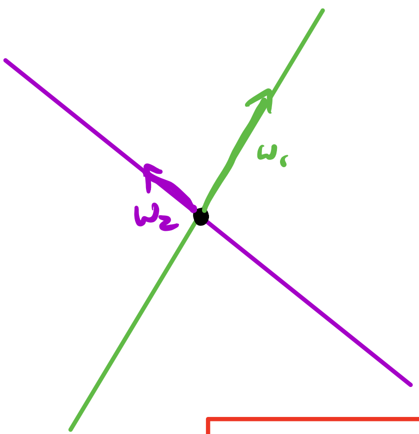
Geometry of the SVD

We have drawn pictures of triple products before:

Diagonalization:

$$A = \frac{1}{10}\begin{pmatrix} 11 & 6 \\ 9 & 14 \end{pmatrix} = CDC^{-1}$$

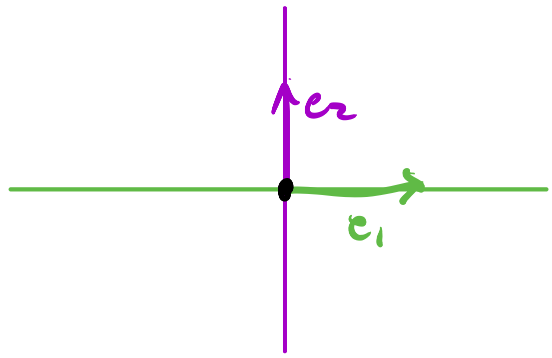
$$\text{for} \quad C = \begin{array}{cc} w_1 & w_2 \\ \end{array} \begin{pmatrix} 2 & -1 \\ 3 & 1 \end{pmatrix} \qquad D = \begin{array}{cc} \lambda_1 & \lambda_2 \\ \end{array} \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$



$$\xrightarrow{C^{-1}}$$

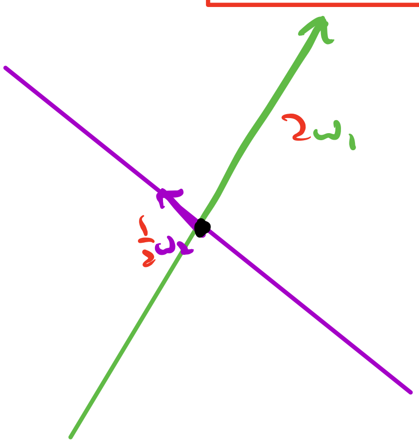
$C^{-1}w_1 = e_1$

$C^{-1}w_2 = e_2$



$\downarrow$ $\boxed{\begin{aligned} Aw_1 &= 2w_1 \\ Aw_2 &= \tfrac{1}{2}w_2 \end{aligned}}$

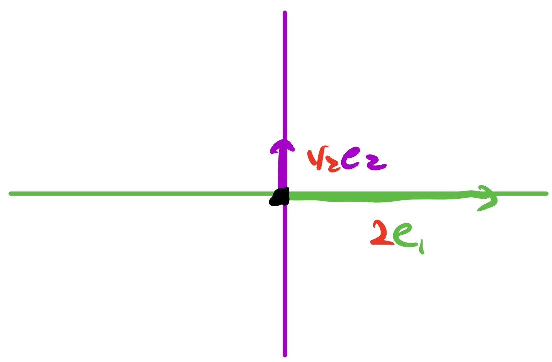
$\downarrow D$



$$\xleftarrow{C}$$

$Ce_1 = w_1$

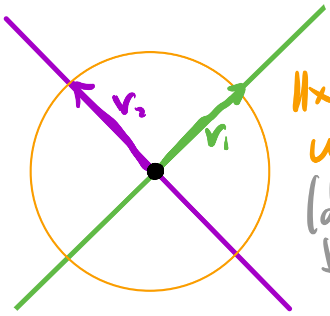
$Ce_2 = w_2$



SVD:

$$A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix} = U\Sigma V^T \quad \text{for}$$

$$\begin{array}{ccc} & u_1 & u_2 \\ U = \frac{1}{\sqrt{10}} & \multicolumn{2}{c}{\begin{pmatrix} 1 & -3 \\ 3 & 1 \end{pmatrix}} \end{array} \qquad \begin{array}{ccc} & v_1 & v_2 \\ V = \frac{1}{\sqrt{2}} & \multicolumn{2}{c}{\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}} \end{array} \qquad \begin{array}{ccc} & \sigma_1 & \sigma_2 \\ \Sigma = & \multicolumn{2}{c}{\begin{pmatrix} 3\sqrt{5} & 0 \\ 0 & \sqrt{5} \end{pmatrix}} \end{array}$$

