Geometry of the SVD: Matrix Form

We have drawn pictures of a triple product decomposition before.

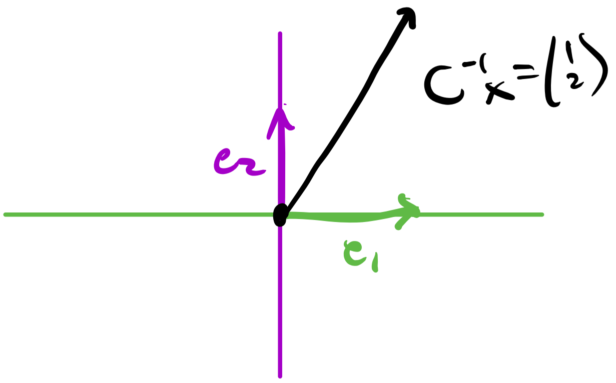
Diagonalization:

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

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

To evaluate $Ax = CDC^{-1}x$:

(1) multiply by C^{-1} (2) multiply by D (3) multiply by C

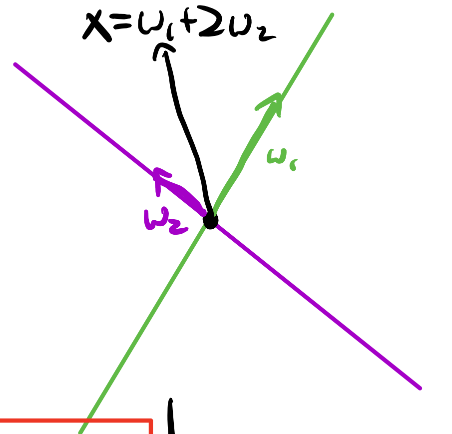


(1) $\xleftarrow{C^{-1}}$

$C^{-1}w_1 = e_1$

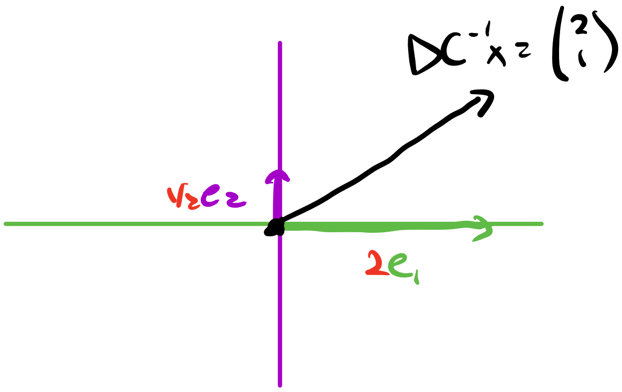
$C^{-1}w_2 = e_2$

$C^{-1}x = \begin{pmatrix}1\\2\end{pmatrix}$



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

(2) $\downarrow D$

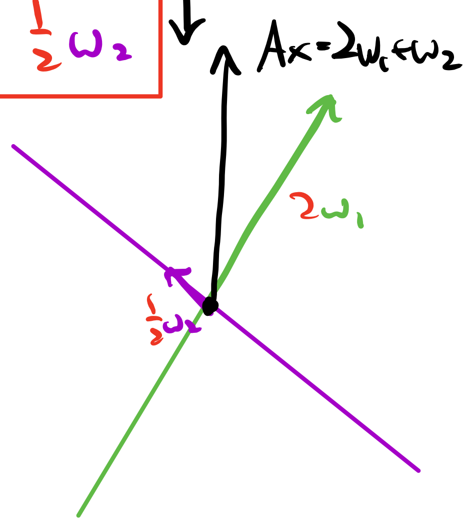


(3) $\xrightarrow{C}$

