Geometry of the SVD: Matrix Form
We have drawn pictures of a triple product decomposition before.

Diagonalization:

$$
A=\frac{1}{10}\left(\begin{array}{cc}
11 & 6 \\
9 & 14
\end{array}\right)=C D C^{-1}
$$

for $C=\left(\begin{array}{cc}\omega_{1} & \omega_{2} \\ 3 & 1\end{array}\right) \quad D=\left(\begin{array}{cc}\lambda_{1} & \lambda_{2} \\ 2 & 0 \\ 0 & 1 / 2\end{array}\right)$
To evaluate $A_{x}=C C^{-1} x$ :
(1) multiply by $\mathrm{C}^{-1}$
(2) multiply by $D$
(3) multiply by $C$


$$
\begin{array}{r}
\begin{array}{c}
(1) \\
C^{-1}
\end{array}{ }^{C^{-1}=e_{1}}=e_{1} \\
C^{-1} \omega_{2}=e_{2} \\
C^{-1} x=\binom{1}{2}
\end{array}
$$


(2) $\downarrow D$


$$
\left.\xrightarrow\left[\begin{array}{l}
C e_{1}=w_{1} \\
C e_{2}=w_{2} \\
C \\
C
\end{array} l_{1}^{2}\right)=2 \omega_{1}+w_{2}\right]{\substack{(3)}}
$$

$$
A_{\omega_{2}}=\frac{1}{2} \omega_{2} \downarrow \uparrow A_{x}-2 \omega_{1}+\omega_{2}
$$


$S V D: A=\left(\begin{array}{ll}3 & 0 \\ 4 & 5\end{array}\right)=U \Sigma V^{\top} \quad$ for

$$
U=\frac{1}{\sqrt{10}}\left(\begin{array}{cc}
u_{1} & u_{2} \\
1 & -3 \\
3 & 1
\end{array}\right) \quad V=\frac{1}{\sqrt{2}}\left(\begin{array}{cc}
v_{1} & v_{2} \\
1 & -1 \\
1 & 1
\end{array}\right) \quad \sum=\left(\begin{array}{cc}
\sigma_{2} & \sigma_{2} \\
\sqrt[35]{5} & 0 \\
0 & \sqrt{5}
\end{array}\right)
$$

To evaluate $A_{x}=U \Sigma V^{\top} x$ :
(1) multiply by $V^{\top}$
(2) multiply by $\Sigma$
(3) multiply by $U$

But $U$ and $V^{T}$ are orthogond, so these just rotate/ flip.
$A_{x}=(1)$ rotate/ flip (2) stretch (3) rotate/flip

(ovate $\mathrm{CW} 45^{\circ}$ )

$$
\begin{aligned}
& V^{\top}=V^{\top} \\
& V^{\top} v_{v_{1}}=e_{1} \\
& V^{\top} V_{2}=e_{2} \\
& V_{x}=\binom{-1}{1}
\end{aligned}
$$



Indole ccu o by $\left.\arctan (31)=シ 5^{\circ}\right)$


Notes / caveats:

- Diagonalization: start \& end in $\left\{\omega_{1}, \omega_{2}\right\}$ basis SVD: start with $\left\{v_{1}, v_{2}\right\}$ \& end with $\left\{u_{1} u_{2}\right\}$ basis $\rightarrow$ Different bases!
- The $V^{\top} \& U$ steps preserve lengths \& angles (rotations / flips) $u$ easier to visualize.
- The $\sum$ step can flatten a sphere in the same $\mathbb{R}^{n}$ :


$$
\frac{\Sigma=\left(\begin{array}{lll}
2 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{array}\right)}{\sum\left(\begin{array}{l}
x \\
z \\
z
\end{array}\right)=\left(\begin{array}{l}
2 x \\
y \\
0
\end{array}\right)}
$$


"project onto the $x y$-plane, then stretch"

- The $\sum$ step can change dimensions:


$$
\frac{\sum=\left(\begin{array}{lll}
2 & 0 & 0 \\
0 & 1 & 0
\end{array}\right)}{\sum\left(\begin{array}{l}
x \\
y \\
z
\end{array}\right)=\binom{2 x}{y}}
$$

"project onto the $x y$-planes 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

$$
\begin{aligned}
A=\left(d_{r} \cdots d_{n}^{\prime}\right) & \quad \text { SUD } \quad A=\sigma_{i} u_{1} v_{i}^{\top}+\cdots+\sigma_{r} u v_{1} v_{r}^{\top} \\
& \Rightarrow A v_{i}=\sigma_{i} u_{i} \quad A^{\top} u_{i}=\sigma_{i} v_{i}
\end{aligned}
$$

