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 2:

samples $x_1, \dots, x_n$

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

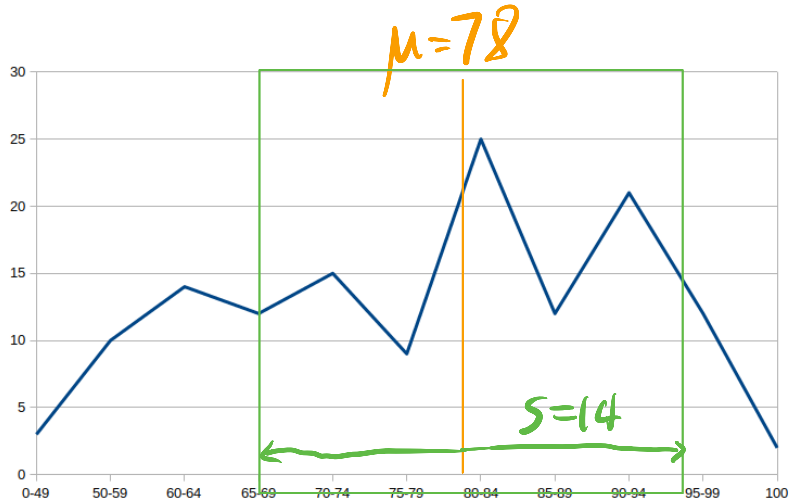
This tells you how "spaced out" the samples are:

$\approx 68\%$ 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 2 scores from Fall '20:



$\approx 68\%$ of scores are in this range

Two Values ($m=2$):

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

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

Mean scores:

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

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

Recenter to compute variance:

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

Variance:

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

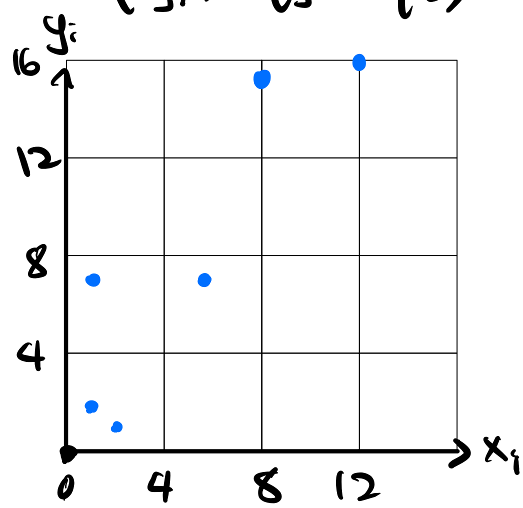
Problem 2: $s_2^2 = \frac{1}{n-1}\left(\bar{y}_1^2 + \cdots + \bar{y}_n^2\right)$

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

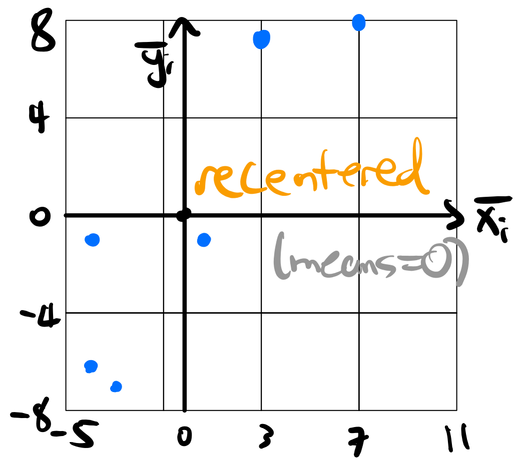
NB: These are just statistics for Problem 1 (x_i) and Problem 2 (y_i) individually — so far we've ignored the fact they might be related. This is what PCA does.

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



$\xrightarrow[\text{means}]{\text{subtract}}$



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:

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

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

Covariance Matrix

columns sum to zero

A: $m \times n$ recentered data matrix

$S = \frac{1}{n-1} AA^T$ covariance matrix

(i,i) entry $= s_i^2$ is the variance of the i^{th} measurement

(i,j) entry is the covariance of the i^{th} & j^{th} measurements

$Tr(S) = s_1^2 + \cdots + s_m^2 =$ total variance

$S = \left(\frac{1}{\sqrt{n-1}} A\right)\left(\frac{1}{\sqrt{n-1}} A\right)^T$, so this looks like we can...