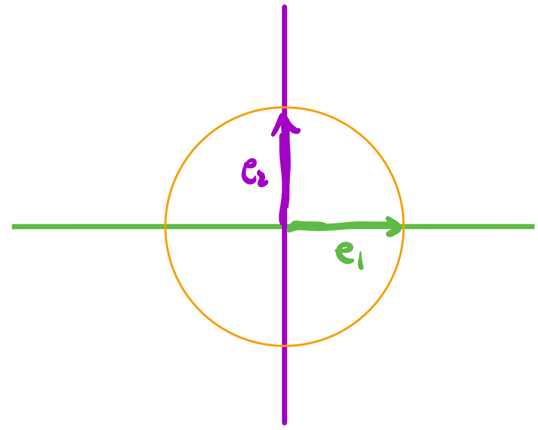


$\|x\|=1$:
unit circle
(draw this to help visualize)

$$\xrightarrow{V^T = V^{-1}}$$

$V^T v_1 = e_1$

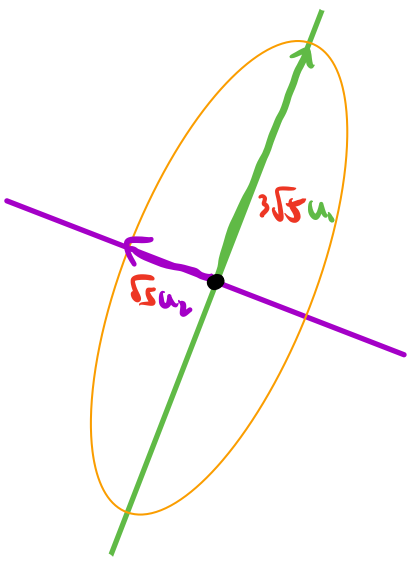
$V^T v_2 = e_2$



$A \downarrow$

$$\boxed{\begin{aligned} A v_1 &= 3\sqrt{5}\, u_1 \\ A v_2 &= \sqrt{5}\, u_2 \end{aligned}}$$

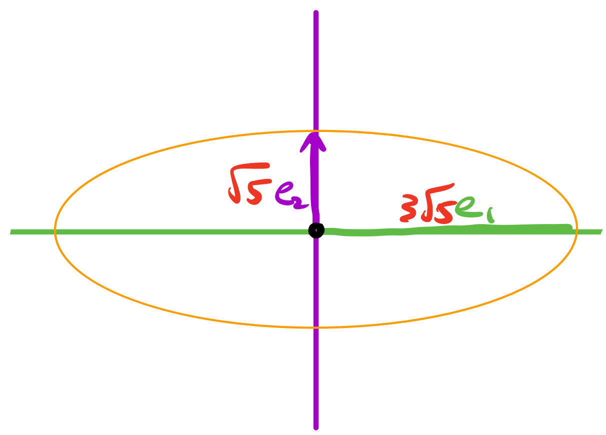
$\downarrow \Sigma$



$$\xleftarrow{U}$$

$U e_1 = u_1$

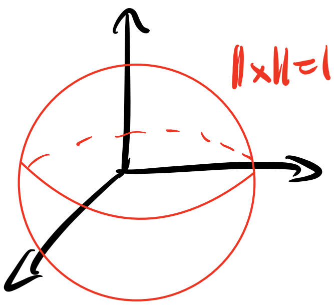
$U e_2 = u_2$



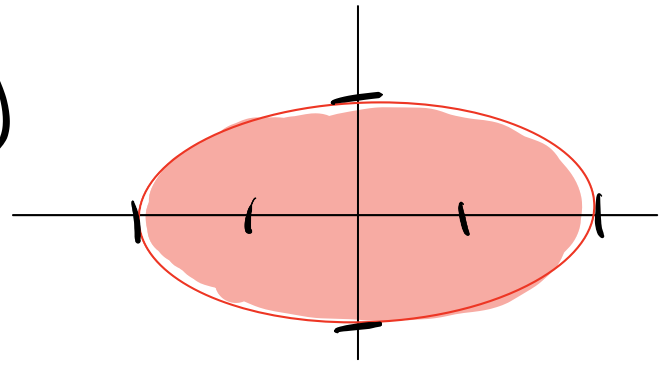
A = (first rotate/flip) (then stretch) (then rotate/flip)