$Ce_1 = w_1$

$Ce_2 = w_2$

$C\begin{pmatrix}2\\1\end{pmatrix} = 2w_1 + w_2$



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

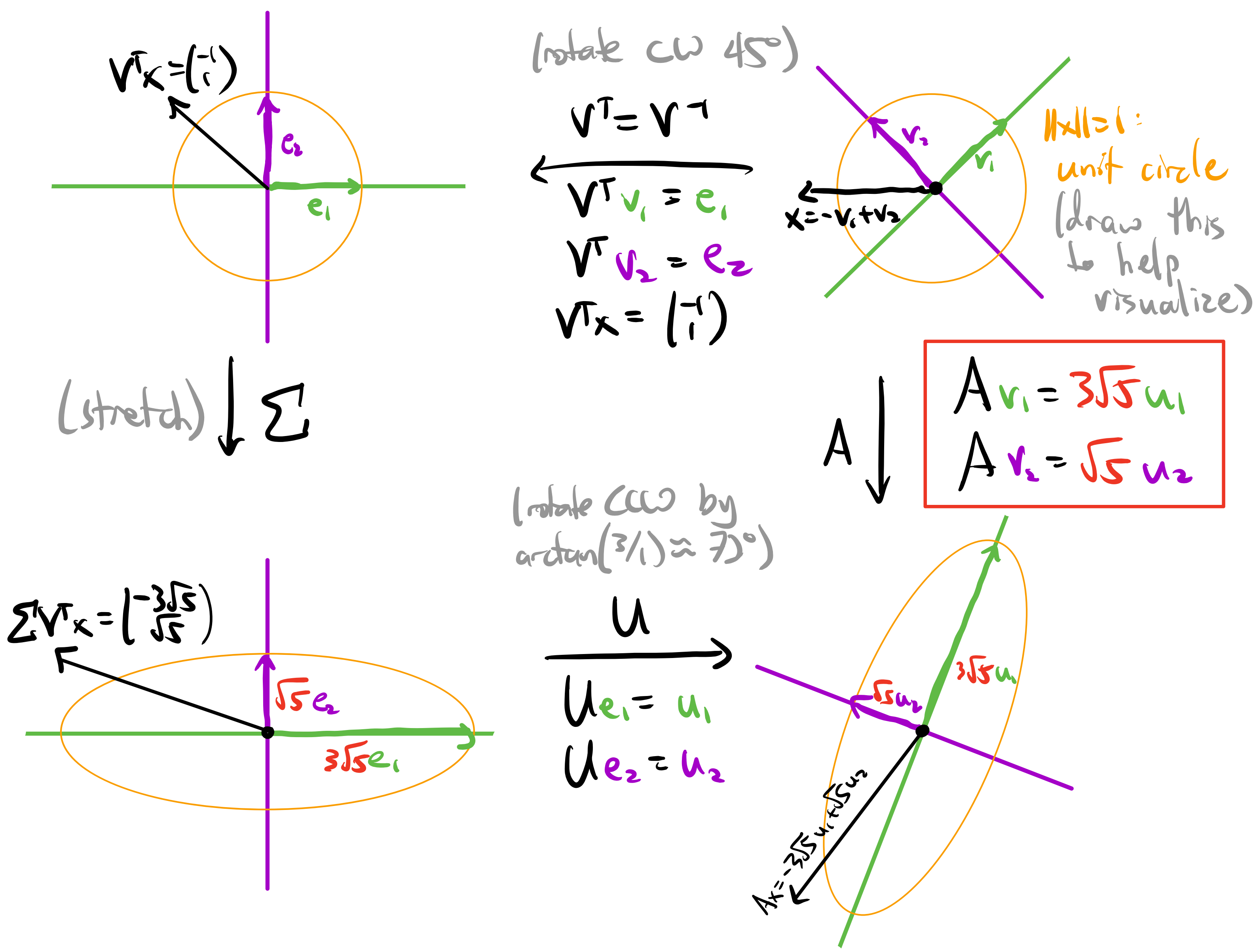
$$U = \frac{1}{\sqrt{10}} \begin{array}{c} \begin{array}{cc} u_1 & u_2 \end{array} \\ \begin{pmatrix} 1 & -3 \\ 3 & 1 \end{pmatrix} \end{array} \qquad V = \frac{1}{\sqrt{2}} \begin{array}{c} \begin{array}{cc} v_1 & v_2 \end{array} \\ \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \end{array} \qquad \Sigma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

To evaluate $Ax = U\Sigma V^T x$:

(1) multiply by V^T (2) multiply by Σ (3) multiply by U

But U and V^T are orthogonal, so these just rotate/flip.