Apply the SVD: Eigenvalues & eigenvectors of

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

compute the SVD of $\frac{1}{\sqrt{n-1}}A$!

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

- $\sigma_1^2 \geq \cdots \geq \sigma_r^2 > 0$ are the nonzero eigenvalues of S
- $u_1, \ldots, u_r$ = orthonormal eigenvectors of S
  = left-singular vectors of $\frac{1}{\sqrt{n-1}}A$ (& of A)
  ↳ $S = \left(\frac{1}{\sqrt{n-1}}A\right)\left(\frac{1}{\sqrt{n-1}}A\right)^T$, not $\left(\frac{1}{\sqrt{n-1}}A\right)^T\left(\frac{1}{\sqrt{n-1}}A\right)$
- $v_i = \frac{1}{\sigma_i} \cdot \frac{1}{\sqrt{n-1}} A^T u_i$
  = right-singular vectors of $\frac{1}{\sqrt{n-1}}A$ (& of A)

NB: the SVD of A is

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

$\Rightarrow$ the singular values of A are $\sqrt{n-1}\,\sigma_1, \ldots, \sqrt{n-1}\,\sigma_r$

Fact: The trace of a square matrix is the sum of its eigenvals
(HW)

$$\Rightarrow \boxed{\text{total variance} = s^2 = \mathrm{Tr}(S) = \sigma_1^2 + \cdots + \sigma_r^2}$$

What do the singular values & singular vectors tell us about our data? They give the directions and amounts of largest & smallest variance.

(columns sum to 0)

Def: Let A be a recentered data matrix

$A = (\vec{d}_1 \cdots \vec{d}_n)$ where $\vec{d}_i = \begin{pmatrix} \bar{x}_{i1} \\ \vdots \\ \bar{x}_{im} \end{pmatrix}$ = i^{th} 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: $s(u)^2 = u^T\left(\frac{1}{n-1}AA^T\right)u = \frac{1}{n-1}(u^TA)(A^Tu) = \frac{1}{n-1}(A^Tu)^T(A^Tu)$

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

Since $A^Tu = \begin{pmatrix} -\vec{d}_1^T- \\ \vdots \\ -\vec{d}_n^T- \end{pmatrix} u = \begin{pmatrix} \vec{d}_1 \cdot u \\ \vdots \\ \vec{d}_n \cdot u \end{pmatrix}$ 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: $\vec{d}_1 + \cdots + \vec{d}_n = 0$ for a recentered data matrix A.

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

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

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

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

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

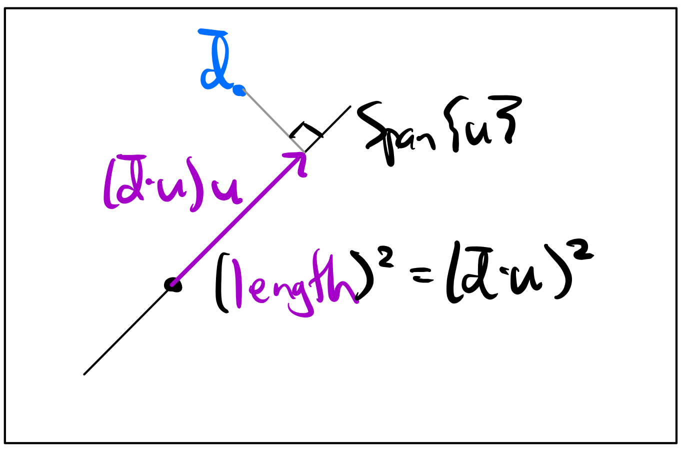
This is just the variance of the xis.

