A Little Bit of Statistics

Principal Component Analysis is basically an interpretation of the SVD + QO in the language of statistics.

→ This is often how the SVD, or "linear algebra", is used in statistics and data science.

→ It makes precise statements about lines/planes/etc. of best fit, and how good the fit is.

To that end, we need a bit of terminology from statistics.

Idea: an $m \times n$ matrix stores n samples, each containing m values or measurements.

One value ($m=1$):

Let's record everyone's scores on midterm 1:

samples $x_1, x_2, \ldots, x_n$ (n = # students)

- The **mean** (average) of the samples is
$$\mu = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$
- The **variance** of the samples is
$$s^2 = \frac{1}{n-1}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2\right]$$
- The **standard deviation** is $s = \sqrt{s^2}$.

The standard deviation tells you how "spread out" your values are from the mean:

$\approx$ 68% of samples will be within $\pm s$ of μ

$\approx$ 95% of samples will be within $\pm 2s$ of μ

$\approx$ 99% of samples will be within $\pm 3s$ of μ

(if your data are normally distributed...)

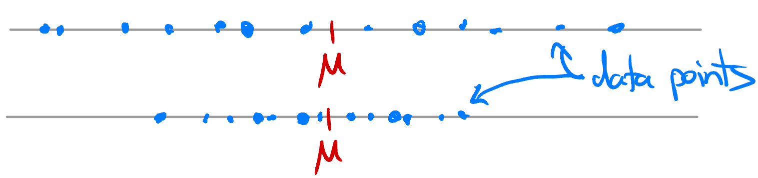
$\to$ Where are these formulas from?

A statistics class!

NB: The recentered values $(x_1-\mu), (x_2-\mu), \ldots, (x_n-\mu)$ have mean $\mu - \mu = 0$.

Eg: large s:

small s:



Eg: Here is a histogram of midterm 2 scores from fall '20:

Two Values ($m=2$):

Now let's record everyone's scores on problem 1 and problem 2 on midterm 2: samples

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

x_i = score on problem 1
y_i score on problem 2

- **Mean scores:**

$$\mu_1 = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \text{mean of problem 1}$$

$$\mu_2 = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \text{mean of problem 2}$$

- **Recenter** to compute variance:

$$\bar{x}_i = x_i - \mu_1 \qquad \bar{y}_i = y_i - \mu_2 \qquad (\text{now mean} = 0)$$

- **Variances:**

$$s_1^2 = \frac{1}{n-1}\left(\bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_n^2\right)$$

$$s_2^2 = \frac{1}{n-1}\left(\bar{y}_1^2 + \bar{y}_2^2 + \cdots + \bar{y}_n^2\right)$$

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

NB: Except for the total variance, these are just statistics for Problems 1 and 2 individually — so far we've ignored the fact that they might be correlated. This is what the PCA does.

Running Example: Suppose the problem 1 & 2 scores are:

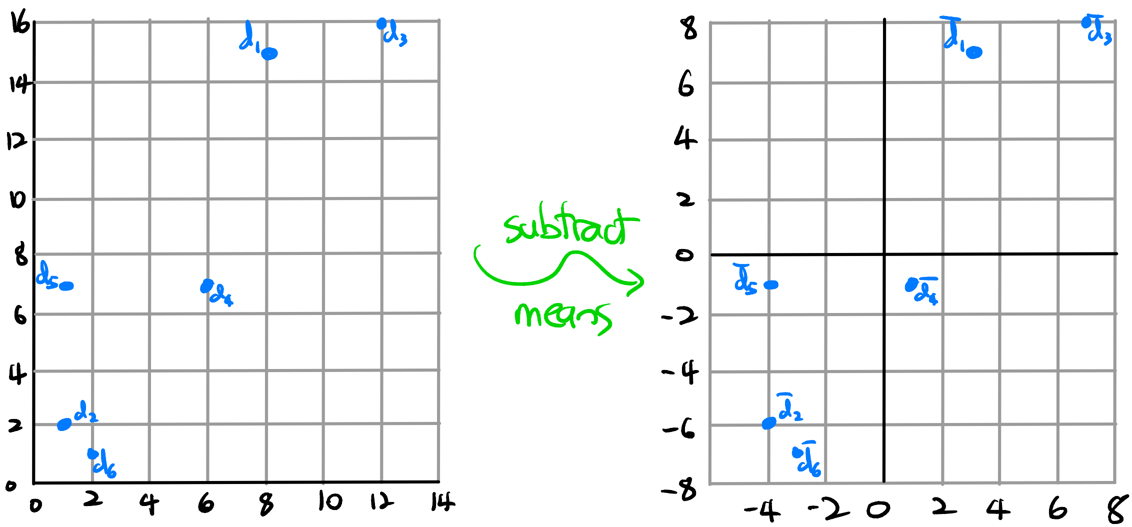
$$d_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} 8 \\ 15 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 12 \\ 16 \end{pmatrix}, \begin{pmatrix} 6 \\ 7 \end{pmatrix}, \begin{pmatrix} 1 \\ 7 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \mu_1 = 5 \quad m = 2$$
$$\mu_2 = 8 \quad n = 6$$