$Ax =$ (1) rotate/flip (2) stretch (3) rotate/flip

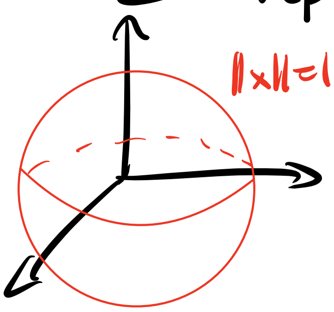


Notes / caveats:

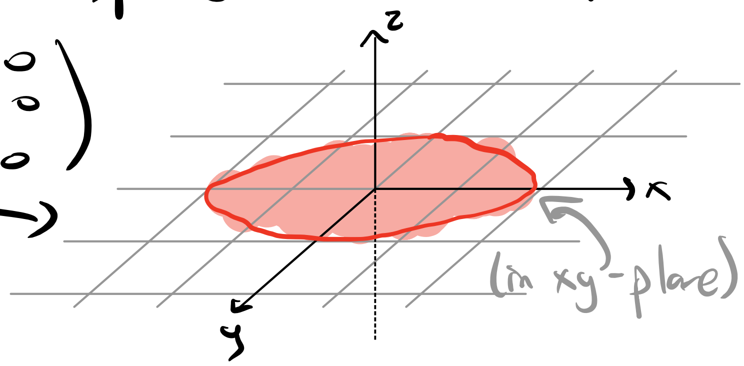
- Diagonalization: start & end in $\{w_1, w_2\}$ basis
  SVD: start with $\{v_1, v_2\}$ & end with $\{u_1, u_2\}$ basis
  → Different bases!

- The V^T & U steps preserve lengths & angles (rotations / flips) ⇝ easier to visualize.

- The Σ step can flatten a sphere in the same $\mathbb{R}^n$:

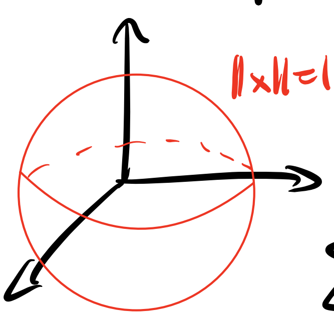


$$\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \Sigma\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2x \\ y \\ 0 \end{pmatrix}$$

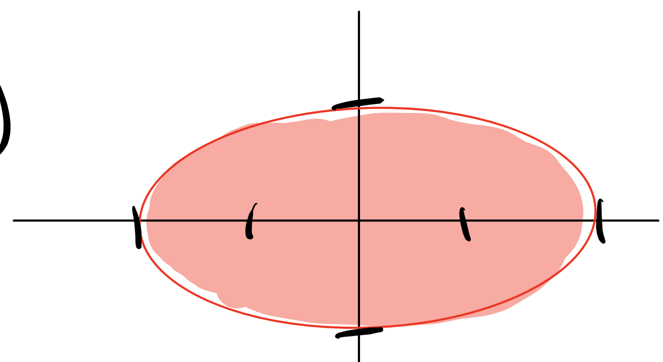


"project onto the xy-plane, then stretch"

- The Σ step can change dimensions:



$$\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \qquad \Sigma\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2x \\ y \end{pmatrix}$$



"project onto the xy-plane, forget the z-coordinate, then stretch"