Expand out $A^{\top} u_{i}=\sigma_{i} v_{i}$ :

$$
\begin{aligned}
\sigma_{i} v_{i} & =A^{\top} u_{i}=\binom{-d_{i}^{\top}-}{-d_{i}^{T}-} u_{i}=\binom{d_{i} u_{i}}{d_{i} u_{i}} \\
\Rightarrow \sigma_{i} u_{i} v_{i}^{\top} & =u_{i}\left(\sigma_{i} v_{i} v^{\top}=u_{i}\left(d_{i} u_{i} \cdots d_{i} \cdot u_{i}\right)\right. \\
& =\left(\begin{array}{ccc}
\left(d_{i}^{\prime} u_{i}\right) u_{i} & \cdots & \left(d_{i} u_{i} u_{i} u_{i}\right. \\
1
\end{array}\right)
\end{aligned}
$$

NB: $\left(d \cdot u_{i}\right) u_{i}=$ ortheganal projection of $d$ onto Span $\left\{u_{i}\right\}$ (since $u_{i} \cdot u_{i}=\left\|u_{i}\right\|^{2}=1$ ).
The columns of $a_{i} u_{i} v_{i}^{\top}$ are the orthogonal projections of the columns of $A$ onto Span \{is\} . ~
Now look at the sum:

$$
A=\sigma_{i} u_{1} v_{c}^{\top}+\cdots+\sigma_{r} u_{n} v_{r}{ }^{\top}
$$

The $i^{\text {th }}$ colum of this sum is:

$$
\operatorname{lin}_{\text {of }}{ }^{\circ 1} \rightarrow d_{i}=\left(d_{i} \cdot u_{i}\right) u_{1}+\cdots+\left(d_{i} \cdot u_{r}\right) u_{r}
$$

Sine $\left\{u_{1}, \ldots w_{3}\right\}$ is an orthonomal basis of $\operatorname{Col}(A)$, this is just the projection formula as applied to $d_{i}$ : the projection of $d_{i}$ onto $\operatorname{Col}(A)$ is just $d_{i}$ since $d_{i} \in C_{0}(A)$ lit is the th column of $A$ ).

$$
\begin{gathered}
E g: A=\left(\begin{array}{ccccc}
3 & -4 & 7 & -4 \\
7 & -6 & 8 & -1 & -1 \\
E & -7
\end{array}\right) r=2 \\
A=o_{1} u_{1} v_{1}^{\top}+o_{2} u_{2} v_{2}^{\top} \\
0_{1} \approx 16.9 \\
o_{2} \approx 3.92
\end{gathered}
$$

- = columns of $\sigma_{1} u_{1} v_{1}^{\top}$

$=$ projections of $\cdot$ onto $/=\operatorname{Span}\{u . i$
$=$ columns of $\sigma_{1} u_{2} v_{2}^{\top}$
$=$ projections of $\cdot$ onto $\rangle=\operatorname{Span}\left\{u_{2}\right\}$

$$
N B: \quad=+\lambda
$$

So SVD "pulls apart" the columns of $A$ in $u_{1, \ldots, u_{1}-}$ components

Principal Component Analysis (PCA)
This B "SVD $+Q 0$ in stats language".
$\rightarrow$ it's often how SVD (or "(near algebra") is used in statistics 4 data analysis.
$\rightarrow$ it makes precise statements about fitting data to ines/planes/etc and how good the fit is
Idea: If you have $n$ samples of $m$ values each $\leadsto$ 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): $\mu=\frac{1}{n}\left(x_{1}+\cdots+x_{n}\right)$
Variance: $s^{2}=\frac{1}{n-1}\left[\left(x_{1}-\mu\right)^{2}+\cdots+\left(x_{n}-\mu\right)^{2}\right]$
Standard Deviation: $s=\sqrt{\text { variance }}$
This tells you how "spaced out" the samples are: $\because 68 \%$ of samples are within $\pm s$ of the mean.2
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 $\binom{x_{1}}{y_{1}}, \ldots,\binom{x_{n}}{y_{n}}$
$x_{i}=$ score on problem 1 $y_{i}=$ score on problem 2

Mean scores:
Problem 1: $\mu_{1}=\frac{1}{n}\left(x_{1}+\cdots+x_{n}\right)$
Problem 2: $\mu_{2}=\frac{1}{n}\left(y_{1}+\cdots+y_{n}\right)$
Recenter to compute variance:

$$
\bar{x}_{i}=x_{i}-\mu_{1} \quad \bar{y}_{i}=y_{i}-\mu_{2} \quad \text { (subtract means) }
$$

Variance
Problem 1: $\quad s_{1}^{2}=\frac{1}{n-1}\left(\bar{x}_{1}^{2}+\cdots+\bar{x}_{n}^{2}\right)$
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 $1\left(x_{i}\right)$ and Problem $2\left(y_{i}\right)$ individually - so far we've ignored the fact they might be rebated. This is what PCA does.

