covariance, correlation

If you'd like to, go play around with:

https://tylervigen.com/spurious-correlations

What are two things that have a positive correlation in your experience?

↳ amount of time on HW + amount of $$ on take-out

↳ time studying + time at the gym

↳ negative correlation

Covariance

- A **covariance** measurement tells us about how two random variables vary *together*.

- variance told us about 1 r.v.

- no claims about correlation or causation

Covariance

- A **covariance** measurement is calculated with the formula
- $cov(X, Y) = E[(X - E[X])(Y - E[Y])]$

expected val of X

r.v.s

vals of X

- For a specific sample of data points, this becomes:
- $\hat{\sigma}^2_{x,y} = \frac{1}{N-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

individual data points

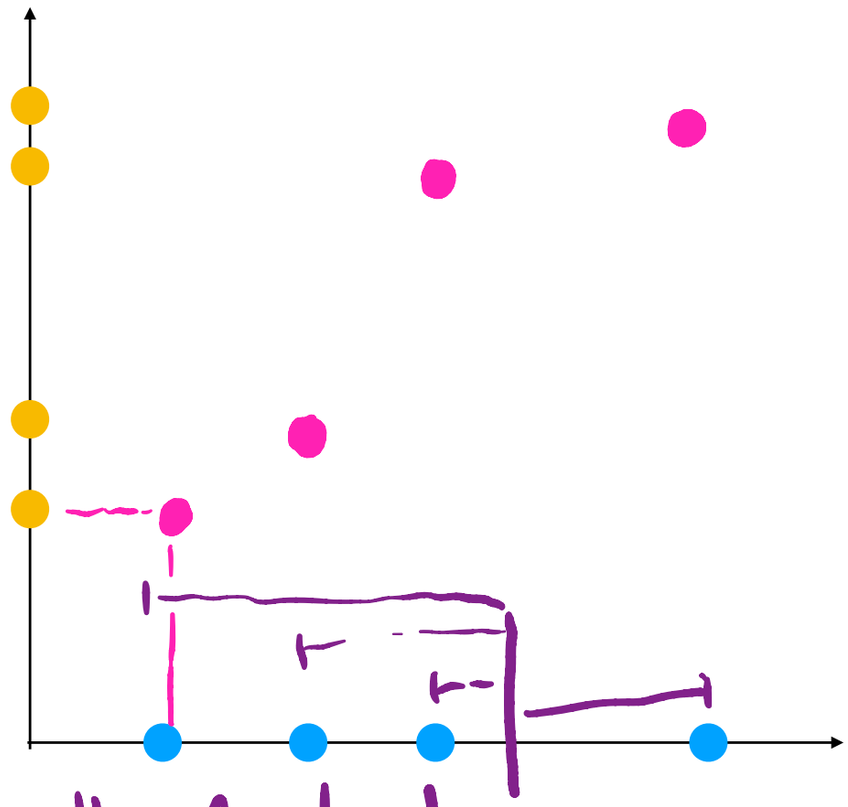