Geometry of the SVD: Outer Product Form

Here is a geometric interpretation of the SVD that will be useful for the PCA. Let

$$A = \begin{pmatrix} | & & | \\ d_1 & \cdots & d_n \\ | & & | \end{pmatrix} \quad \text{SVD} \quad A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T$$

$$\Rightarrow A v_i = \sigma_i u_i \qquad A^T u_i = \sigma_i v_i$$

Expand out $A^T u_i = \sigma_i v_i$:

$$\sigma_i v_i = A^T u_i = \begin{pmatrix} — d_1^T — \\ \vdots \\ — d_n^T — \end{pmatrix} u_i = \begin{pmatrix} d_1 \cdot u_i \\ \vdots \\ d_n \cdot u_i \end{pmatrix}$$

$$\Rightarrow \sigma_i u_i v_i^T = u_i (\sigma_i v_i)^T = u_i \begin{pmatrix} d_1 \cdot u_i & \cdots & d_n \cdot u_i \end{pmatrix}$$

$$= \begin{pmatrix} | & & | \\ (d_1 \cdot u_i) u_i & \cdots & (d_n \cdot u_i) u_i \\ | & & | \end{pmatrix}$$

NB: $(d \cdot u_i) u_i$ = orthogonal projection of d onto $\operatorname{Span}\{u_i\}$ (since $u_i \cdot u_i = \|u_i\|^2 = 1$).

> The columns of $\sigma_i u_i v_i^T$ are the orthogonal projections of the columns of A onto $\operatorname{Span}\{u_i\}$.

Now look at the sum:

$$A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T$$

The i^{th} column of this sum is:

$$\overset{i^{th}\text{ col of } A}{\longrightarrow} d_i = (d_i \cdot u_1)u_1 + \cdots + (d_i \cdot u_r)u_r$$

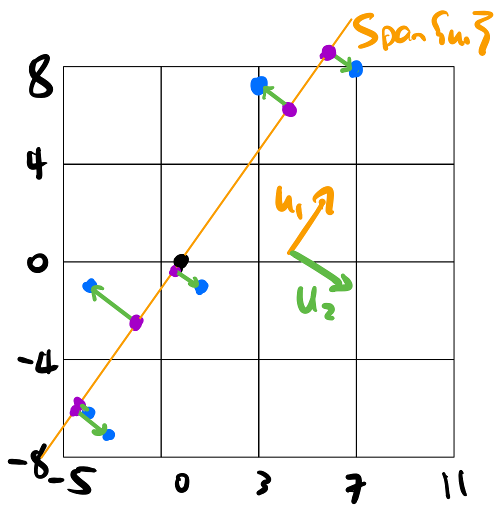
Since $\{u_1, \ldots, u_r\}$ is an orthonormal basis of $\mathrm{Col}(A)$, this is just the projection formula as applied to d_i: the projection of d_i onto $\mathrm{Col}(A)$ is just d_i since $d_i \in \mathrm{Col}(A)$ (it is the i^{th} column of A).

Eg: $A = \begin{pmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{pmatrix}$ $r=2$

$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$

$\sigma_1 \approx 16.9$ $\sigma_2 \approx 3.92$

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



• $= d_i = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \begin{pmatrix} -4 \\ -6 \end{pmatrix}, \ldots$ (columns)

• = columns of $\sigma_1 u_1 v_1^T$

= projections of • onto = Span$\{u_1\}$

↖ = columns of $\sigma_2 u_2 v_2^T$

= projections of • onto = Span$\{u_2\}$

NB: • = • + ↖

So SVD "pulls apart" the columns of A in $u_1, \ldots, u_r$-components

Principal Component Analysis (PCA)

This is "SVD + QO in stats language".

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

→ it makes precise statements about fitting data to lines/planes/etc and how good the fit is.