Recenter: subtract $\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} \rightsquigarrow$

$$\bar{d}_i = \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \begin{pmatrix} -4 \\ -6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -4 \\ -1 \end{pmatrix}, \begin{pmatrix} -3 \\ -7 \end{pmatrix} \qquad s_1^2 = 20 \quad s_2^2 = 40 \quad s^2 = 60$$

total variance

Geometrically, subtracting $\begin{pmatrix} 5 \\ 8 \end{pmatrix}$ moves the origin to $\begin{pmatrix} 5 \\ 8 \end{pmatrix}$:



NB: The recentered values have mean zero, so

$$\bar{d}_1 + \cdots + \bar{d}_6 = \begin{pmatrix} 3 - 4 + 7 + 1 - 4 - 3 \\ 7 - 6 + 8 - 1 - 1 - 7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Principal Component Analysis

Now we do linear algebra.

Suppose we have n samples of n values: data points

$$d_1, d_2, \ldots, d_n$$

Store in a matrix:

$$A_0 = \begin{pmatrix} | & & | \\ d_1 & \cdots & d_n \\ | & & | \end{pmatrix}$$

Recenter the data points (subtract the means of each row):

$$A = \begin{pmatrix} | & & | \\ \bar{d}_1 & \cdots & \bar{d}_n \\ | & & | \end{pmatrix}$$

NB: Each value now has mean zero. This means

$$\boxed{\bar{d}_1 + \bar{d}_2 + \cdots + \bar{d}_n = 0.}$$

Def: The covariance matrix is $S = \frac{1}{n-1} A A^T$.

This contains the dot products of the rows.

For example, if $m=2$ and

$$A = \begin{pmatrix} | & & | \\ \vec{d}_1 & \cdots & \vec{d}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} \overline{x}_1 & \cdots & \overline{x}_n \\ \overline{y}_1 & \cdots & \overline{y}_n \end{pmatrix}$$

then

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

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

The diagonal entries are the variances:

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

The trace is the total variance:

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

The off-diagonal entries are called covariances.

Essentially, if the $(1,2)$ entry

$$\frac{1}{n-1}\left(\overline{x}_1\overline{y}_1+\cdots+\overline{x}_n\overline{y}_n\right)$$

is large then $\overline{x}_i$ and $\overline{y}_i$ tend to have the same sign: so if the first measurement is above average, then the second probably is too. Likewise, if the $(1,2)$-entry is large negative, then the opposite is true.

We won't use covariances directly for anything.

Running Example:

$$d_i = \begin{pmatrix}8\\15\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}12\\16\end{pmatrix}, \begin{pmatrix}6\\7\end{pmatrix}, \begin{pmatrix}1\\7\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix} \rightsquigarrow A_0 = \begin{pmatrix}8 & 1 & 12 & 6 & 1 & 2\\15 & 2 & 16 & 7 & 7 & 1\end{pmatrix}$$

$$\hat{d}_i = \begin{pmatrix}3\\7\end{pmatrix}, \begin{pmatrix}-4\\-6\end{pmatrix}, \begin{pmatrix}7\\8\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}-4\\-1\end{pmatrix}, \begin{pmatrix}-3\\-7\end{pmatrix} \rightsquigarrow A = \begin{pmatrix}3 & -4 & 7 & 1 & -4 & 3\\7 & -6 & 8 & -1 & -1 & -7\end{pmatrix}$$

$$S = \frac{1}{6-1} AA^T = \begin{pmatrix}20 & 25\\25 & 40\end{pmatrix} \qquad s_1^2 = 20 \quad s_2^2 = 40$$

Total variance: $\mathrm{Tr}(S) = 20 + 40 = 60$

Covariance $= 25$: the values are correlated.

Covariance Matrix: Summary

A: $m \times n$ recentered data matrix.

$$\boxed{S = \frac{1}{n-1} AA^T} = \text{covariance matrix } (m \times m)$$

The (i,i)-entry of S is

s_i^2 = variance of the i^{th} value

The trace of S is the total variance:

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

The (i,j)-entry of S is the covariance of the i^{th} & j^{th} values.