actually use this calculation

N: number of paired data points

Covariance

$\bar{x} = \frac{(2+3+4+6)}{4} = 3.75$

$\bar{y} = 3.25$



	1	2	3	4
# of bedrooms	2	3	4	6
rent	2	2.5	4	4.5

$$\frac{1}{N-1}\sum_i (x_i-\bar{x})(y_i-\bar{y})$$

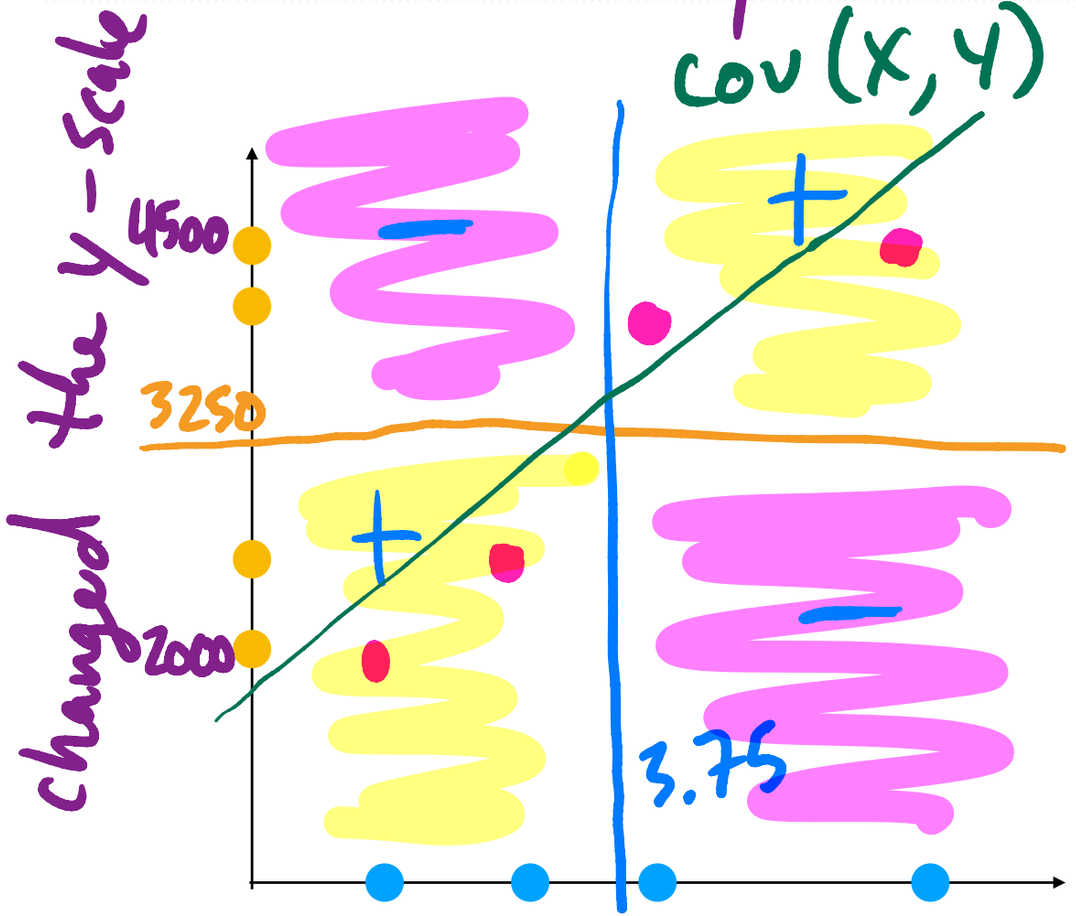
$$\frac{1}{4-1}\left((2-3.75)(2-3.25)+(3-3.75)(2.5-3.25)+\dots\right) = 1.91$$

Covariance

$\bar{x} = 3.75$

$\bar{y}$ = was 3.25, now 3250

cov(X,Y) was 1.91, now 1910



	1	2	3	4
# of bedrooms	2	3	4	6
rent	2000	2500	4000	4500

$\frac{1}{3}\left((2-3.75)(2000-3250) + \dots\right)$

by 1000 times what we had before

ICA Question 1: Calculating Covariance

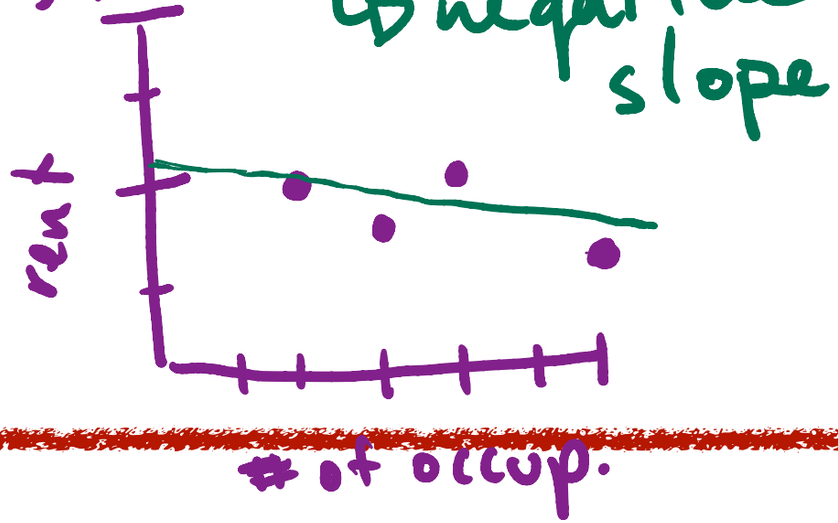
Calculate the covariance for the given data points.

$$\hat{\sigma}^2_{x,y} = \frac{1}{N-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$\bar{x} = 3.75$