Eg cares $\binom{x_{i}}{g_{i}}=\binom{8}{15},\left(\begin{array}{l}1 \\ 2 \\ 2\end{array},\binom{12}{16},\binom{6}{7},\binom{1}{7},\binom{2}{1} \quad \begin{array}{l}\mu=5 \\ \mu_{2}=8\end{array}\right.$ recenter: $\binom{\overline{x_{i}}}{y_{i}}=\binom{x_{i}}{y_{i}}-\binom{1}{8}=\binom{3}{7},\binom{-4}{-6},\binom{7}{8},\binom{1}{1},\binom{-4}{-1},\binom{-3}{-7}$



Store in matrices:

$$
\begin{aligned}
& A_{0}=\left(\begin{array}{lll}
n & \text { dances } & \cdots \\
y_{1} & \cdots & y_{n}
\end{array}\right)=\left(\begin{array}{cccccc}
8 & 1 & 12 & 6 & 1 & 2 \\
15 & 2 & 16 & 7 & 7 & 1 \\
3 & -4 & 7 & 1 & -4 & -3 \\
7 & -6 & 8 & -1 & -1 & -7
\end{array}\right) \\
& A=\left(\begin{array}{llll}
x_{1} & \cdots & x_{n} \\
y_{1} & & y_{n}
\end{array}\right)=\left(\begin{array}{ccc} 
&
\end{array}\right)
\end{aligned}
$$

$N B:$ Recentered means $\bar{x}_{1}+\cdots+\bar{x}_{n}=0=\bar{y}_{1}+\cdots+\bar{y}_{n}$ :
The sum of the columns of the reentered data matrix $A$ zero.

Covariance Matrix:

$$
\begin{aligned}
S & =\frac{1}{n-1} A A^{\top}=\frac{1}{n-1}\left(\begin{array}{cc}
(\text { row } 1) \cdot(\text { row } 1) & (r o w 1) \cdot(\text { row } 2) \\
(\text { row 2) }) \cdot(\text { row } 1) & (n, 02) \cdot(\text { row } 2)
\end{array}\right) \\
& =\frac{1}{n-1}\left(\begin{array}{cc}
\bar{x}_{1}^{2}+\cdots+\bar{x}_{n}^{2} & \bar{x}_{1} y_{1}+\cdots+\bar{x}_{n} y_{n} \\
\bar{x}_{1} y_{1}+\cdots+\bar{x}_{n} y_{n} & \bar{y}_{1}^{2}+\cdots+\bar{y}_{n}^{2}
\end{array}\right)
\end{aligned}
$$

The diagonal entries are the variances:

$$
s_{1}^{2}=\frac{1}{n-1}\left(\bar{x}_{1}^{2}+\cdots+\bar{x}_{n}^{2}\right) \quad s_{2}^{2}=\frac{1}{n-1}\left(\bar{y}_{1}^{2}+\cdots+\bar{y}_{n}^{2}\right)
$$

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

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

- If this is positive then $\bar{x}_{i} \& \bar{y}_{i}$ generally have the same sign: if you did above average on $P 1$ then you likely did above average on $P 2$ too, \& vice-versa. The values are comelated.
- 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:

$$
S=\frac{1}{5} A A^{\top}=\left(\begin{array}{ll}
20 & 25 \\
25 & 40
\end{array}\right) \quad \begin{aligned}
& S_{1}^{2}=20 \\
& S_{2}^{2}=40
\end{aligned}
$$