APPLY THE SVD:

The eigenvalues and 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 $\frac{1}{\sqrt{n-1}} A$ (and $\frac{1}{\sqrt{n-1}} A^T$).

$$\boxed{\frac{1}{\sqrt{n-1}} A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T}$$

where:

- $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 > 0$ are the nonzero eigenvalues of S.
- $u_1, u_2, \ldots, u_r$ are orthonormal eigenvectors of S $=$ left-singular vectors of $\frac{1}{\sqrt{n-1}} A$

  $\hookrightarrow S = \left(\frac{1}{\sqrt{n-1}} A\right)\left(\frac{1}{\sqrt{n-1}} A\right)^T$, not $\left(\frac{1}{\sqrt{n-1}} A\right)^T\left(\frac{1}{\sqrt{n-1}} A\right)$
- $v_1, v_2, \ldots, v_r$ are the right-singular vectors of $\frac{1}{\sqrt{n-1}} A$

As always:

$$u_i = \frac{1}{\sigma_i} \cdot \frac{1}{\sqrt{n-1}} A v_i \qquad v_i = \frac{1}{\sigma_i} \cdot \frac{1}{\sqrt{n-1}} A^T u_i$$

NB: The SVD of A is just

$$\boxed{A = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,\sigma_2 u_2 v_2^T + \cdots + \sqrt{n-1}\,\sigma_r u_r v_r^T}$$

$\rightsquigarrow$ same singular vectors of $\frac{1}{\sqrt{n-1}} A$, but the singular values are $\sqrt{n-1}\,\sigma_1, \ldots, \sqrt{n-1}\,\sigma_r$.

We need to keep the $\frac{1}{n-1}$'s around so that, for instance, we have $Tr(S) = (\text{total variance})$.

Fact: The trace of a square matrix is the sum of its eigenvalues, counted with algebraic multiplicity:

$$Tr\left[C\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{pmatrix}C^{-1}\right] = \lambda_1+\lambda_2+\cdots+\lambda_n$$

(This was an optional HW problem.)

Apply this to S: we know $Tr(S) = (\text{total variance})$, so

$$\boxed{s_1^2+s_2^2+\cdots+s_m^2 = s^2 = Tr(S) = \sigma_1^2+\sigma_2^2+\cdots+\sigma_r^2}$$

Q: Ok, so what do the singular values & singular vectors of $\frac{1}{\sqrt{n-1}}A$ tell us about our data?

A: The directions and magnitudes of largest and smallest variance.

The rest of this lecture is devoted to decoding that sentence.

Def: Let A be a *recentered* data matrix with covariance matrix $S = \frac{1}{\sqrt{n-1}} AA^T$, and let $u \in \mathbb{R}^m$ be a unit vector. The **variance in the u-direction** of our data points is

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

This is a slick definition that obviously suggests quadratic optimization, but let's unpack what it means.

We've seen $x^T (A^T A) x$ before (L21, L23). In this case,

$$s(u)^2 = u^T S u = u^T \cdot \frac{1}{n-1} AA^T u = \frac{1}{n-1} (u^T A)(A^T u)$$

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

If A has columns $\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_n$ then

$$A^T u = \begin{pmatrix} - \vec{d}_1^T - \\ \vdots \\ - \vec{d}_n^T - \end{pmatrix} u = \begin{pmatrix} \vec{d}_1 \cdot u \\ \vdots \\ \vec{d}_n \cdot u \end{pmatrix}, \quad \text{so}$$

$$\boxed{s(u)^2 = \frac{1}{n-1} \|A^T u\|^2 = \frac{1}{n-1} \left[(\vec{d}_1 \cdot u)^2 + (\vec{d}_2 \cdot u)^2 + \cdots + (\vec{d}_n \cdot u)^2 \right]}$$

NB: Since A is a recentered data matrix, we have

$$\bar{d}_1 + \bar{d}_2 + \cdots + \bar{d}_n = 0,$$

so

$$0 = 0 \cdot u = (\bar{d}_1 + \bar{d}_2 + \cdots + \bar{d}_n) \cdot u = \bar{d}_1 \cdot u + \bar{d}_2 \cdot u + \cdots + \bar{d}_n \cdot u$$

Therefore, $s(u)^2$ is the variance of the numbers $\bar{d}_1 \cdot u, \bar{d}_2 \cdot u, \ldots, \bar{d}_n \cdot u$ with mean zero.