Notes / caveats:

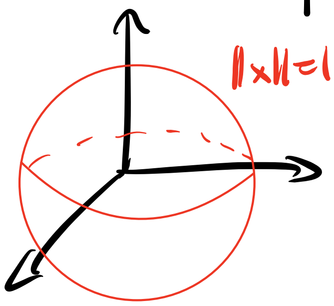
- Diagonalization: start & end in $\{w_1, w_2\}$ basis
  SVD: start with $\{v_1, v_2\}$ & end with $\{u_1, u_2\}$ basis
- The V^T & U steps preserve lengths & angles (rotations / flips)
- The Σ step can change dimensions:



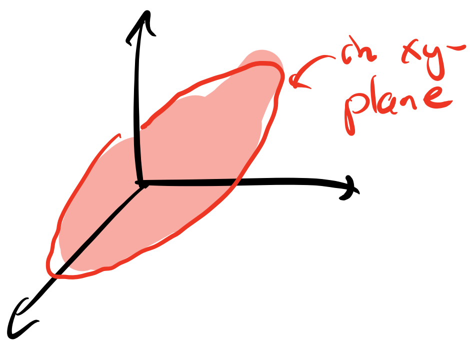
$$\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$



- The Σ step can flatten a sphere:



$$\Sigma' = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$



Principal Component Analysis (PCA)

This is "SVD in stats language".

→ it's often how SVD (or "linear algebra") is used in statistics & data analysis.

Idea: If you have n samples of m values each ⇝ columns of an $m \times n$ data matrix

Let's introduce some terminology from statistics.

One Value ($m=1$):

Let's record everyone's scores on Midterm 3: samples $x_1, \ldots, x_n$

Mean (average): $\mu = \frac{1}{n}(x_1 + \cdots + x_n)$

Variance: $s^2 = \frac{1}{n-1}\left[(x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right]$

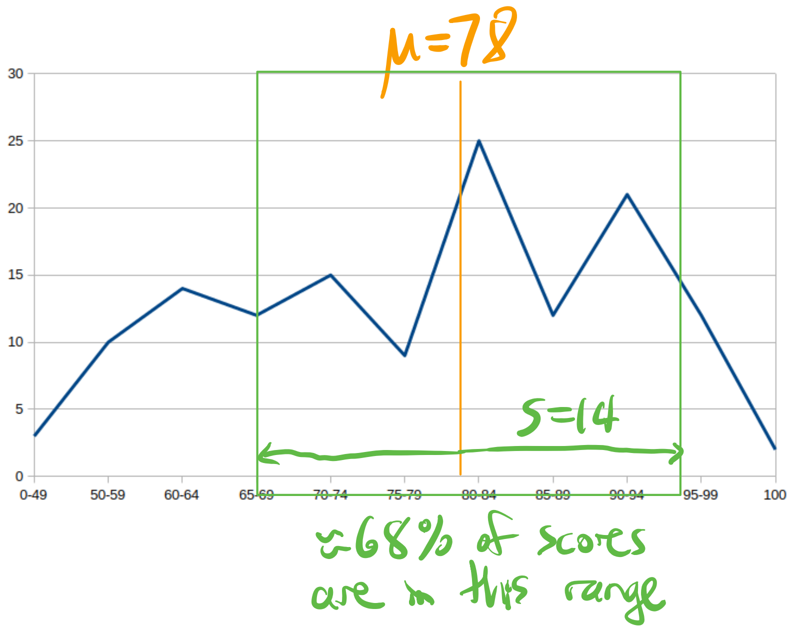
Standard Deviation: $s = \sqrt{\text{variance}}$

This tells you how "spaced out" the samples are: $\approx 68\%$ of samples are within $\pm s$ of the mean.

Where do these formulas come from?

Take a stats class!