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

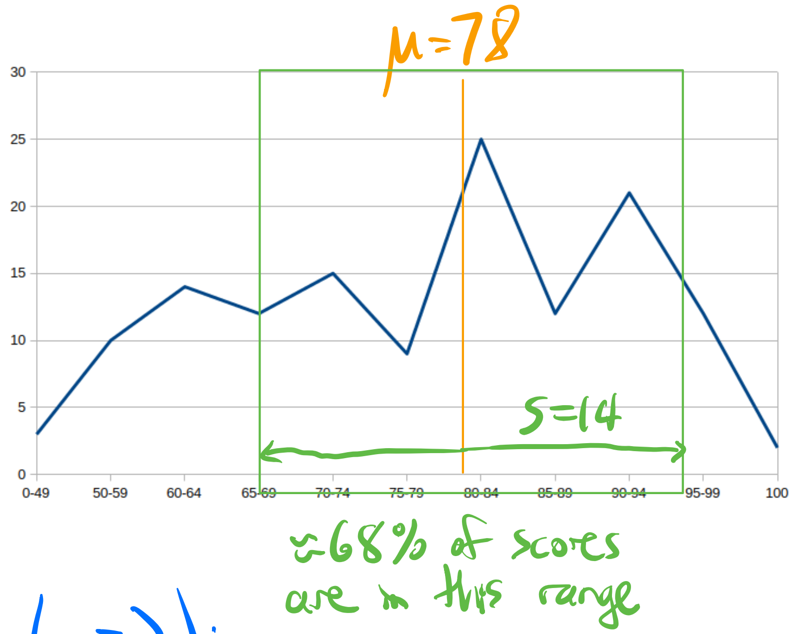
This tells you how "spaced out" the samples are:

$\approx 68\%$ of samples are within $\pm s$ of the mean. (if normally distributed)

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 $\binom{x_1}{y_1}, \dots, \binom{x_n}{y_n}$ $\quad x_i$ = 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: $s_1^2 = \frac{1}{n-1}\left(\bar{x}_1^2 + \dots + \bar{x}_n^2\right)$

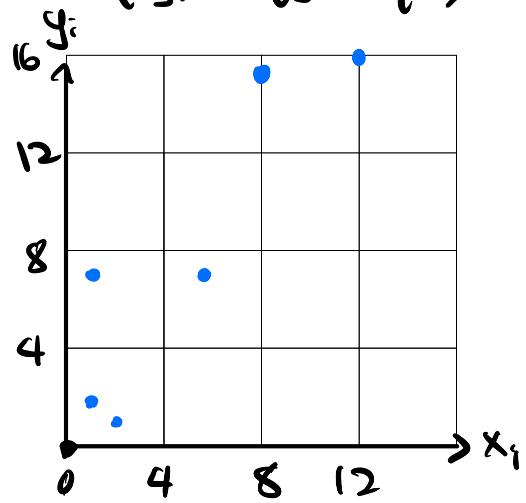
Problem 2: $s_2^2 = \frac{1}{n-1}\left(\bar{y}_1^2 + \dots + \bar{y}_n^2\right)$

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

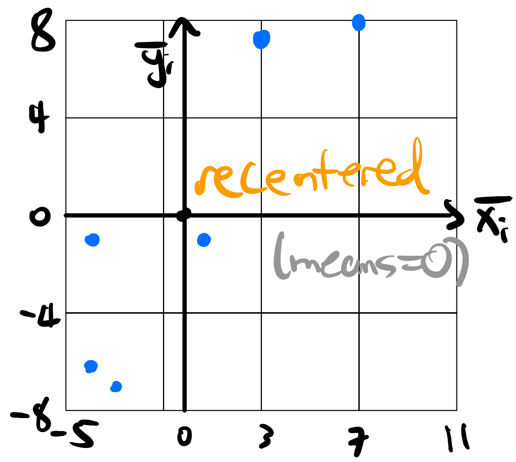
NB: These are just statistics for Problem 1 (x_i) and Problem 2 (y_i) individually – so far we've ignored the fact they might be related. This is what PCA does.

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

recenter: $\begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} 5 \\ 8 \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \begin{pmatrix} -4 \\ -6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -4 \\ -1 \end{pmatrix}, \begin{pmatrix} -3 \\ -7 \end{pmatrix}$



subtract means



Store in matrices:

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

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

NB: Recentered means $\bar{x}_1 + \cdots + \bar{x}_n = 0 = \bar{y}_1 + \cdots + \bar{y}_n$:

The sum of the columns of the recentered data matrix A is zero.

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:

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

The off-diagonal entries are called covariances.

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

$$(\text{row } 1)\cdot(\text{row } 2) = \frac{1}{n-1}\left(\bar{x}_1\bar{y}_1 + \cdots + \bar{x}_n\bar{y}_n\right)$$

- 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.

- If this is almost zero then the values are not correlated.