Eg: If $u = e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ then

$$\bar{d}_i \cdot e_1 = \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \bar{x}_i$$

so
$$s(e_1)^2 = \frac{1}{n-1}\left[(\bar{d}_1 \cdot e_1)^2 + (\bar{d}_2 \cdot e_1)^2 + \cdots + (\bar{d}_n \cdot e_1)^2 \right]$$
$$= \frac{1}{n-1}\left(\bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_n^2 \right) = s_1^2$$

More generally,

$$\boxed{s(e_i)^2 = s_i^2 = \text{variance of the } i^{\text{th}} \text{ value}}$$

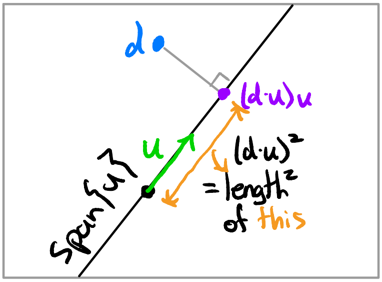
For a general unit vector u, recall that the orthogonal projection of d onto $\mathrm{Span}\{u\}$ is $(d \cdot u)u$, so that

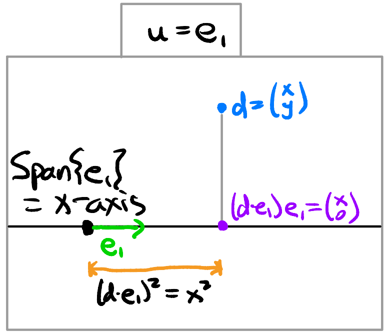
$$\|(d \cdot u)u\|^2 = (d \cdot u)^2 \|u\|^2 = (d \cdot u)^2.$$

In other words,

$$(d \cdot u)^2 = \text{length}^2 \text{ of the projection of } d \text{ onto } \mathrm{Span}\{u\}.$$

Picture:

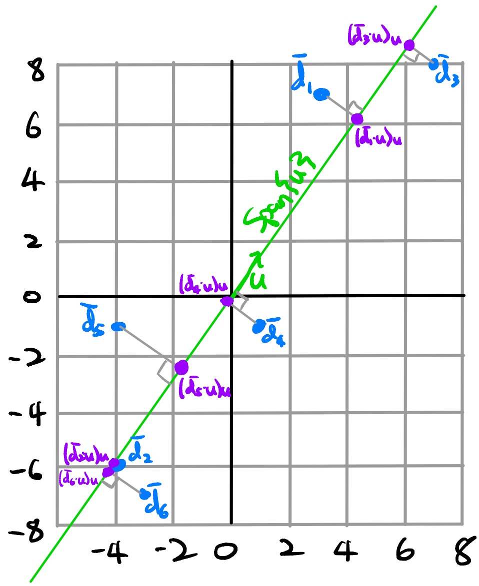




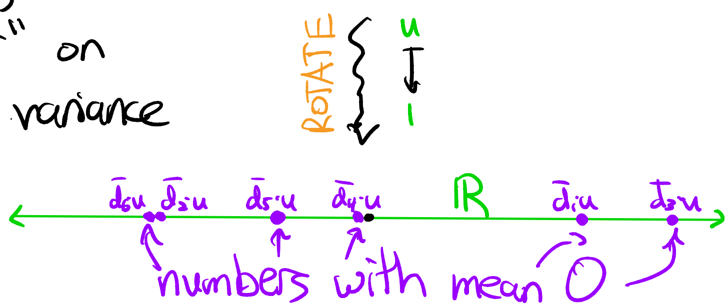
Running Example:

In this case,

$s(u)^2$ = sum of squares of distances of • from zero.



Here's another way to think about it: Span{u} is a "number" line, and $(\bar d_i\cdot u)u$ is a "number" on it. Then $s(u)^2$ = the variance of these numbers.



APPLY QUADRATIC OPTIMIZATION

The quadratic form $s(u)^2 = u^T S u$ has maximum value (subject to $\|u\|=1$) $= \sigma_1^2 =$ largest eigenvalue of S. It attains its maximum at $u_1 =$ unit σ_1^2-eigenvector. Therefore:

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

(Remember that σ_1 is the first singular value of $\frac{1}{\sqrt{n-1}}A$ and u_1 is the first left singular vector.)

This says our data points are "stretched out" the most in the u_1-direction.

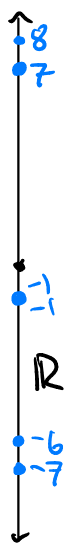
Running Example:

In our example, $\sigma_1^2 = 56.9$ and $u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$.
The variance is maximized in the u_1-direction with max variance 56.9.