$\bar{y} = 0.895\ldots$

$cov(x,y) = -0.145$ ← negative slope



	1	2	3	4
# of occupants	2	3	4	6
rent per person	1	0.83	1	0.75

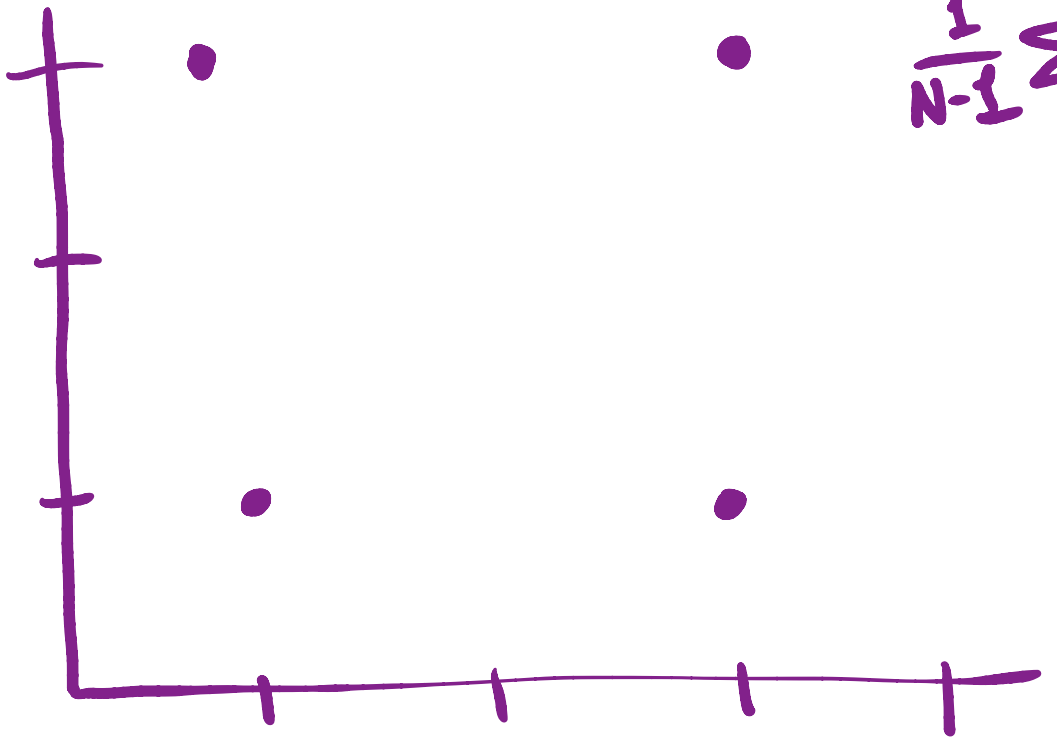
ICA Question 2: Calculating Covariance

Give an example data set for which the covariance is 0.

$$\hat{\sigma}^2_{x,y} = \frac{1}{N-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- the slope of the line is 0
  - one of our vars. has 0 for its variance
- if for every point there is an "opposite" point

	1	2	3	4
X	1	7	-3	2
Y	4	4	4	4



$$\frac{1}{N-1}\sum (x_i-\bar{x})(y_i-\bar{y})$$

$$\downarrow$$

$$1+-1+1+-1$$

Covariance

- Positive covariance = relationship w/ pos. slope
- Negative covariance = neg. slope
- zero covariance = no relationship
- Covariance is sensitive to the scale of the underlying data
- The magnitude of the covariance tells us nothing about the slope of the line and nothing about the degree of fit to the line

Covariance

- Some properties of covariance:
  - The covariance of a variable with itself is equal to its variance
  - $cov(X, X) = Var(X) = \sigma_X^2$
- Random variables whose covariance is zero are uncorrelated, but not necessarily independent
- Random variables that are independent have a covariance of zero

if you get above 100 on a test, I'll give you a gold star

Covariance Matrices

- Given two (or more) variables, we can define a matrix to contain information about the linear relationships between these variables

