Principle Component Analysis (PCA)

- When? Data consists of large sets of correlated variables

- Why? Allows you to summarize complex data with a set of smaller (in number) representative variables. These vars are supposed to explain most of the variability in your original dataset

- What? A principal component are a direction in "feature space" along which the data are highly variable. Thus, a Principal Component is a linear combination of our variables that has the highest variance.
  - the following Principal Components are the same, but have the constraint of not being correlated w the first (orthogonal)

- How? Eigenvectors are computed from the covariance matrix.

Possible Framings

- Has this ever happened to you? (huge covariance matrix)

- Why is linear regression interesting? Why do we care about variance in the first place?

Intuitive Explanations

- 2D data w a line maximizing the variance

- demonstrating how getting the linear combination to maximize variance is akin to clustering

- also demonstrate how PCA could be used to look for latent variables in your data.

Explanations for me

- The Covariance Matrix iherently contains some information about the relationship b/t all vars

- If 2 variables covary more, they will have a higher number in that location in the covariance matrix

$$\begin{bmatrix} 1 & .7 & .2 \\ .7 & 1 & 0 \\ .2 & 0 & 1 \end{bmatrix}$$

- Now think about how matrix multiplication works

$$\begin{bmatrix} 1 & .7 & .2 \\ .7 & 1 & 0 \\ .2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 + 0.7 + 0.2 \\ .7 + 1 + 0 \\ .2 + 0 + 1 \end{bmatrix} = \begin{bmatrix} 1.9 \\ 1.7 \\ 1.2 \end{bmatrix}$$

- Multiplying by the covariance matrix "moves" a vector in a direction of the most covariation. That is, if two variables covary a lot, when the matrix multiplication is carried out, the matrix will amplify data in the direction that variables covary the most.

- Since eigen decomposition searches for the direction unchanged by a transformation, and this transformation moves vectors in the direction of most covariance, the eigenvector is the exact direction in which there is the most covariance b/t all variables.

- The intuition for eigenvalues is that they are proportional to the amount the data is pushed in the direction of greatest variance, and are thus proportional to the amount of variance explained by the component pointing in that direction.

How is grouping related to finding the PC?

The gist is that once viewing the data along axis of the greatest variation, we can better see actual structure in our data, like groups that cluster along the axis of most variance.

Summary of Presentation

1.
What I'm Going to Tell You

- when PCA + reduce factor # in future analyses
- why PCA
- what PCA
- how PCA

How to find dir of most var

- correlation matrix part I
- matrix multiplication
- matrices do things
- if matrices do things, there are some objects matrices don't do things to
- eigen vectors and values
- correlation matrix part II

What to do once you find it / so what?

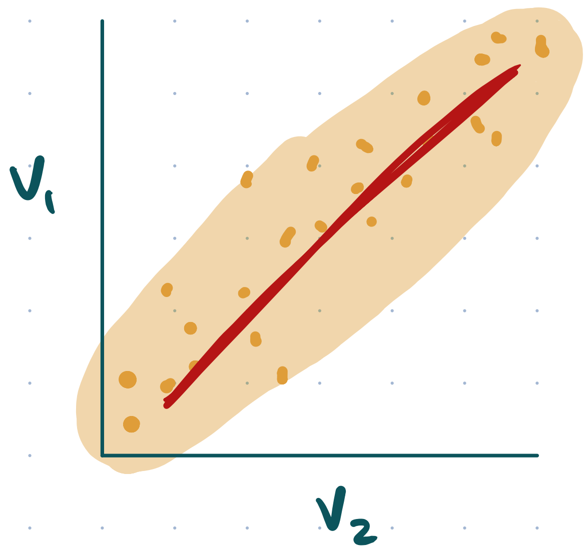
- scree plots + visualize factor loadings
- simplifies data for working intuitively
- e.g. cluster identification
- latent factor identification
- e.g. personality big 5
- exploration vs confirmation