$(1,2)$-covariance $=25>0$. people who did above average on $P 1$ likely did above average on $P 2$.
The SVD will tell as which directions have the largest \& smallest variance.
(column means $=0$ )
Def: Let $A$ be a recentered data matrix
$A=\left(\begin{array}{ll}\bar{d}_{1} & \cdots \\ d_{n}\end{array}\right)$ where $J_{i}=\binom{\bar{x}_{i 1}}{\bar{x}_{i m}}=i-\frac{1-}{}$ recentered data point
Let $S=\frac{1}{n-1} A A^{\top}$ be the covariance matrix.
Let $u \in \mathbb{R}^{m}$ be a unit vector.
The variance in the $u$-direction is

$$
s(u)^{2}=u^{\top} S u
$$

$N B: s(u)^{2}=u^{\top}\left(\frac{1}{n-1} A A^{\top}\right) u=\frac{1}{n-1}\left(u^{\top} A\right)\left(A^{\top} u\right)=\frac{1}{n-1}\left(A^{\top} u\right)^{\top}\left(A^{\top} u\right)$ $=\frac{1}{n-1}\left(A^{\top} u\right) \cdot\left(A^{\top} u\right)=\frac{1}{n-1}\left\|A^{\top} u\right\|^{2}$.
Since $A^{\top} u=\left(\begin{array}{c}-d_{I}^{\top}- \\ \vdots \\ -d_{n}-\end{array}\right) u=\left(\begin{array}{c}\bar{J}_{1} \cdot u \\ \vdots \\ \bar{d}_{n} \cdot u\end{array}\right)$ we get

$$
s(u)^{2}=u^{T} S_{u}=\frac{1}{n-1}\left(\left(d_{i} u\right)^{2}+\cdots+\left(d_{n}-u\right)^{2}\right)
$$

$N B: \bar{d}_{1}+\cdots+\bar{d}_{n}=0$ for a recentered data matrix $A(p .8)$. Hence $0=0 \cdot u=\left(\overline{d_{1}}+\cdots+\bar{d}_{n}\right) \cdot u=\left(\bar{d}_{1} \cdot u\right)+\cdots+\left(\overline{d_{n}} \cdot u\right)$ so it makes sense to compute the variance of these number $\left(\overline{d_{i}} \cdot u\right), \ldots,\left(\bar{d}_{r} \cdot u\right)$ with mean 2 zero:

$$
s(u)^{2}=\frac{1}{n-1}\left(\left(d_{i} u\right)^{2}+\cdots+\left(d_{n}-u\right)^{2}\right)
$$

Eg: If $u=\binom{1}{0}=e_{1}$ then $\bar{d}_{i} \cdot u=\binom{\bar{x}_{i}}{y_{i}} \cdot\binom{1}{0}=\bar{x}_{i}$, so

$$
s(u)^{2}=s(e)^{2}=\frac{1}{n-1}\left(\bar{x}_{1}^{2}+\cdots+\bar{x}_{n}^{2}\right)=s_{1}^{2}
$$

This is just the variance of the $x_{i}$ is.

$$
\text { In general, } s\left(e_{i}\right)^{2}=s_{i}^{2}
$$

Picture: Recall that if $u$ is a unit vector then $\langle v \cdot u\rangle_{u}=$ projection of $v$ onto Span \{u\} ~

$$
\left.\Rightarrow(v \cdot u)^{2}=(v \cdot u)^{2} \| u\right)^{2}=U(v-u) u \|^{2}=\text { length of the }
$$ projection of $r$ onto $S p a n\{u\}$



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


- $=\bar{d}_{i}=\binom{\bar{x}_{i}}{\bar{y}_{i}}$
- $=\left(d_{i} \cdot u\right) u$
$s(u)^{2}=$ sum of squares of distances from the do 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 $\lambda_{1}$-eigenvector.
Quadratic Optimization:
$u$ maximises $s(u)^{2}=u^{\top} S u$ subject to $\|u\|=1$ with maximum value $\sigma_{1}{ }^{2}$
Therefore:
$u_{1}$ is the direction of greatest variance
$\sigma_{1}^{2}=s\left(u_{1}\right)^{2}=$ variance in the $u_{1}$-direction
Our data points are "stretched out" most in the $u$-direction.

In our example:

$$
\begin{array}{ll}
\frac{1}{\sqrt{6-1}} A=\sigma_{1} u_{1} v_{1}^{\top}+\sigma_{2} u_{2} v_{2}^{\top} \quad \text { for } \\
\sigma_{1}^{2} \approx 56.9 & \sigma_{2}^{2} \approx 3.07 \\
u_{1} \approx\binom{0.561}{0.828} & u_{2} \approx\binom{0.828}{-0.561}
\end{array}
$$



- $=\bar{d}_{i} \quad=$ projection of - onto $S_{\text {pan }}\left\{u_{i}\right\}$

So the first principal component is $u_{i}$, and the variance in that direction is $\approx: 56.9$.
(NB this is greater than the Problem 1 variance $=20$ \& the Problem 2 variance $=40$ )