In our case:

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

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

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

(column means $= 0$)

Def: Let A be a recentered data matrix

$$A = (\vec{d}_1 \cdots \vec{d}_n) \quad \text{where} \quad \vec{d}_i = \begin{pmatrix} \bar{x}_{i1} \\ \vdots \\ \bar{x}_{im} \end{pmatrix} = i^{\text{th}} \text{ recentered data point}$$

Let $S = \frac{1}{n-1} AA^T$ be the covariance matrix.

Let $u \in \mathbb{R}^m$ be a unit vector.

The variance in the u-direction is

$$s(u)^2 = u^T S u$$

NB: $s(u)^2 = u^T \left(\frac{1}{n-1} AA^T\right) u = \frac{1}{n-1} (u^T A)(A^T u) = \frac{1}{n-1} (A^T u)^T (A^T u)$

$$= \frac{1}{n-1} (A^T u) \cdot (A^T u) = \frac{1}{n-1} \|A^T u\|^2.$$

Since $A^T u = \begin{pmatrix} - \vec{d}_1^T - \\ \vdots \\ - \vec{d}_n^T - \end{pmatrix} u = \begin{pmatrix} \vec{d}_1 \cdot u \\ \vdots \\ \vec{d}_n \cdot u \end{pmatrix}$ we get

$$\boxed{s(u)^2 = u^T S u = \frac{1}{n-1} \left((\vec{d}_1 \cdot u)^2 + \cdots + (\vec{d}_n \cdot u)^2 \right)}$$

NB: $\bar{d}_1 + \cdots + \bar{d}_n = 0$ for a recentered data matrix A (p.8).

Hence $0 = 0 \cdot u = (\bar{d}_1 + \cdots + \bar{d}_n) \cdot u = (\bar{d}_1 \cdot u) + \cdots + (\bar{d}_n \cdot u)$

so it makes sense to compute the variance of these numbers $(\bar{d}_1 \cdot u), \ldots, (\bar{d}_n \cdot u)$ with mean zero:

$$s(u)^2 = \frac{1}{n-1}\left((\bar{d}_1 \cdot u)^2 + \cdots + (\bar{d}_n \cdot u)^2\right)$$

Eg: If $u = \begin{pmatrix}1\\0\end{pmatrix} = e_1$ then $\bar{d}_i \cdot u = \begin{pmatrix}\bar{x}_i\\\bar{y}_i\end{pmatrix} \cdot \begin{pmatrix}1\\0\end{pmatrix} = \bar{x}_i$, so

$$s(u)^2 = s(e_1)^2 = \frac{1}{n-1}\left(\bar{x}_1^2 + \cdots + \bar{x}_n^2\right) = s_1^2$$

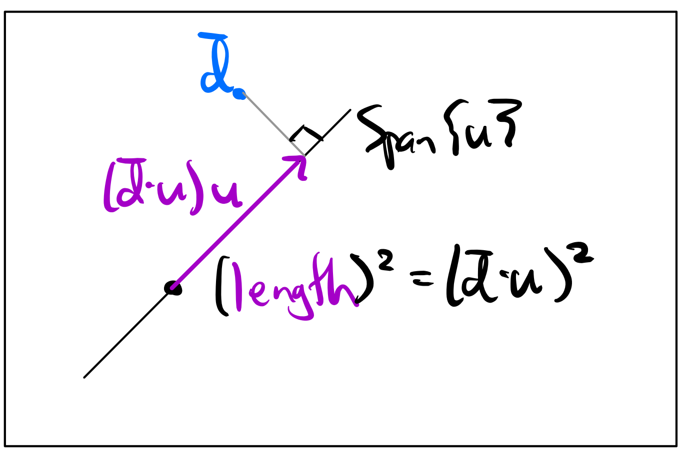
This is just the variance of the xis.

