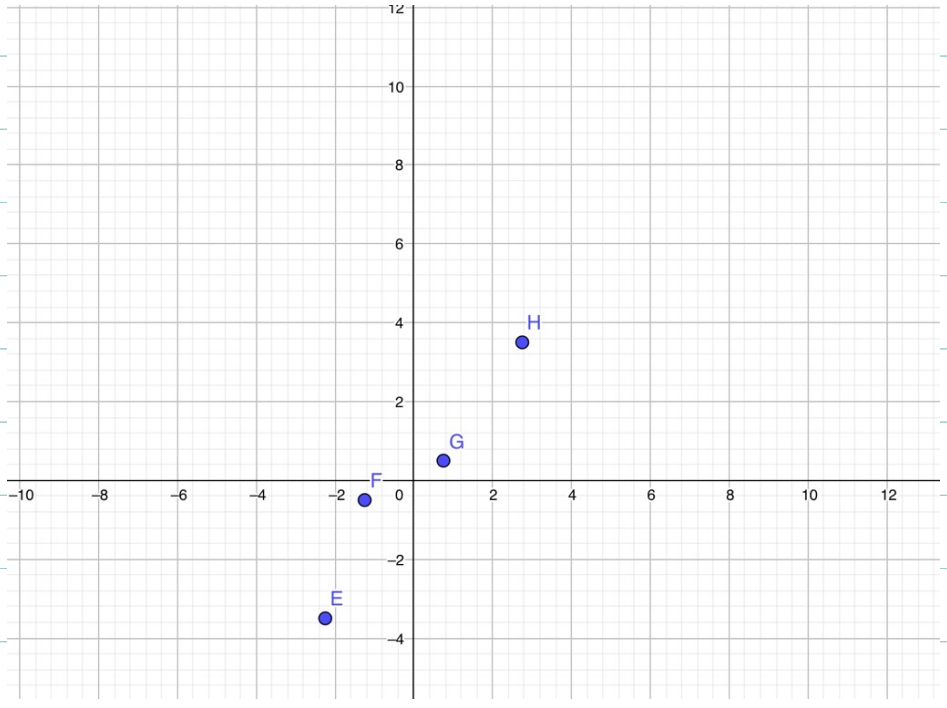
Data is in rows : subtract mean of each column to center around origin.

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 3 \\ 3 & 4 \\ 5 & 7 \end{bmatrix}$$

$$x\text{ mean} = \frac{0+1+3+5}{4} = \frac{9}{4}$$

$$y\text{ mean} = \frac{0+3+4+7}{4} = \frac{14}{4} = \frac{7}{2}$$

$$B = \begin{bmatrix} -\frac{9}{4} & -\frac{7}{2} \\ 1-\frac{9}{4} & 3-\frac{7}{2} \\ 3-\frac{9}{4} & 4-\frac{7}{2} \\ 5-\frac{9}{4} & 7-\frac{7}{2} \end{bmatrix} = \begin{bmatrix} -\frac{9}{4} & -\frac{7}{2} \\ -\frac{5}{4} & -\frac{1}{2} \\ \frac{3}{4} & \frac{1}{2} \\ \frac{11}{4} & \frac{7}{2} \end{bmatrix} \begin{matrix} E \\ F \\ G \\ H \end{matrix}$$