- The diagonal in a covariance matrix is the variance of the variables

$$\Sigma =$$

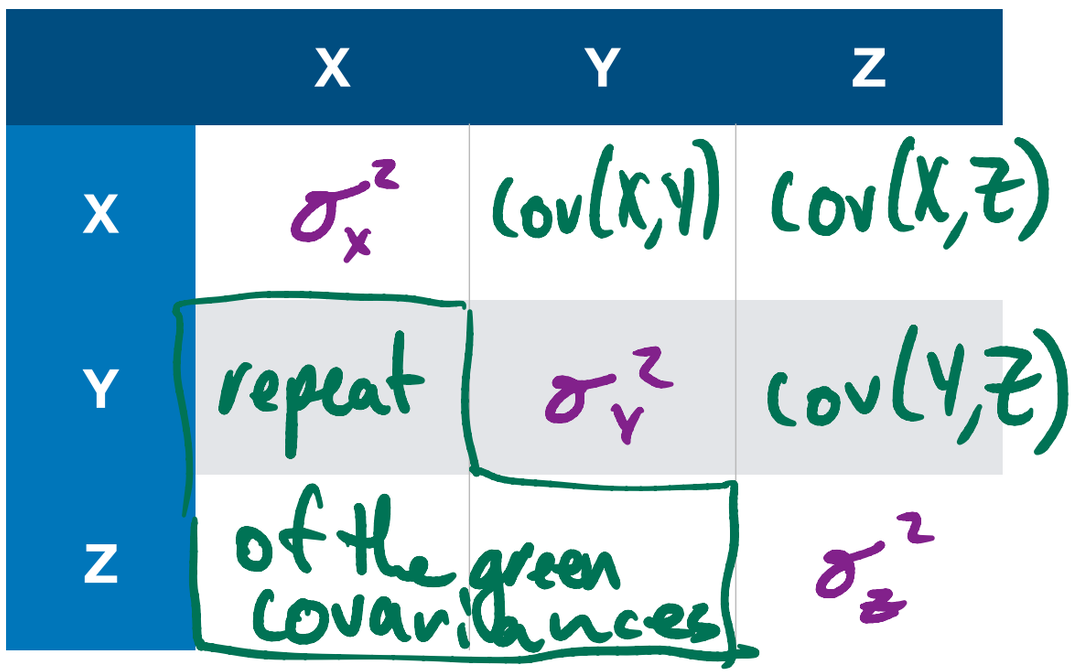
	X	Y
X	cov(X,X)	cov(X,Y)
Y	cov(Y,X)	cov(Y,Y)

cov(X,Y) = cov(Y,X)

Covariance Matrices

- Same deal when we have more than two variables

$$\Sigma =$$



ICA Question 3: Covariance Matrices

What is cov(A, D)?

x

Which pairs of variables might be independent?

A, B

C, D

Which random variable has the smallest expected value?

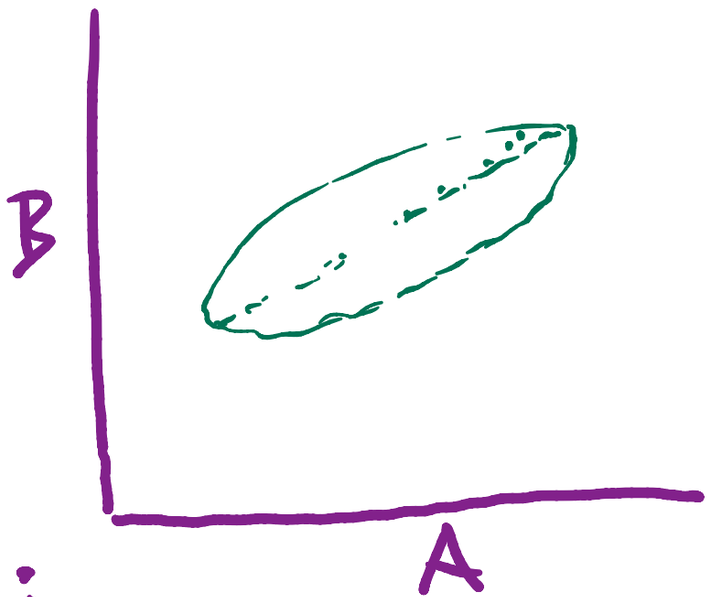
B has the smallest variance

idk?

	A	B	C	D
A	10	0	w	x
B	0	2	y	z
C	w	y	3	0
D	x	z	0	5

ICA Question 4: Covariance Matrices

What might the scatter plot for the given covariance matrix look like?