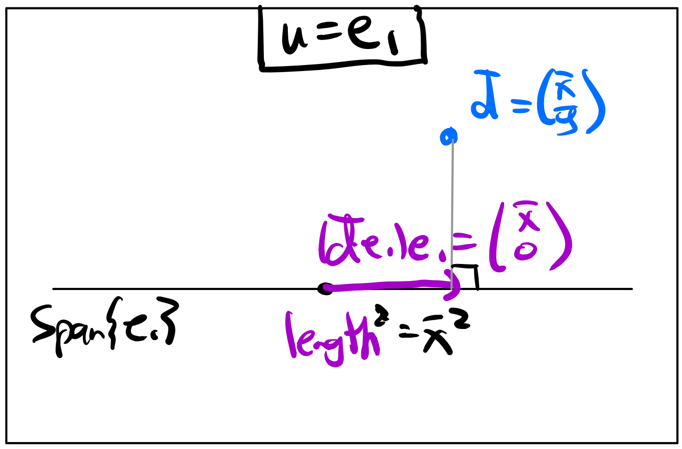
In general, $\boxed{s(e_i)^2 = s_i^2}$

Picture: Recall that if u is a unit vector then

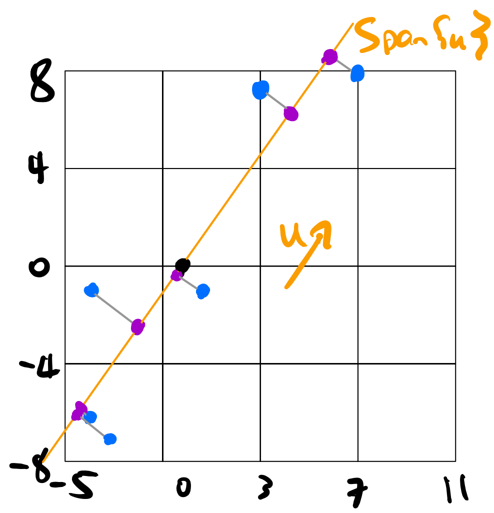
$(v \cdot u)u$ = projection of v onto $\mathrm{Span}\{u\}$

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





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

The largest eigenvalue of $S = \frac{1}{n-1} AA^T$ is σ_1^2, and u_1 is a unit σ_1^2-eigenvector.

Quadratic Optimization:

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

Therefore:

singular vector of $\frac{1}{\sqrt{n-1}} A$ → u_1 is the direction of greatest variance

singular value of $\frac{1}{\sqrt{n-1}} A$ → $\boxed{\sigma_1^2 = s(u_1)^2 = \text{variance in the } u_1\text{-direction}}$

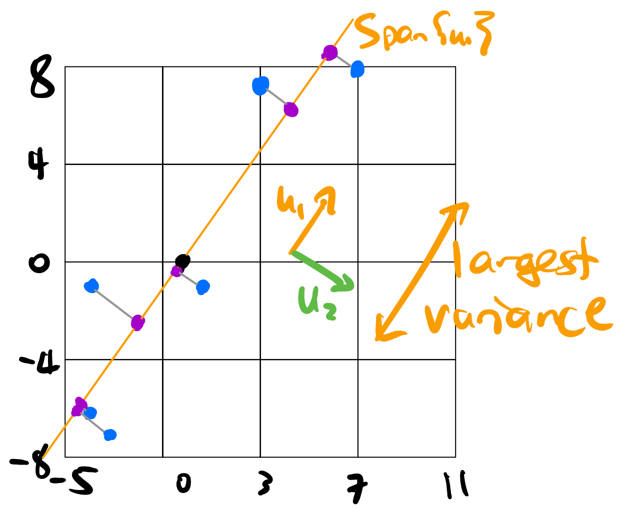
Our data points are "stretched out" most in the u_1-direction.

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 variance is maximized in the u_1 direction, 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 (problem) averages $(\mu_1, \dots, \mu_m)$ to recenter
  $\rightsquigarrow$ matrix $A = \begin{pmatrix} | & & | \\ d_1 & \cdots & d_n \\ | & & | \end{pmatrix}$
- Compute the 1^{st} principal component u_1
- The score for student i is

$$d_i \cdot u_1 + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} \quad \text{(add back the means)}$$

This maximizes the standard deviation by reweighting the problems according to u_1.

We know that u_1 is the direction of largest variance.
What about $u_2, \dots, u_r$?