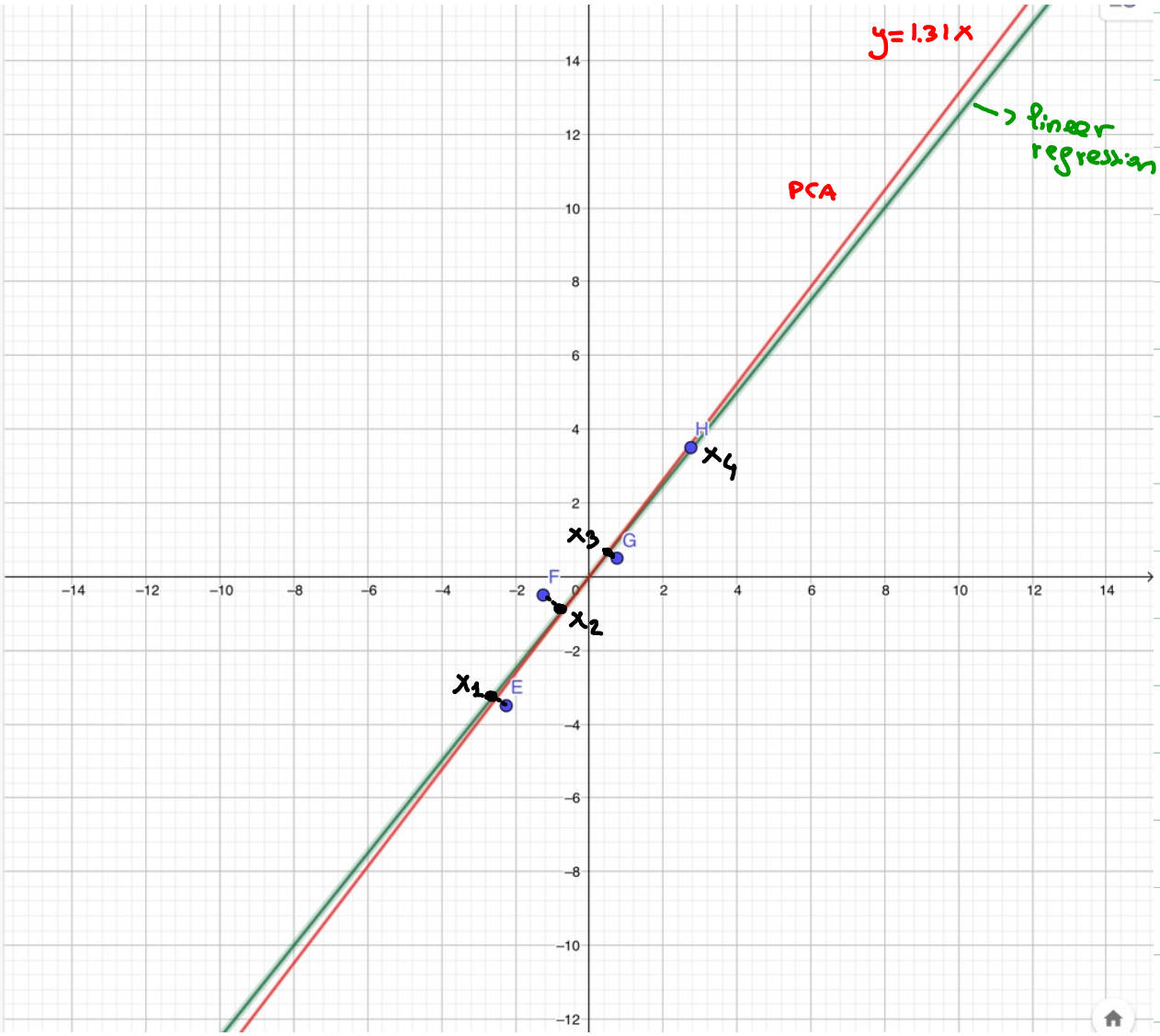


```
[1] top-level scope
  @ none:1

julia> using LinearAlgebra

julia> a=[-9/4 -7/2; -5/4 -1/2; 3/4 1/2; 11/4 7/2]
4×2 Matrix{Float64}:
 -2.25  -3.5
 -1.25  -0.5
  0.75   0.5
  2.75   3.5

julia> svd(a)
SVD{Float64, Float64, Matrix{Float64}}
U factor:
4×2 Matrix{Float64}:
 -0.663581   0.398393
 -0.184732  -0.8405
  0.136306   0.357277
  0.712008   0.0848297
singular values:
2-element Vector{Float64}:
 6.25074081965229
 0.8235527946237668
Vt factor:
2×2 Matrix{Float64}:
 0.605404   0.795918
 0.795918  -0.605404

julia>
```

* v_1

$$\begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} 0.605 \\ 0.796 \end{bmatrix}}_{v_1} \qquad y = \frac{0.796}{0.605}x \qquad y = 1.31x$$

Best fit line is line through origin in the direction of v_1.