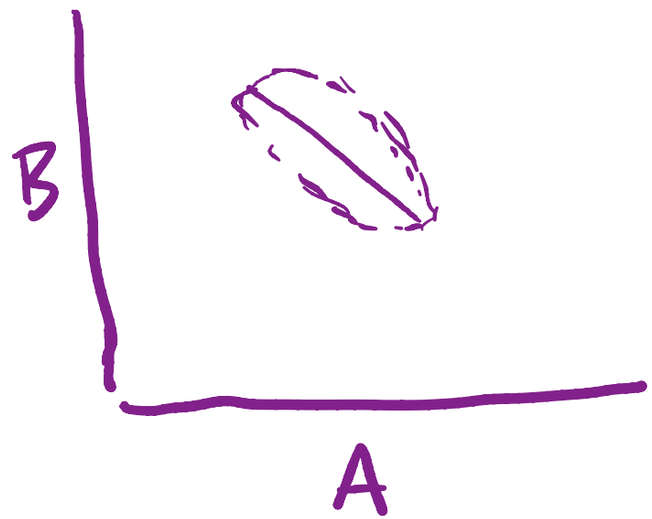


reqs:
- positive relationship (slope)
- A spans more vals than B

	A	B
A	5	1
B	1	2

ICA Question 5: Covariance Matrices

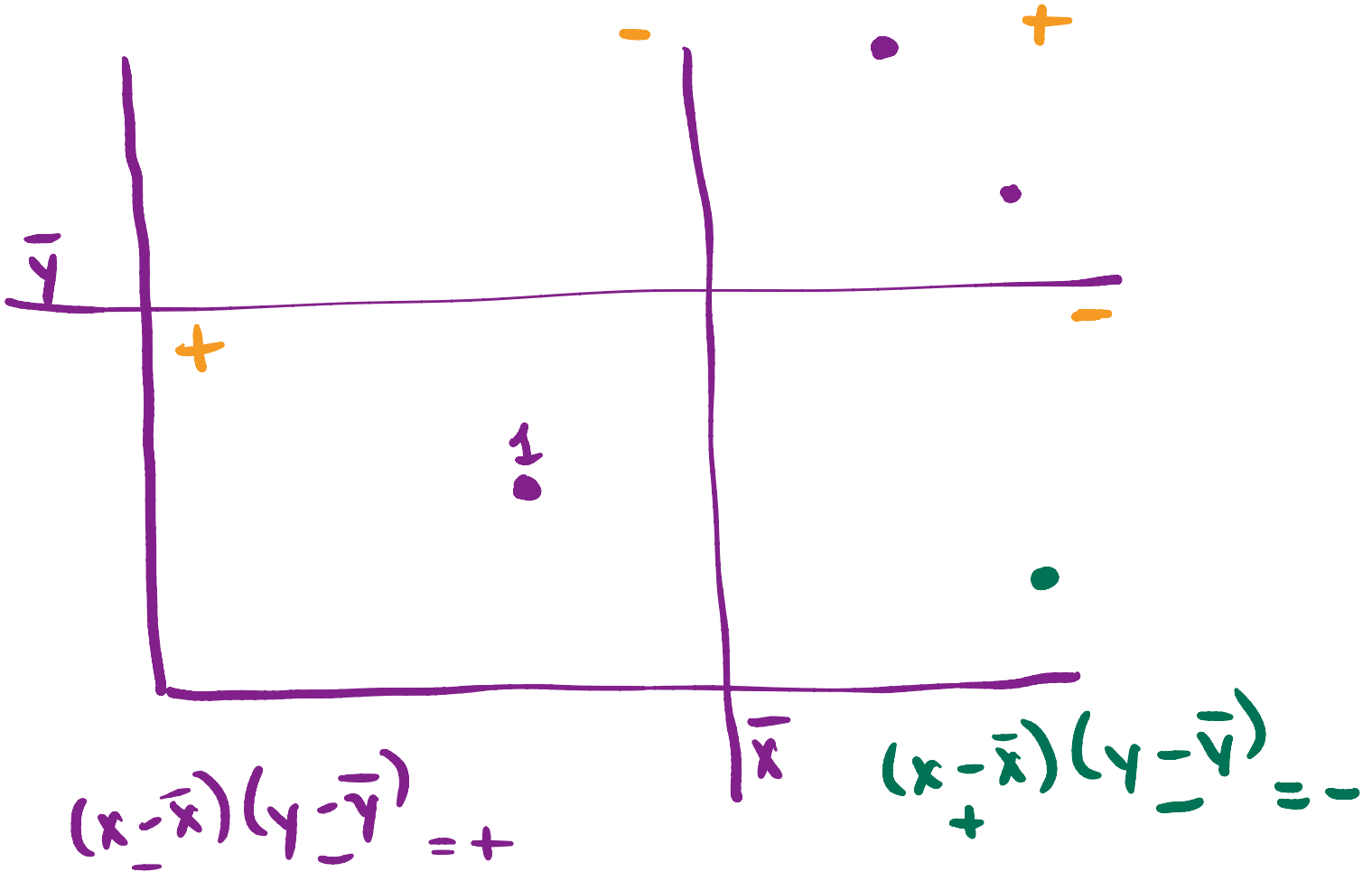
What might the scatter plot for the given covariance matrix look like?