Eg: Actual midterm 3 scores from Fall '20:



Two Values ($m=2$):

Let's record everyone's scores on problems 1 & 2 on Midterm 3:

samples $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ $\quad$ x_i = score on problem 1, y_i = score on problem 2

Mean scores:

Problem 1: $\mu_1 = \frac{1}{n}(x_1 + \dots + x_n)$

Problem 2: $\mu_2 = \frac{1}{n}(y_1 + \dots + y_n)$

Recenter to compute variance:

$\bar{x}_i = x_i - \mu_1 \qquad \bar{y}_i = y_i - \mu_2$ (subtract means)

Variance:

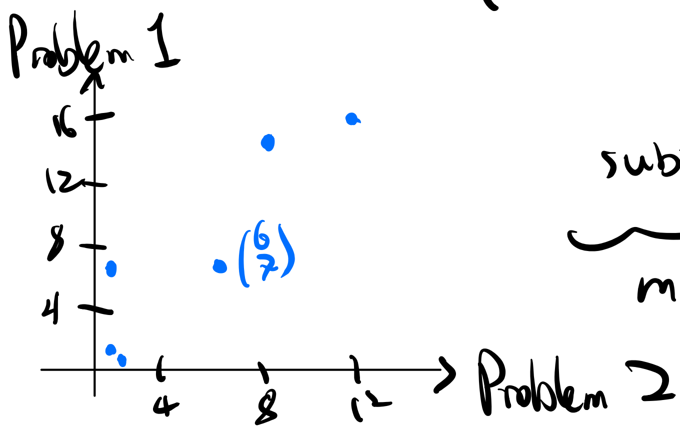
Problem 1: $s_1^2 = \frac{1}{n-1}(\bar{x}_1^2 + \dots + \bar{x}_n^2)$

Problem 2: $s_2^2 = \frac{1}{n-1}(\bar{y}_1^2 + \dots + \bar{y}_n^2)$

Total Variance: $s^2 = s_1^2 + s_2^2$

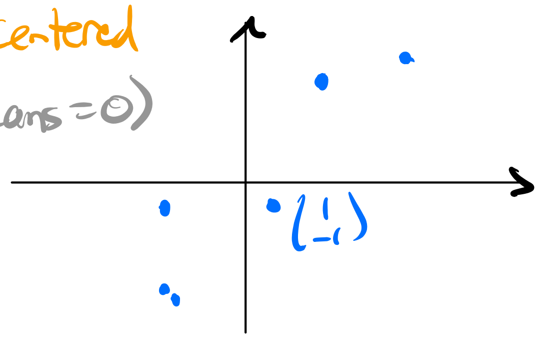
Eg: Scores $\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} 8 \\ 15 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 12 \\ 16 \end{pmatrix}, \begin{pmatrix} 6 \\ 7 \end{pmatrix}, \begin{pmatrix} 1 \\ 7 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\mu_1 = 5$, $\mu_2 = 8$

recenter: $\begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \begin{pmatrix} -4 \\ -6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -4 \\ -1 \end{pmatrix}, \begin{pmatrix} -3 \\ -7 \end{pmatrix}$ $s_1^2 = 20$, $s_2^2 = 40$



subtract mean $\longrightarrow$

centered (means = 0)



Store in matrices:

$$A_0 = \begin{pmatrix} x_1 & \cdots & x_n \\ y_1 & & y_n \end{pmatrix} = \begin{pmatrix} 8 & 1 & 12 & 6 & 1 & 2 \\ 15 & 2 & 16 & 7 & 7 & 1 \end{pmatrix}$$

$$A = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_n \\ \bar{y}_1 & & \bar{y}_n \end{pmatrix} = \begin{pmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{pmatrix}$$

Covariance Matrix:

$$S = \frac{1}{n-1} AA^T = \frac{1}{n-1} \begin{pmatrix} (\text{row }1)\cdot(\text{row }1) & (\text{row }1)\cdot(\text{row }2) \\ (\text{row }2)\cdot(\text{row }1) & (\text{row }2)\cdot(\text{row }2) \end{pmatrix}$$

$$= \frac{1}{n-1} \begin{pmatrix} \bar{x}_1^2 + \cdots + \bar{x}_n^2 & \bar{x}_1\bar{y}_1 + \cdots + \bar{x}_n\bar{y}_n \\ \bar{x}_1\bar{y}_1 + \cdots + \bar{x}_n\bar{y}_n & \bar{y}_1^2 + \cdots + \bar{y}_n^2 \end{pmatrix}$$

The diagonal entries are the variances:

$$s_1^2 = \frac{1}{n-1}\left(\bar{x}_1^2 + \cdots + \bar{x}_n^2\right) \qquad s_2^2 = \frac{1}{n-1}\left(\bar{y}_1^2 + \cdots + \bar{y}_n^2\right)$$

The trace is the total variance:

$$\operatorname{Tr}(S) = s_1^2 + s_2^2 = s^2$$

The off-diagonal entries are called covariances.