$y + \frac{7}{2} = 0.68 + 1.25(x + 9/4)$ Linear regression line

$y = 1.25x$

$y = 1.31x$ Best fit line

Th Given $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ and

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \quad , \text{ then}$$

the rows of A_k are the projections of the rows of A onto $V_k = \text{span}(v_1\ v_2 \cdots v_k)$

$$A = \begin{bmatrix} r_1^T \\ \vdots \\ r_m^T \end{bmatrix} \qquad A_k = \begin{bmatrix} (\text{proj}_{V_k} r_1)^T \\ \vdots \\ (\text{proj}_{V_k} r_m)^T \end{bmatrix}$$

Proof: the i^{th} row of A_k is

$$e_i^T A_k = e_i^T \sum_{j=1}^{k} \sigma_j u_j v_j^T = \sum_{j=1}^{k} \sigma_j e_i^T u_j v_j^T$$

i^{th} row of A is: $r_i^T = e_i^T A$

$$P_{V_k} r_i = \sum_{j=1}^{k} \underbrace{r_i^T v_j}_{\text{scalar}} v_j$$

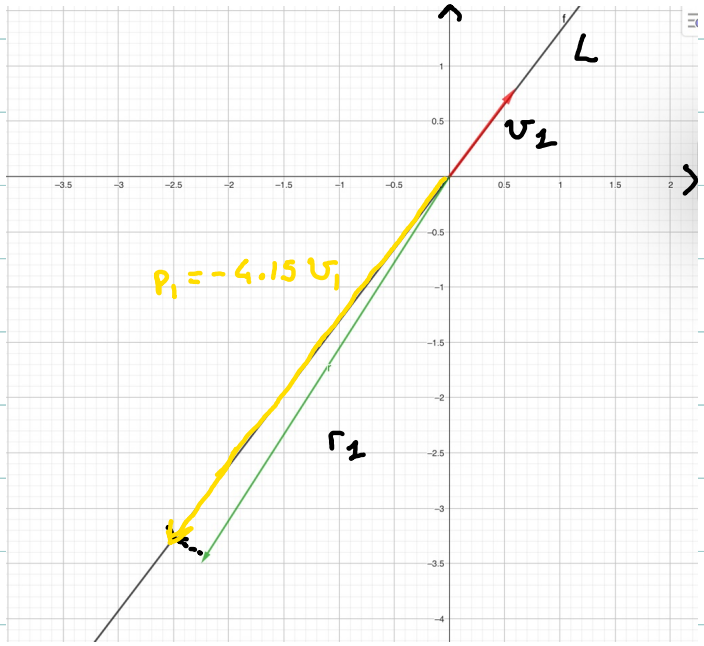
(projection formula starting from orthonormal basis. see hw)

we are transposing to make this into a $1 \times n$ vector

$$(P_{V_k} r_i)^T = \sum_{j=1}^{k} \overbrace{e_i^T A}^{r_i^T} v_j v_j^T = \sum_{j=1}^{k} e_i^T U \Sigma V^T v_j v_j^T =$$

$$= \sum_{j=1}^{k} e_i^T U \sigma_j e_j v_j^T = \sum_{j=1}^{k} \sigma_j e_i^T u_j v_j^T$$

Back to our example



$$A = \begin{bmatrix} -9/4 & -7/2 \\ -5/4 & -11/2 \\ 3/4 & 11/2 \\ 11/4 & 7/2 \end{bmatrix} \begin{matrix} E \\ F \\ G \\ H \end{matrix}$$

From SVD: $v_1 = \begin{bmatrix} 0.605 \\ 0.796 \end{bmatrix}$, $u_1 = \begin{bmatrix} -0.66 \\ -0.18 \\ 0.14 \\ 0.71 \end{bmatrix}$, $\sigma_1 = 6.25$.