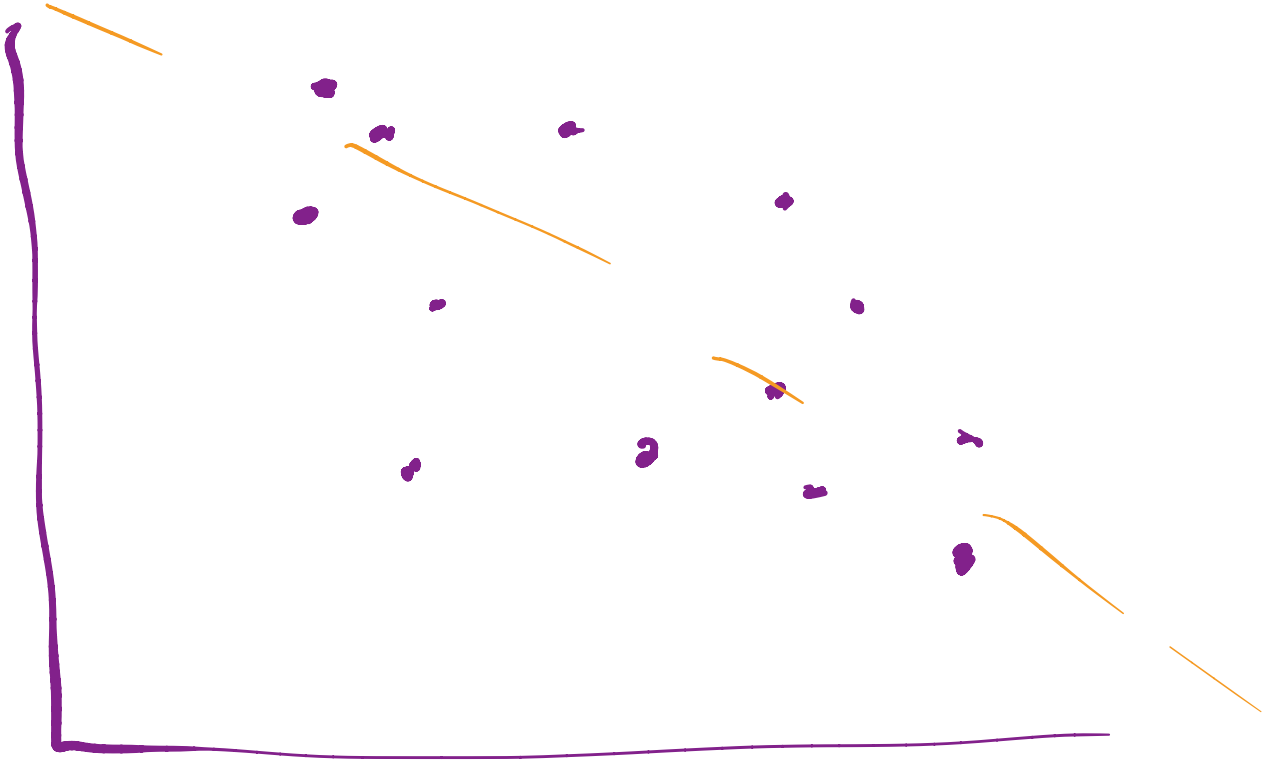


pos cov = pos slope

neg cov = neg slope

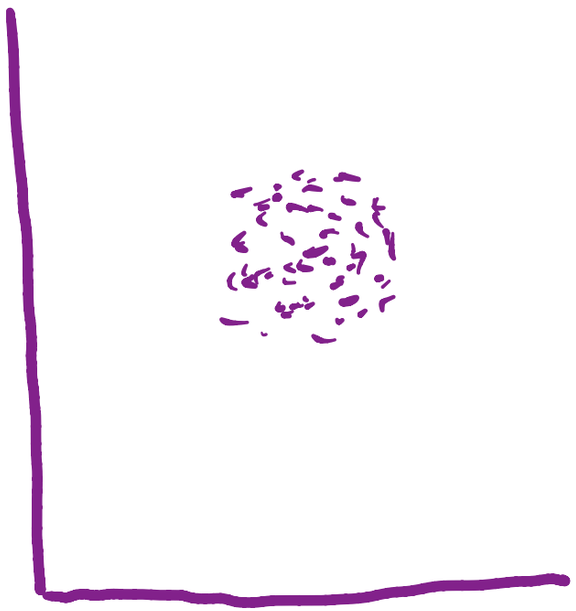
	A	B
A	2	-4
B	-4	2


$\bar{y}$
$\bar{x}$
1
$(x-\bar{x})(y-\bar{y}) = +$