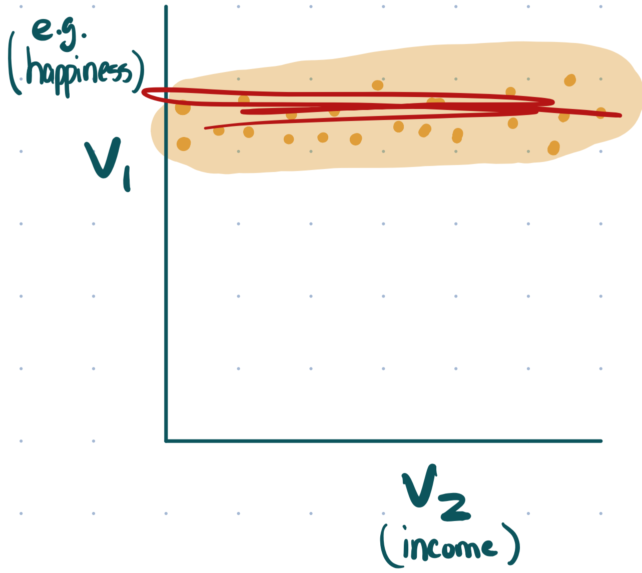
Principle Component Analysis

Branson Byers
PSY 540
4/10/2022

Why care about variance, spiritually?

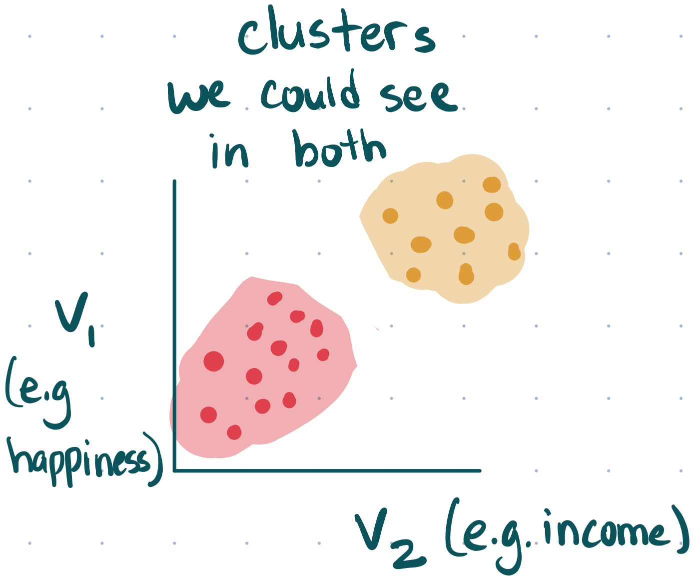
- Which of the following two plots are more interesting?

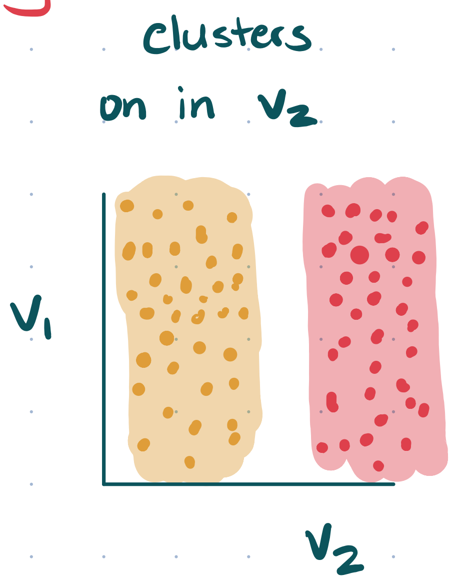




- When we are looking for a relationship b/t 2 variables, this requires variance in those 2 variables.

What can variance actually reveal?

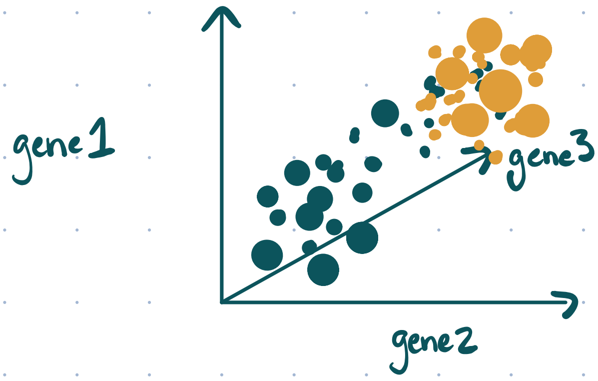




[ultimately, visualizing the variance that matters most in our data helps us visualize the relationships and clustering in our data]