$L = \text{span}(v_1)$

$$\text{Proj}_L\left(\begin{bmatrix} -9/4 \\ -7/2 \end{bmatrix}\right) = x_1 v_1$$

$$x_1 = \begin{bmatrix} -9/4 & -7/2 \end{bmatrix} \underbrace{\begin{bmatrix} 0.605 \\ 0.796 \end{bmatrix}}_{v_1} \approx -4.15$$

$$\text{Proj}_{v_1}(r_1) = -4.15 \begin{bmatrix} 0.605 \\ 0.796 \end{bmatrix} \approx \begin{bmatrix} -2.51 \\ -3.3 \end{bmatrix}$$

$$A_1 = 6.251 \begin{bmatrix} -0.664 \\ -0.18 \\ 0.14 \\ 0.71 \end{bmatrix} \begin{bmatrix} 0.605 & 0.796 \end{bmatrix} = 6.251 \begin{bmatrix} -0.664 \cdot 0.605 & -0.664 \cdot 0.796 \\ & \end{bmatrix}$$

$4 \times 1 \qquad 1 \times 2$

$$= \begin{bmatrix} -2.51 & -3.3 \\ = & = \\ = & = \end{bmatrix}$$

Two aspects of PCA :

1) dimension reduction

2) maximum variation of data

Example of PCA:

In a study, 200 gene expressions were studied in 96 tissue samples

Data are 96 vectors in $\mathbb{R}^{200}$

A = 96 x 200 matrix of data

To visualize the data we want to project it onto 2 dimensional subspace: best fit 2 subspace.

1) Preprocessing like centering data $A \to A'$

2) $A' = U \Sigma V^T$

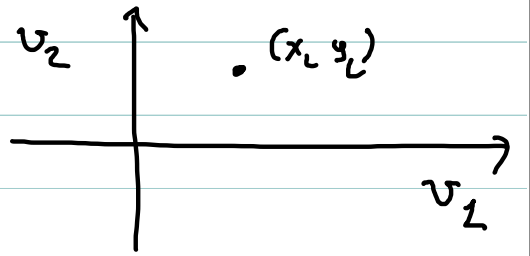
3) Best fit 2 dimensional space $= \text{span}(v_1, v_2) = V_2$

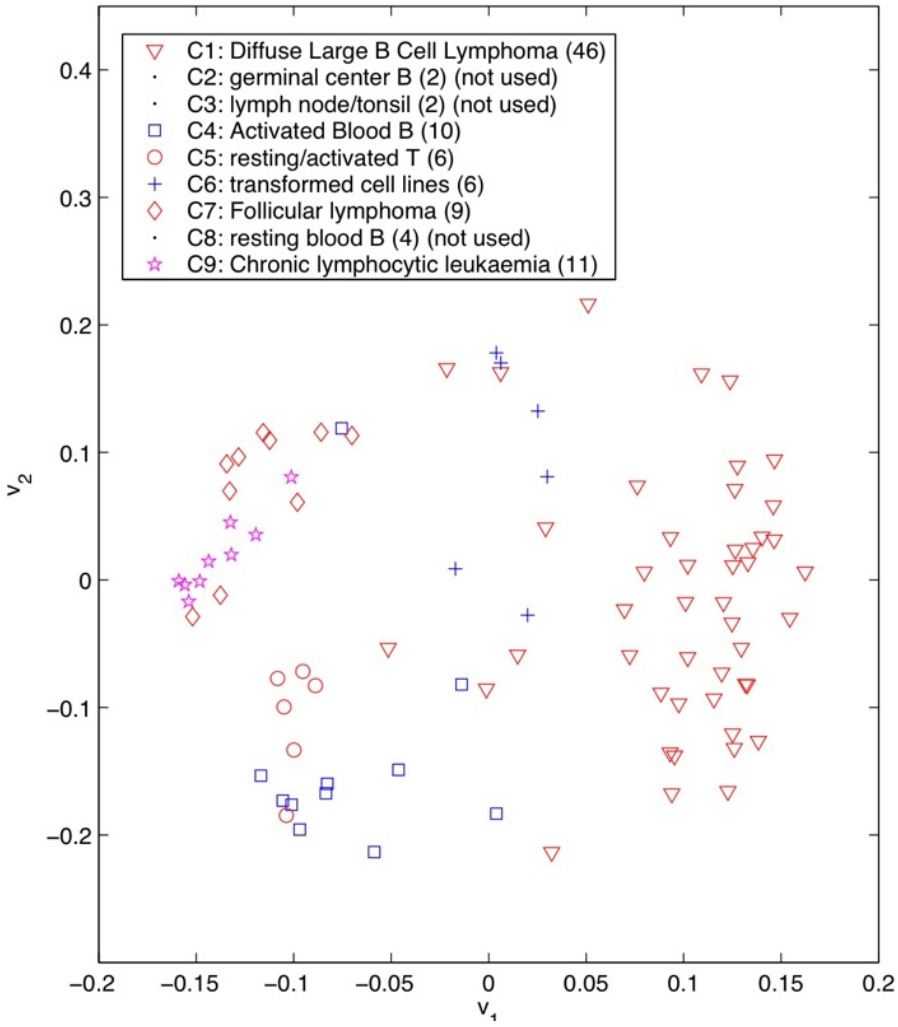
4) Project the rows of A' onto V_2:

$$\text{Proj}_{V_2} r_i = x_i v_1 + y_i v_2$$

$x_i = r_i^T v_2$, $y_i = r_2^T v_2$.

5) plot all points (x_i, y_i) in 2D space





Plot of the points (x_i, y_i)

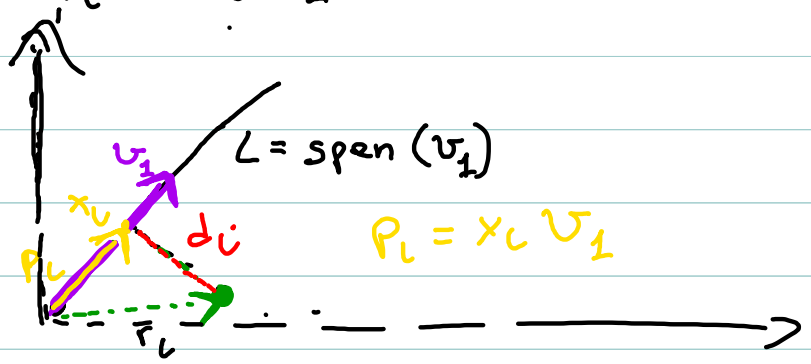
Maximum variation

See animation at: setosa.io

data centered around origin

P_i projection of r_i on v_1 (principal component = first right singular vector)

$P_i = x_i v_1$ $\quad x_i = r_i^T v_1$ $\quad x_i = \|P_i\|$



① $\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} r_i^T v_1 = 0$ since $\sum_{i=1}^{m} r_i^T = 0$

(data is centered) so x_i are also centered around 0. How spread out are they?

② In statistics variance is a measure of how data is spread out around mean

$\frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2 = \frac{1}{m-1} \sum_{i=1}^{m} x_i^2 = \frac{1}{m-1} \sum_{i=1}^{m} \|P_i\|^2$

since are best fit line maximizes this sum it maximizes the variance of the data

Example to explain proof of Eckert Young th.