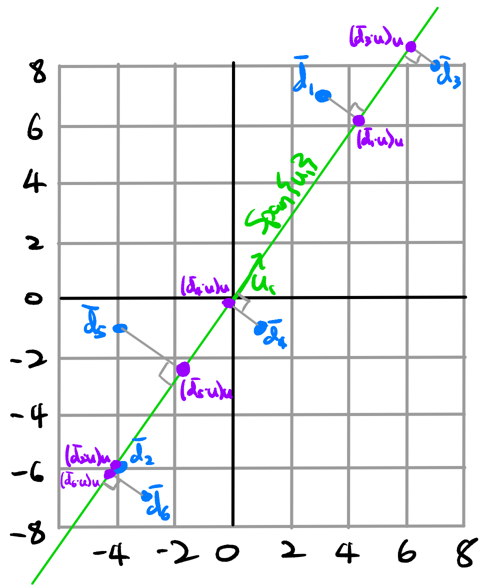
Note this is greater than

$$s_1^2 = 20 = \text{problem 1 variance}$$
$$s_2^2 = 40 = \text{problem 2 variance.}$$

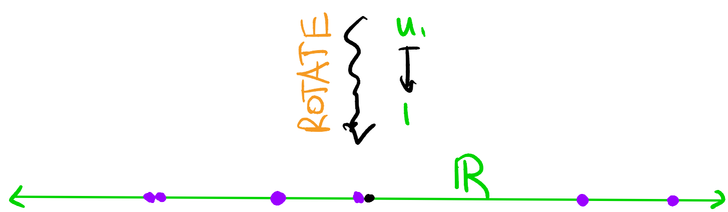
$s(c_2)^2 = s_2^2$ = variance = 40



y-coords
(Problem 2)



these are more spread out



$s(u_1)^2 = \text{variance} \approx 56.9$

Eg: 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 to recenter

$$\rightsquigarrow \text{matrix } A = \begin{pmatrix} | & & | \\ \bar{d}_1 & \cdots & \bar{d}_n \\ | & & | \end{pmatrix}$$

- Compute the first left singular vector u_1
- The score for student i is

$$\bar{d}_i \cdot u_1 + (\text{mean scores})$$

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

Of course, this isn't necessarily fair. For instance, if the j^{th} coordinate of u_1 is negative, then you're penalized for getting problem j correct!

Minimum Variance:

If A has FRR, then $s(u)^2$ has minimum value (subject to $\|u\|=1$) $= \sigma_r^2 =$ smallest eigenvalue of S.
It is minimized at $u_r =$ unit σ_r^2-eigenvector.
Therefore:

$\boxed{\begin{array}{l} u_r \text{ is the direction of smallest variance} \\ \sigma_r^2 = s(u_r)^2 = \text{variance in the } u_r\text{-direction} \end{array}}$

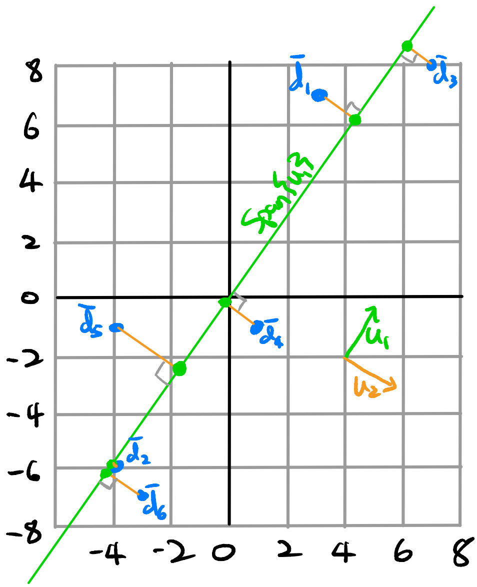
(If A does not have FRR then $s(u)^2$ has minimum value zero, attained at any unit vector in $\mathrm{Nul}(A^T)$.)

Running Example:

In our case,

$\sigma_2^2 \approx 3.07 \quad u_2 \approx \begin{pmatrix} .828 \\ -.561 \end{pmatrix}$

The variance in the u_2-direction is minimized $\Longrightarrow$ the sum of the length^2 of the projections is minimized.



But the length of is the orthogonal distance of the data point from $= \mathrm{Span}\{u_1\} = \mathrm{Span}\{u_2\}^\perp$.

Conclusion: In this case,

the direction of maximum variance

$=$ the line of best fit in the sense of orthogonal least squares

and the $\text{error}^2 = \sum \text{distance}^2$ from $= s(u_2)^2$