Eg. the $(1,2)$-entry is

$$(\text{row }1)\cdot(\text{row }2) = \bar{x}_1\bar{y}_1 + \cdots + \bar{x}_n\bar{y}_n$$

- If this is positive then $\bar{x}_i$ & $\bar{y}_i$ generally have the same sign: if you did above average on P1 then you likely did above average on P2 too, & vice-versa. The values are correlated.
- If this is negative then $\bar{x}_i$ & $\bar{y}_i$ generally have opposite signs: if you did above average on P1 then you likely did below average on P2, & vice-versa. The values are anti-correlated.

In our case:

$$S = \frac{1}{5}AA^T = \begin{pmatrix} 20 & 25 \\ 25 & 40 \end{pmatrix} \qquad \begin{aligned} s_1^2 &= 20 \\ s_2^2 &= 40 \end{aligned}$$

$(1,2)$-covariance $= 25 > 0$: people who did above average on P1 likely did above average on P2.

So far we've done statistics.

Now we apply the SVD to A.

This will tell us which directions have the largest & smallest variance.

SVD of the Centered Data Matrix A (times $\frac{1}{\sqrt{n-1}}$):

$$\frac{1}{\sqrt{n-1}} A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T \rightsquigarrow \frac{1}{\sqrt{n-1}} A^T = \sigma_1 v_1 u_1^T + \cdots + \sigma_r v_r u_r^T$$

The σ_i^2 are the nonzero eigenvalues of the covariance matrix:

$$S = \frac{1}{n-1} A A^T = \left(\frac{1}{\sqrt{n-1}} A\right)\left(\frac{1}{\sqrt{n-1}} A\right)^T$$

Recall: $\mathrm{Tr}(S) = \sigma_1^2 + \cdots + \sigma_r^2$ (sum of (nonzero) eigenvalues)

$\rightarrow$ HW 11 #12(b)

But $\mathrm{Tr}(S) = s_1^2 + \cdots + s_n^2 = s^2 =$ total variance, so:

$$\boxed{\text{total variance } \; s^2 = s_1^2 + \cdots + s_n^2 = \sigma_1^2 + \cdots + \sigma_r^2}$$

Recall: u_1 maximizes $\left\| \frac{1}{\sqrt{n-1}} A^T x \right\|^2$ subject to $\|x\| = 1$ with maximum value σ_1^2

$\rightarrow$ HW 14 #9

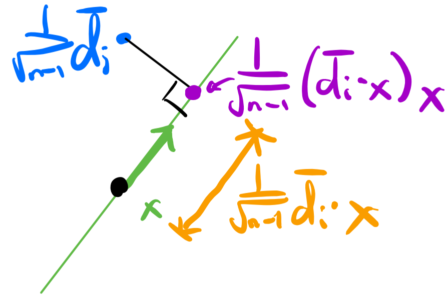
What is $\frac{1}{\sqrt{n-1}} A^T x$? Assume $\|x\| = 1$.

If $d_1, \ldots, d_n$ are our recentered data points / columns of A

$$\rightsquigarrow \frac{1}{\sqrt{n-1}} A^T x = \frac{1}{\sqrt{n-1}} \begin{pmatrix} - d_1 - \\ \vdots \\ - d_n - \end{pmatrix} x = \frac{1}{\sqrt{n-1}} \begin{pmatrix} d_1 \cdot x \\ \vdots \\ d_n \cdot x \end{pmatrix}$$

Since x is a unit vector, the projection of $\frac{1}{\sqrt{n-1}} d_i$ onto $\mathrm{Span}\{x\}$ is $\frac{1}{\sqrt{n-1}} (d_i \cdot x) x$ and $\frac{1}{\sqrt{n-1}} d_i \cdot x$ is the amount of $\frac{1}{\sqrt{n-1}} d_i$ in the x-direction (picture below)