$(x-\bar{x})(y-\bar{y}) = -$



ICA Question 6: Covariance Matrices

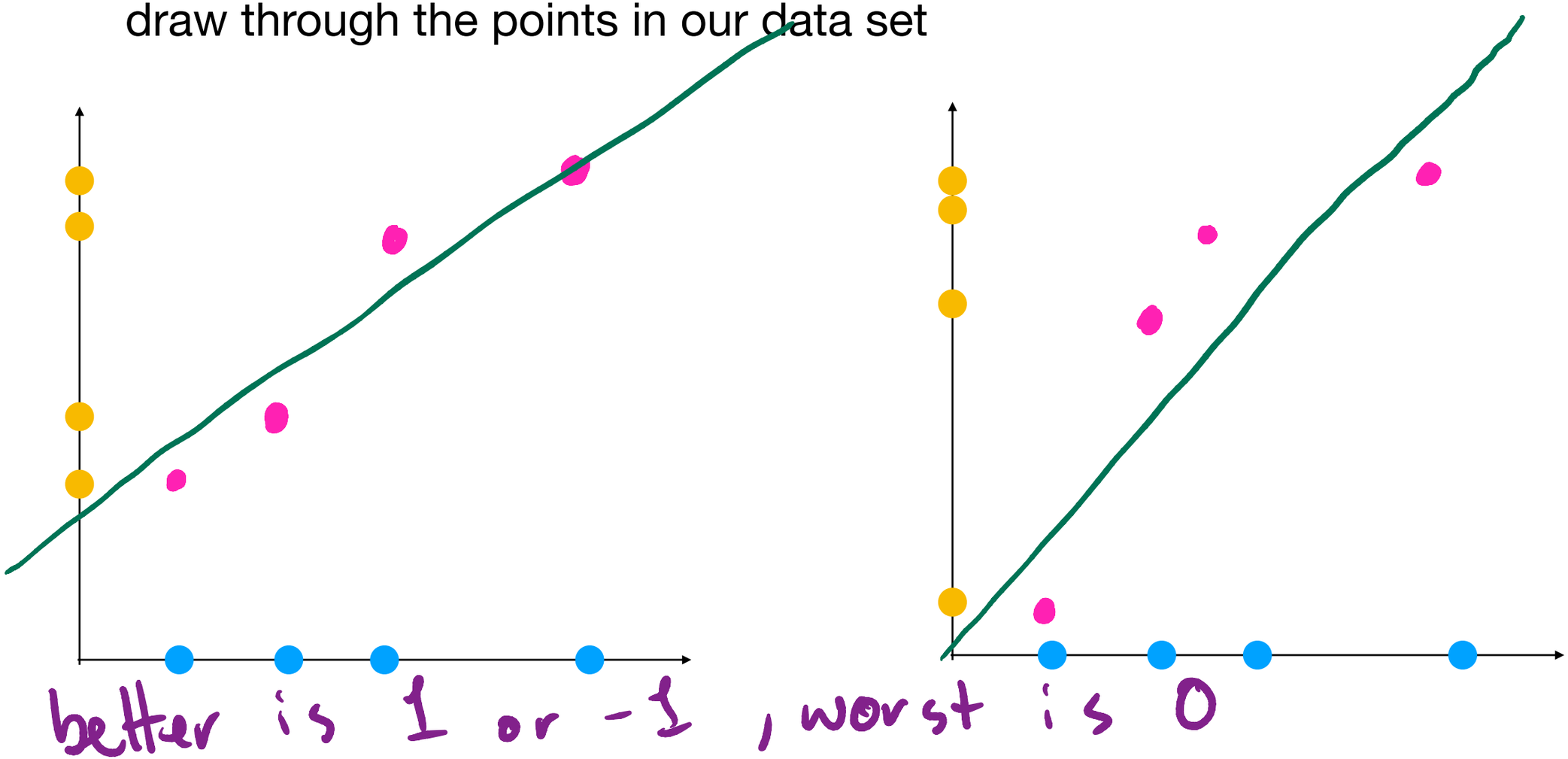
What might the given covariance matrix be for the given data?