QO with Extra Constraints:

- $s(u)^2 = u^T S u$ is maximized
  subject to $\|u\|=1$
  at u_1 with $s(u_1)^2 = \sigma_1^2$
  $\rightarrow$ u_1 is the direction with largest variance

- $s(u)^2$ is maximized
  subject to $\|u\|=1$ and $u \perp u_1$
  at u_2 with $s(u_2)^2 = \sigma_2^2$
  $\rightarrow$ u_2 is the direction with 2^{nd}-largest variance

- $s(u)^2$ is maximized
  subject to $\|u\|=1$ and $u \perp u_1, \ldots, u \perp u_{i-1}$
  at u_i with $s(u_i)^2 = \sigma_i^2$
  $\rightarrow$ u_i is the direction with i^{th}-largest variance

NB: if A has full row rank $(r=m)$ then

- $s(u)^2 = u^T S u$ is minimized
  subject to $\|u\|=1$
  at u_r with $s(u_r)^2 = \sigma_r^2$
  $\rightarrow$ u_r is the direction with smallest variance

(If A does not have full row rank then $s(u)^2 = 0$ for any $u \in \mathrm{Nul}(A^T) \neq \{0\}$.)

Principal Components

The columns of $\sqrt{n-1}\,\sigma_i u_i v_i^T$ are the orthogonal projections of the columns of A onto $\text{Span}\{u_i\}$.

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

"breaks apart" your data points into principal components.

Def: Let A be a recentered data matrix with SVD

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

The i^{th} principal component of A is $\sqrt{n-1}\,\sigma_i u_i v_i^T$.

The columns of the i^{th} principal component of A are the orthogonal projections of the columns of A onto $\text{Span}\{u_i\}$ = direction of i^{th}-largest variance.

The variance of the lengths of the i^{th} principal component of A is $\sigma_i^2 = i^{th}$-largest variance.

Eg: In our example, $\frac{1}{\sqrt{6-1}} A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$

$\sigma_1^2 \approx 56.9$ $\qquad$ $\sigma_2^2 \approx 3.07$ $\qquad$ $S = \begin{pmatrix} 20 & 25 \\ 25 & 40 \end{pmatrix}$ $\qquad$ $s_1^2 = 20$

$u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$ $\qquad$ $u_2 \approx \begin{pmatrix} 0.828 \\ -0.561 \end{pmatrix}$ $\qquad\qquad\qquad\qquad$ $s_2^2 = 40$

Total variance: $\sigma_1^2 + \sigma_2^2 = 56.9 + 3.1 = 60 = 20 + 40$

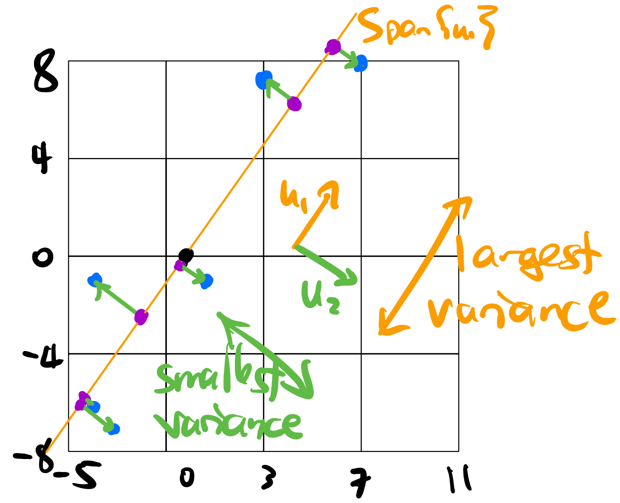
● $= \bar{d}_i$

● = columns of $\sqrt{5}\,\sigma_1 u_1 v_1^T$

= projections of ● onto Span$\{u_1\}$

↖ = columns of $\sqrt{5}\,\sigma_2 u_2 v_2^T$

= projections of ● onto Span$\{u_2\}$



NB: In this case, $s(u)^2$ is minimized at u_2 with minimum value σ_2^2 = smallest eigenvalue of S.

$$s(u_2)^2 = \frac{1}{n-1}\left[(d_1 \cdot u_2)^2 + \cdots + (d_n \cdot u_2)^2\right]$$
$$= \frac{1}{n-1}\left[\text{sum of squares of lengths of } \nwarrow\right]$$

Conclusion: The direction of largest variance is the line of best fit in the sense of orthogonal least squares, and the

$$(\text{error})^2 = (\text{sum of squares of lengths of } \nwarrow)$$
$$= (n-1)s(u_2)^2 = (n-1)\sigma_2^2$$