Today: • Principal Component Analysis (PCA)
• Singular Value Decomposition (SVD)

Problem: Sort through a set of a lot of complicated data with potentially irrelevant data and noisy data, to find out what the most important parts of the data are.

Ex: Netflix movie recommendations ~

GOAL: they want to recommend to you a movie you will watch and like

To do this, they have a ton of data:
- previous watch history
- categories / genres
- writers / directors
- when it was written / general popularity
- specific actors
- user ratings
- demographic info
... and lots more info!

Netflix needs to know what things are most important to consider...

This kind of problem happens so frequently that there are lots of algorithms for deciding the most important information. (in data science: feature selection)

One popular method: principal component analysis (PCA)

Steps in PCA:

Start with an $n \times p$ matrix of data X, which we think of as "experimental data".

Rows of X = different experiments

Cols of X = different measurements

(Step 1) Create a matrix Y whose columns are the same as the columns of X, minus their mean values

MATLAB: `Y = X - mean(X,1);`

(Step 2) Create the covariance matrix $C = Y^T Y$ ($p \times n$ times $n \times p$ $\leftarrow p \times p$)

MATLAB: `C = transpose(Y)*Y;`

(Step 3) Calculate the eigenvectors/eigenvalues of C

MATLAB: `[P,D] = eig(C);`

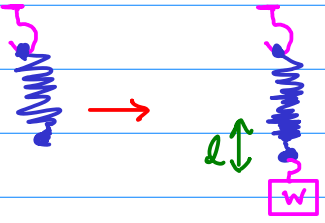
`P`: columns are eigenvectors

`D`: diagonal matrix whose entries are eigenvalues

(Step 4) Grab the eigenvector whose eigenvalue is largest. When the eigenvalue is much larger than the rest, the associated eigenvector corresponds to a more important measurement (linear combination of the base measurements)

Mass-Spring System Example:

Imagine an experiment:



Attach different weights w to a spring and seeing how long l it stretches

Note: in my experiment, I will have measurement errors

Results of experiment:

	weight attached (lbs)	stretch length (ft)
experiment 1	1.0	0.009
experiment 2	1.2	0.013
experiment 3	1.4	0.014
experiment 4	1.6	0.017
experiment 5	1.8	0.018
experiment 6	2.0	0.019

Using PCA, without knowing any physics, can determine the most important components in the mass-spring system.

$$X = \begin{bmatrix} 1.0 & 0.009 \\ 1.2 & 0.013 \\ 1.4 & 0.014 \\ 1.6 & 0.017 \\ 1.8 & 0.018 \\ 2.0 & 0.019 \end{bmatrix}$$

$$C = \begin{bmatrix} 0.7 & 0.0068 \\ 0.0068 & 0.00007 \end{bmatrix}$$

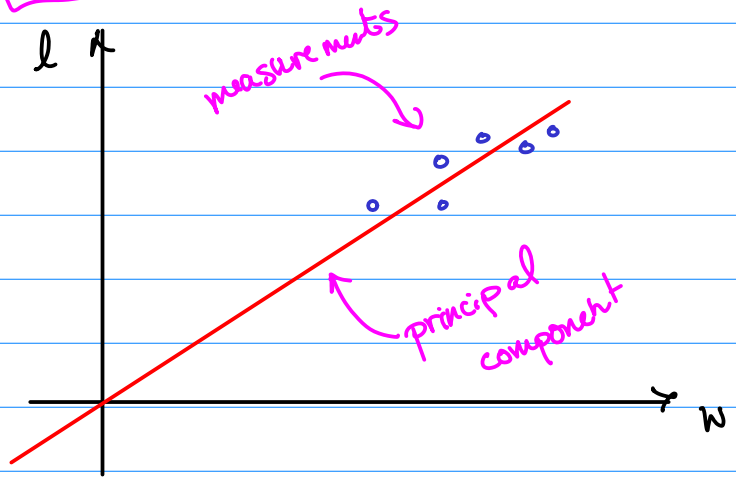
Eigendata: eigenvalue 0.0000039425 w/ eigvector $\begin{bmatrix} 0.009717 \\ -0.9995 \end{bmatrix}$

eigenvalue 0.700068 w/ eigenvector $\begin{bmatrix} -0.9999528 \\ -0.0097139 \end{bmatrix}$

The measurement given by

$$\begin{bmatrix} -0.9999528 \\ -0.0097139 \end{bmatrix} \cdot \begin{bmatrix} w \\ l \end{bmatrix}$$

$$= \boxed{-0.9999528\,w - 0.0097139\,l}$$



eigenvectors w/ eigenvalue 0.7000666 are $\begin{bmatrix} -0.9999528c \\ -0.0097139c \end{bmatrix}$ for $c \neq 0$.

Housing Prices:

Some data on some houses sold in a certain area

Cost C	Interior Sq. Footage x	Exterior Sq. Footage z	# Bathrooms m	# Bedrooms n
234641	1000	3000	2	1
449749	2125	0	1.5	2
648350	3025	1000	2	3
352657	1500	2100	3	2
592788	2800	500	2	2
427948	1900	500	2	3

Let's do PCA! Make a data matrix X (6×5)

From MATLAB:

eigenvalues of the covariance matrix:

$$0.00000000 \cdot 10^{11}$$
$$0.00008000 \cdot 10^{11}$$
$$0.00000012 \cdot 10^{11}$$
$$0.00003418 \cdot 10^{11}$$
$$1.16072403 \cdot 10^{11}$$

Eigenvector:

$$\begin{bmatrix} 0.999974\,a \\ 0.00502618\,a \\ -0.0051617\,a \\ -0.0000008\,a \\ 0.00000337\,a \end{bmatrix} \qquad a \neq 0$$

Interpretation:

$$\begin{bmatrix} c \\ x \\ z \\ m \\ n \end{bmatrix} \approx a \begin{bmatrix} 0.999974 \\ 0.00502618 \\ -0.0051612 \\ -0.0000008 \\ 0.00000337 \end{bmatrix}$$

$$a = 0.999974c + 0.00502618x - 0.0051612z - 0.0000008m + 0.00000337n$$