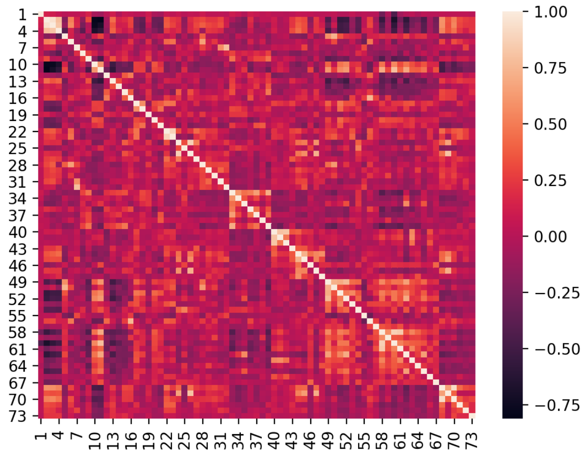
Example: Too Many Genes

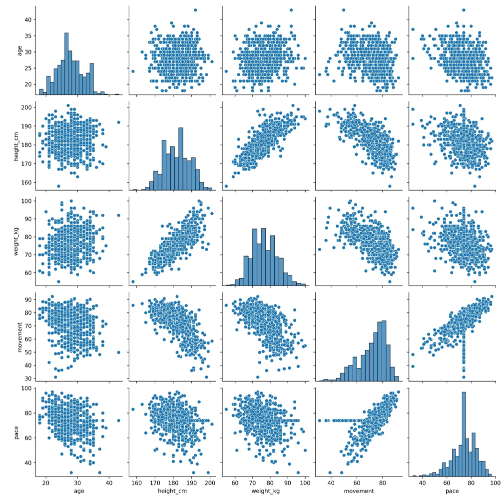
- 3 genes?



- 4 genes???

Has this ever happened to you?





"Just one more measure"

"Why not add it? We have room."

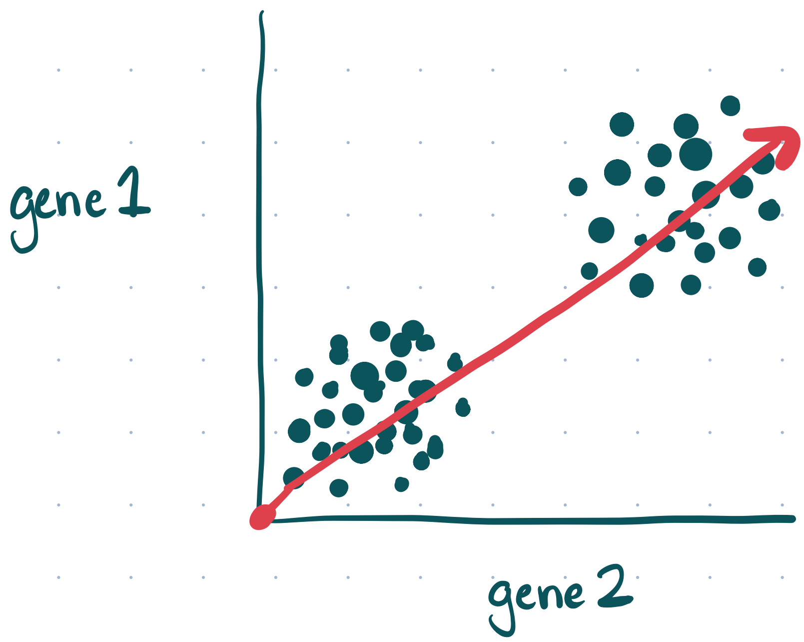
- When we have lots of data, specifically, many measures/features/dimensions, it can be hard to tell what to care about.