Taking $x = u_1$:



$$\frac{1}{\sqrt{n-1}}\begin{pmatrix} \bar{d}_1 \cdot u_1 \\ \vdots \\ \bar{d}_n \cdot u_1 \end{pmatrix} = \frac{1}{\sqrt{n-1}} A^T u_1 = \sigma_1 v_1$$

$$\Rightarrow \sigma_1 u_1 v_1^T = u_1 (\sigma_1 v_1)^T = \frac{1}{\sqrt{n-1}} u_1 (\bar{d}_1 \cdot u_1 \ \cdots \ \bar{d}_n \cdot u_1)$$

$$= \frac{1}{\sqrt{n-1}}\left((\bar{d}_1 \cdot u_1)u_1 \ \cdots \ (\bar{d}_n \cdot u_1)u_1\right)$$

$=$ projections of $\frac{1}{\sqrt{n-1}}\bar{d}_i$ onto $\mathrm{Span}\{u_1\}$

Upshot:

$$\sigma_1^2 = \left\|\frac{1}{\sqrt{n-1}} A^T u_1\right\|^2 = \frac{1}{n-1}\left[(\bar{d}_1 \cdot u_1)^2 + \cdots + (\bar{d}_n \cdot u_1)^2\right]$$

$=$ variance in the u_1-direction

This quantity is maximized at u_1, so

u_1 is the direction of greatest variance

$=$ the first principal component

The summand $\sigma_1 u_1 v_1^T$ in the SVD of $\frac{1}{\sqrt{n-1}} A$ is:

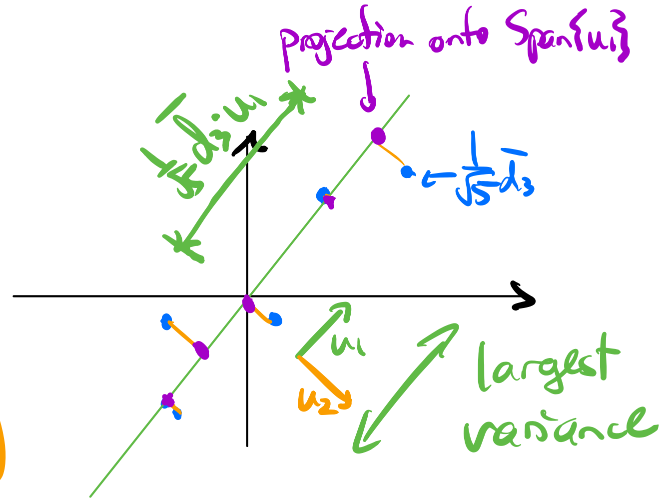
$\sigma_1 u_1 v_1^T = \Bigg($ projection of $\frac{1}{\sqrt{n-1}}\bar{d}_1$ onto $\mathrm{Span}\{u_1\}$, $\ldots$, projection of $\frac{1}{\sqrt{n-1}}\bar{d}_n$ onto $\mathrm{Span}\{u_1\}$ $\Bigg)$

In our case:

$$\frac{1}{\sqrt{6-1}}A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T \quad \text{for}$$

$\sigma_1^2 \approx 56.9$ $\qquad$ $\sigma_2^2 \approx 3.07$

$u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$ $\qquad$ $u_2 \approx \begin{pmatrix} 0.828 \\ -0.561 \end{pmatrix}$



The purple dots are the columns of $\sigma_1 u_1 v_1^T$.

So the first principal component is u_1, and the variance in that direction is ≈ 56.9.

(NB this is greater than the Problem 1 variance $= 20$ & the Problem 2 variance $= 40$)

NB: Here's how I should (but won't) grade the final exam:

- Put the scores of each problem in an $m \times n$ matrix A_0 ($m = \#$problems, $n = \#$students)
- Subtract row averages to recenter $\rightsquigarrow$ matrix A
- Compute the SVD of $\frac{1}{\sqrt{n-1}}A$ $\rightsquigarrow$ 1^{st} principal component u_1
- The score for student i is (mean score) $+ \frac{1}{\sqrt{n-1}} d_i \cdot u_1$

This gives the same average but maximizes the standard deviation by weighting the problems.