	A	B
A	X	0
B	0	X

Correlation - Pearson's Correlation

- Correlation measures the "goodness of fit" about the line that we can draw through the points in our data set



Correlation

- Correlation of positive one means that we can draw a straight line with a positive slope through all of the points
- it tells us nothing about the steepness of the line
- Any two points always have a correlation of: 1 or -1
- The more data that we have, the more confidence we can have in our predictions, but the correlation number doesn't explicitly tell you how much data you have

ONE

negative one = neg slope through all points

Correlation

- Covariance is sensitive to the scale of the data
- Correlation is not sensitive to the scale of the data

↳ mapping cov into [-1, 1]
↳ (also indicates goodness of fit)

Correlation

- Correlation always produces a number in the range of: [-1, 1]
- If a straight line cannot go through all of the data points, the correlation gets closer to zero

Correlation

- Calculating Pearson's Correlation Coefficient:
- $$\frac{cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$
- This will produce a correlation of 0 if:

ICA Question 7: correlation

$$\text{correlation} = \frac{cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

What is the correlation coefficient for A and B?

$$A,B = \frac{0.25}{\sqrt{0.25}\sqrt{1}} = 0.5$$

Which line would we "trust the most" for making a prediction of one variable based on the other?

$A,B = 0.5$

$A,C = -0.62$

$B,C = -0.99$

	A	B	C
A	0.25	0.25	-1.25
B		1	-4
C			16.3

Correlation

- Correlation is still tricky to interpret!
  - A line with a correlation of 0.9 might be twice as good to make predictions with as a line with the correlation 0.64, for instance
    - (We'll talk about this more when we talk about R^2)
- Want to know how much to trust your correlation?
  - Calculate a p-value!
  - We actually do this by getting the t-score and then calculating the p-value the same way we did before—looking up where we are in the t-distribution

↳ yes, but how much data?

p takes into account "how much data"

Admin

- Test 4: if you have a conflict because of Eid celebrations, send Felix an email **now** so that we can get you set up with an alternate time

Schedule

HW 8 is released

Turn in **ICA 20** on Canvas (make sure that this is submitted by 2pm!) - passcode is "hi"

Test 4: May 4th from 1 - 3pm in Snell Engineering 108

Mon	Tue	Wed	Thu	Fri	Sat	Sun
April 4th Lecture 19 - chi-square test, multiple comparison correction	**Felix OH Calendly**	**Felix OH Calendly**	**Felix OH Calendly** Lecture 20 - covariance, correlation			
April 11th Lecture 21 - conditional probabilities, bayes	**Felix OH Calendly**	**Felix OH Calendly**	**Felix OH Calendly** Lecture 22 - conditional ind., bayes nets			**HW 8 due @ 11:59pm**
April 18th **No lecture - Patriot's Day**	**Felix OH Calendly**	**Felix OH Calendly**	**Felix OH Calendly** Lecture 23 - Regression: R^2 & F			
April 25th Lecture 24 - presentations, wrap-up **Mini-project due @ 11:45am**		**HW 9 due @ 11:59pm**				

More recommended resources on these topics

- Some slightly aggressive youtube videos (there's a lot of "bam!" sound effects?)
  - StatQuest: Covariance, Clearly Explained!!!
  - StatQuest: Pearson's Correlation, Clearly Explained!!!
- YouTube: Brandon Foltz, Statistics 101: The Covariance Matrix
- Website: Statology: How to find the p-value for a correlation coefficient in Excel