Finding the Direction of Max Variance is Helpful for Visualization

1. Looking at features with variance is most informative for seeing relationships in data.

2. If we can ignore features/directions with less variance, we can visualize the data more easily :)

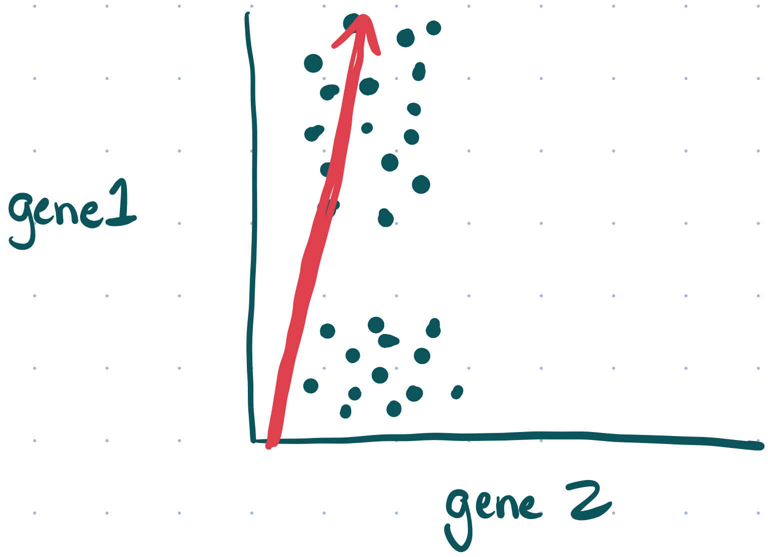
Simpler Example



- direction of most variance
- think pythagoras here

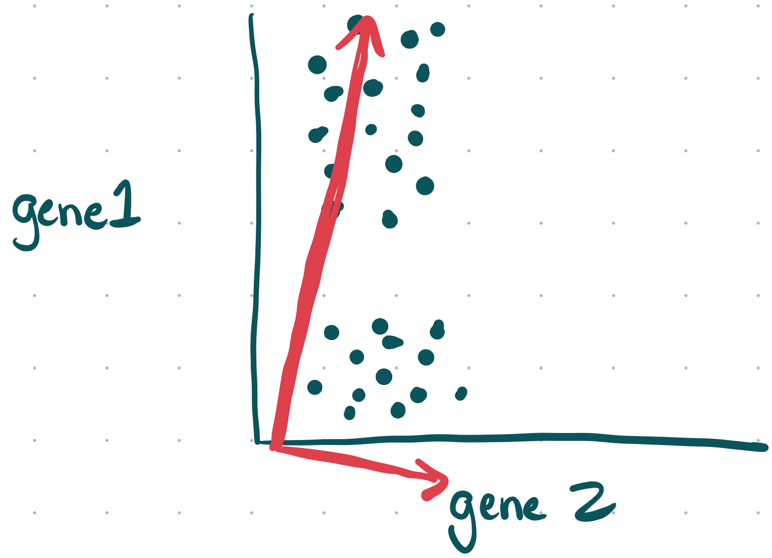
- in terms of gene 1 and 2, they both contribute lots of variance

· we can imagine a situation where 1 does and 2 doesn't



[explain factor loading]

[It's how much a factor's variance contributes to the direction of most variance]

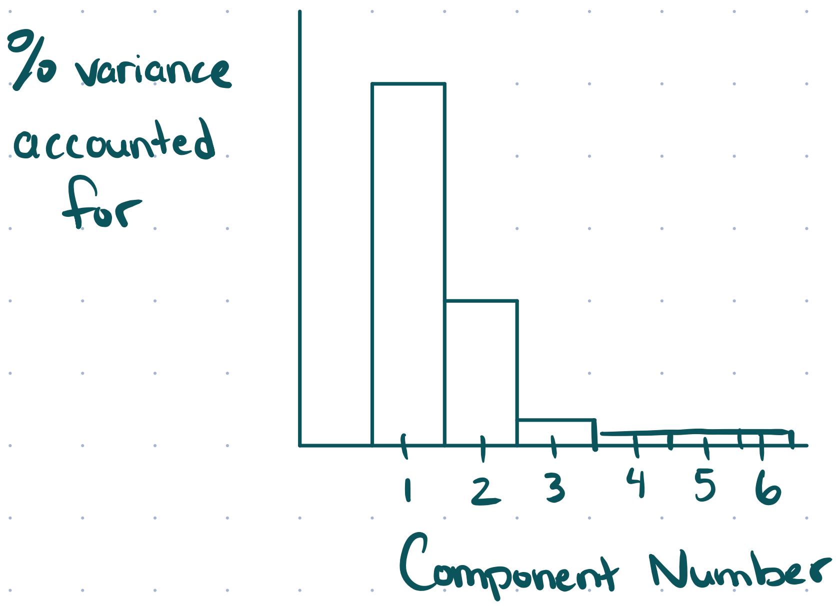


you can have a direction of second most variance, but it should be 90° to the first one so that they do not double count any variance (are uncorrelated)