Higher Principal Components

$$\frac{1}{\sqrt{n-1}}A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T \rightsquigarrow \frac{1}{\sqrt{n-1}}A^T = \sigma_1 v_1 u_1^T + \cdots + \sigma_r v_r u_r^T$$

"Recall": $\left\|\frac{1}{\sqrt{n-1}}A^T x\right\|^2$ is maximized [subject to $x \perp u_1$ and $\|x\| = 1$] at u_2, with maximum value σ_2^2. → HW 15 #15

More generally, $\left\|\frac{1}{\sqrt{n-1}}A^T x\right\|^2$ is maximized subject to $\|x\|=1$, $x \perp u_1, \ldots, x \perp u_{i-1}$ at u_i, with maximum value σ_i^2. By the same reasoning as before:

$$\boxed{\begin{aligned}\sigma_i^2 = \left\|\tfrac{1}{\sqrt{n-1}}A^T u_i\right\|^2 &= \frac{1}{n-1}\left[(\bar{d}_1 \cdot u_i)^2 + \cdots + (\bar{d}_n \cdot u_i)^2\right] \\ &= \text{variance in the } u_i\text{-direction}\end{aligned}}$$

This quantity is maximized at u_i subject to $x \perp u_1, \ldots, x \perp u_{i-1}$, so

$$\boxed{\begin{aligned}u_i &\text{ is the direction of } i^{th} \text{ greatest variance} \\ &= \text{the } i^{th} \text{ principal component}\end{aligned}}$$

The summand $\sigma_i u_i v_i^T$ in the SVD of $\frac{1}{\sqrt{n-1}}A$ is:

$$\boxed{\sigma_i u_i v_i^T = \begin{pmatrix} \text{projection of } \frac{1}{\sqrt{n-1}}\bar{d}_1 \text{ onto } \operatorname{Span}\{u_i\}, & \cdots, & \text{projection of } \frac{1}{\sqrt{n-1}}\bar{d}_n \text{ onto } \operatorname{Span}\{u_i\}\end{pmatrix}}$$

NB: u_r is the direction of smallest variance.

So if σ_r is small then $u_r \cdot \bar{d}_i$ is small for all i

$\rightsquigarrow$ the data points almost lie on $\mathrm{Span}\{u_r\}^\perp$.

Likewise, if σ_{r-1} & σ_r are small

$\rightsquigarrow$ the data points almost lie on $\mathrm{Span}\{u_{r-1}, u_r\}^\perp$

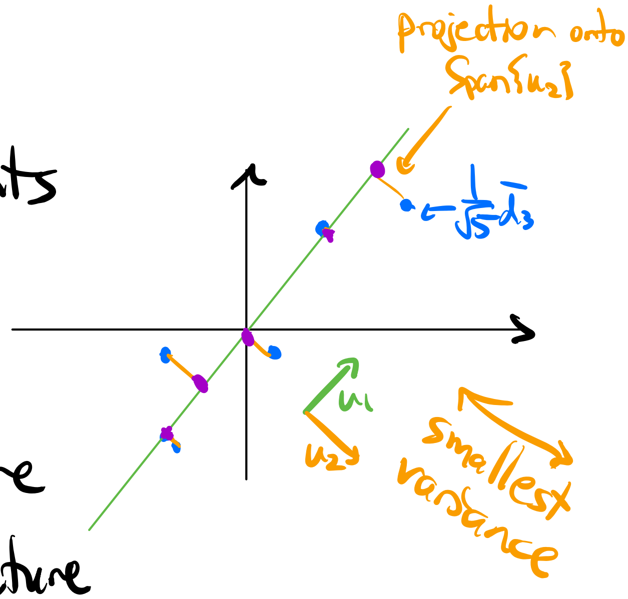
etc. This is dimension reduction.

In Our Case:

$\sigma_2^2 \approx 3.07 \Rightarrow$ our data points almost lie on

$$\mathrm{Span}\{u_2\}^\perp = \mathrm{Span}\{u_1\}$$

The columns of $\sigma_2 u_2 v_2^T$ are the orange lines in the picture (projection onto $\mathrm{Span}\{u_2\}$)



$u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$

Interpretation:

- Total variance is $60 = 20$ on prob $1 + 40$ on prob 2
- $\sigma_1^2 \approx 56.9$ is variance in the u_1-direction
- $\sigma_2^2 \approx 3.1$ is variance in the u_2-direction.

This says: the variance on Problem 2 is $.828/.560 \approx 1.48$ times more than on Problem 1, so I should weight Problem 2 more heavily!