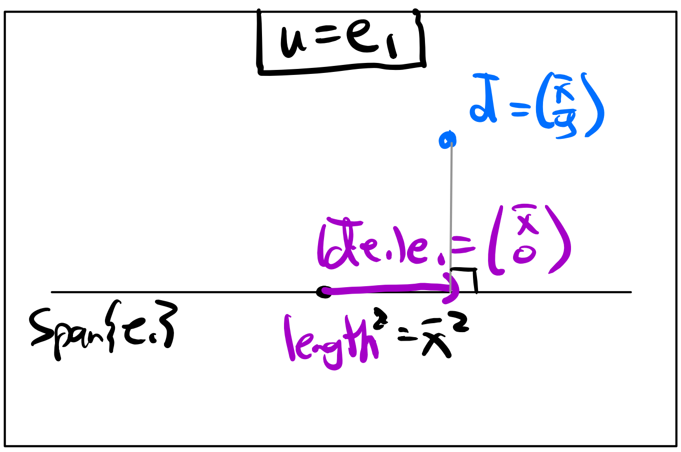
In general, $\boxed{s(e_i)^2 = s_i^2}$

Picture: Recall that if u is a unit vector then

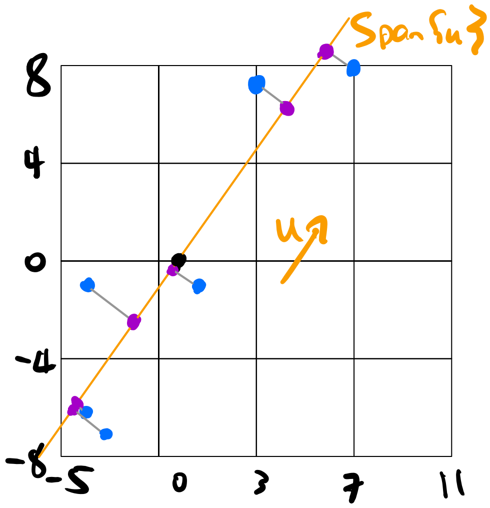
$(v \cdot u)u =$ projection of v onto Span$\{u\}$

$\Longrightarrow (v \cdot u)^2 = (v \cdot u)^2 \|u\|^2 = \|(v \cdot u)u\|^2 =$ length^2 of the projection of v onto Span$\{u\}$





Eg: With our data before, take u in the picture.



● $= \bar{d}_i = \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix}$

● $= (\bar{d}_i \cdot u) u$

$s(u)^2 =$ sum of squares of distances from the ● to zero ●.

Now we apply quadratic optimization to $s(u) = u^T S u$.

Let $\lambda_1 = \sigma_1^2$ be the largest eigenvalue of $S = \frac{1}{n-1} A A^T$.
Let u_1 be a unit λ_1-eigenvector.

Quadratic Optimization:

u_1 maximizes $s(u)^2 = u^T S u$ subject to $\|u\| = 1$
with maximum value σ_1^2

Therefore:

u_1 is the direction of greatest variance
$\sigma_1^2 = s(u_1)^2 =$ variance in the u_1-direction

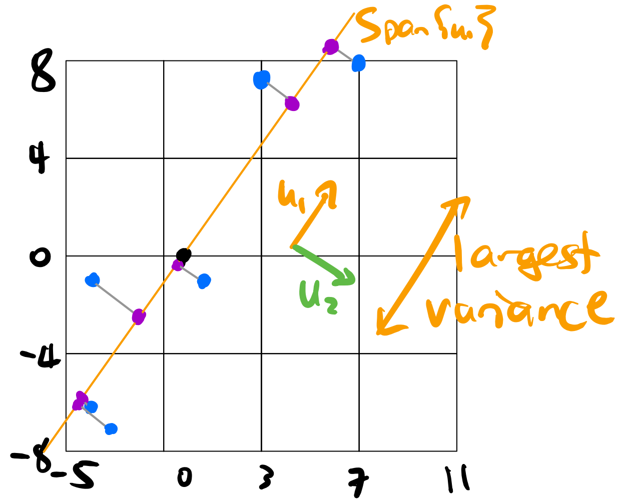
Our data points are "stretched out" most in the u_1-direction.

In our example:

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

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

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



● $= \bar{d}_i$ $\qquad$ ● = projection of ● onto Span$\{u_1\}$

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)