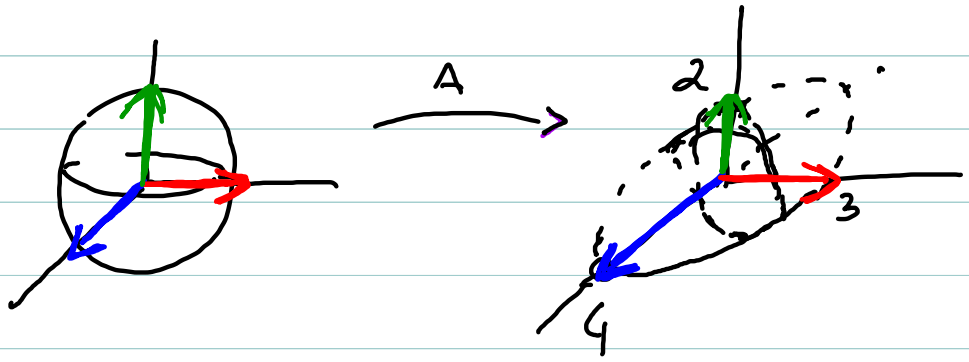
$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} = I \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} I^T =$$

$$= 4 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} [100] + 3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} [010] + 2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} [001]$$

$$A_1 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 3.5 & 3.5 & 0 \\ 3.5 & 3.5 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

We know $\|A - A_1\|_2 = 3$ so we must have $\|A - B\|_2 \geq 3$

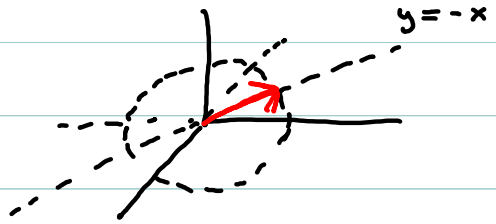


want w, $\|w\|=1$ s.t $\|Aw-Bw\| \geq 3$

Null(B) is the plane $x+y=0$

Span$\left(\begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}0\\1\\0\end{bmatrix}\right)$ is the plane $z=0$

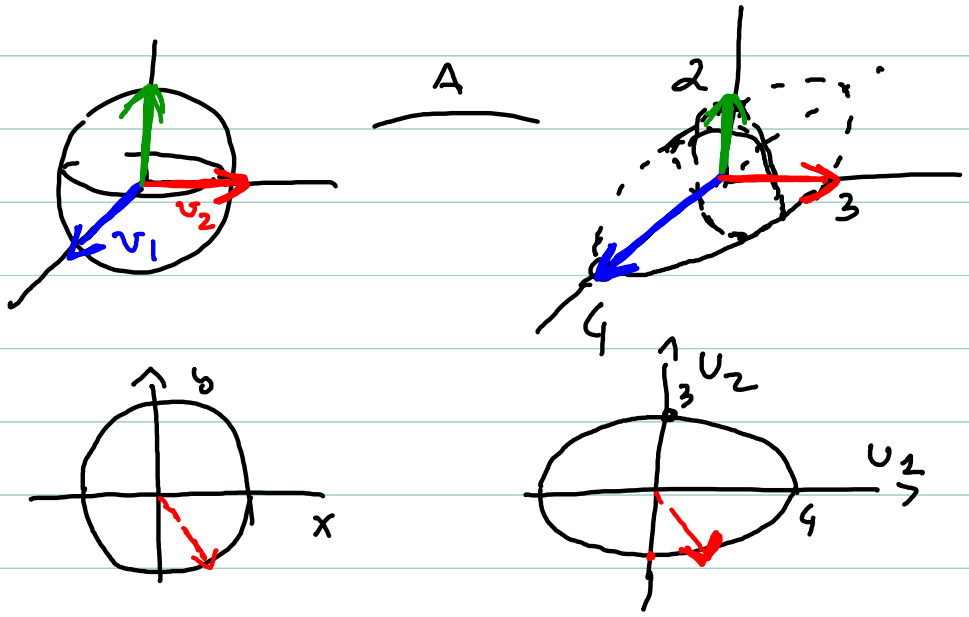
want w on both of these planes



$$w = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}$$

$$\left\| \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix} \right\| = \left\| \begin{bmatrix} 4/\sqrt{2} \\ -3/\sqrt{2} \\ 0 \end{bmatrix} \right\| = \sqrt{\frac{16}{2}+\frac{9}{2}} = \frac{5}{\sqrt{2}}$$

≈ 3.54


A
v_2
v_1
2
3
4

b
x
3
U_2
U_1
4