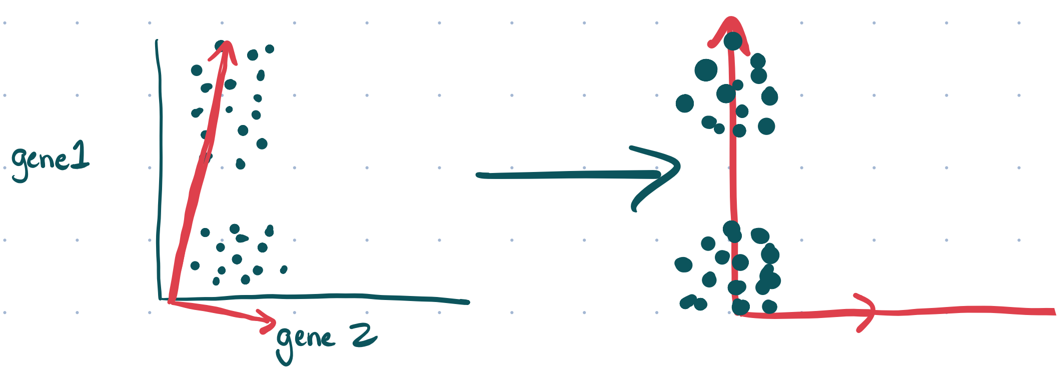
notice the size of each corresponds to how much of the data's total variance is accounted for by that component

Eventually the amount of variance becomes increasingly accounted for

Even if we had many genes...



- Here we wouldn't expect to get much information out of more than 2 components



Factor Analysis
+
Principle Component Analysis

When? 

- Your data has many measures. Almost too many to holistically interpret.
- Your measures are correlated.
- My phenomenon cannot be directly measured and thus how do I get it to vary with other things? (it's latent!)

Why?

- Allows you to summarize complex data with a set of smaller (in number) representative variables.
- Can help identify / confirm latent structure in data.

What?

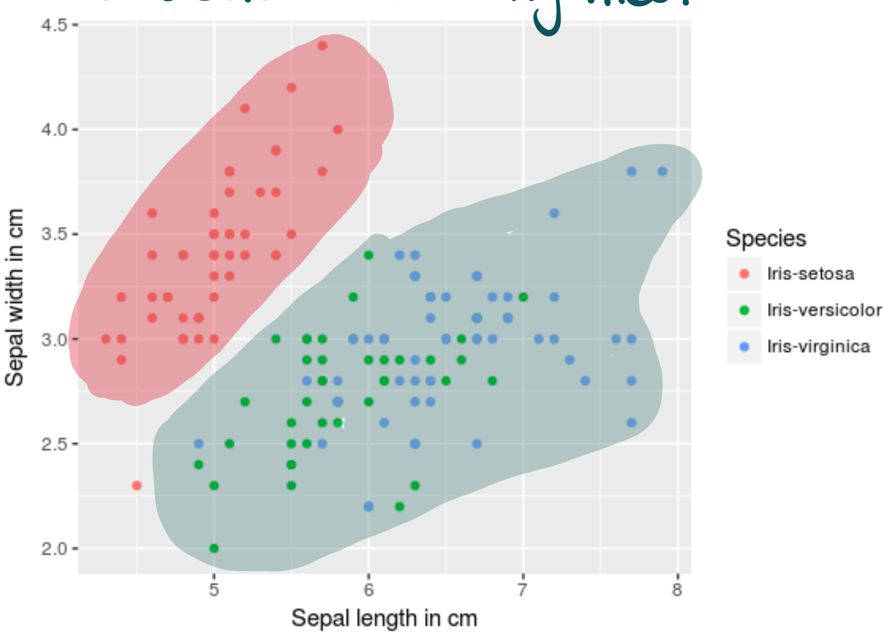
- Factor Analysis identifies where the most variance is in your data.
- This allows you to
  (1) hone your focus on what actually matters

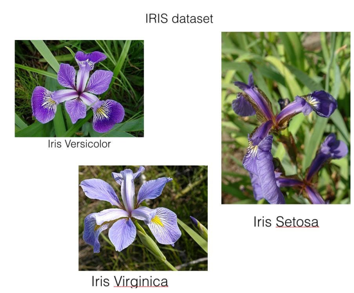
by looking at some measures + not others
(2) visualize complex data by ignoring where this is no variance and only showing where there is
(3) Identify if some of your measures are either redundant, or capturing *potentially* the same phenomenon.

How? and then How?+™

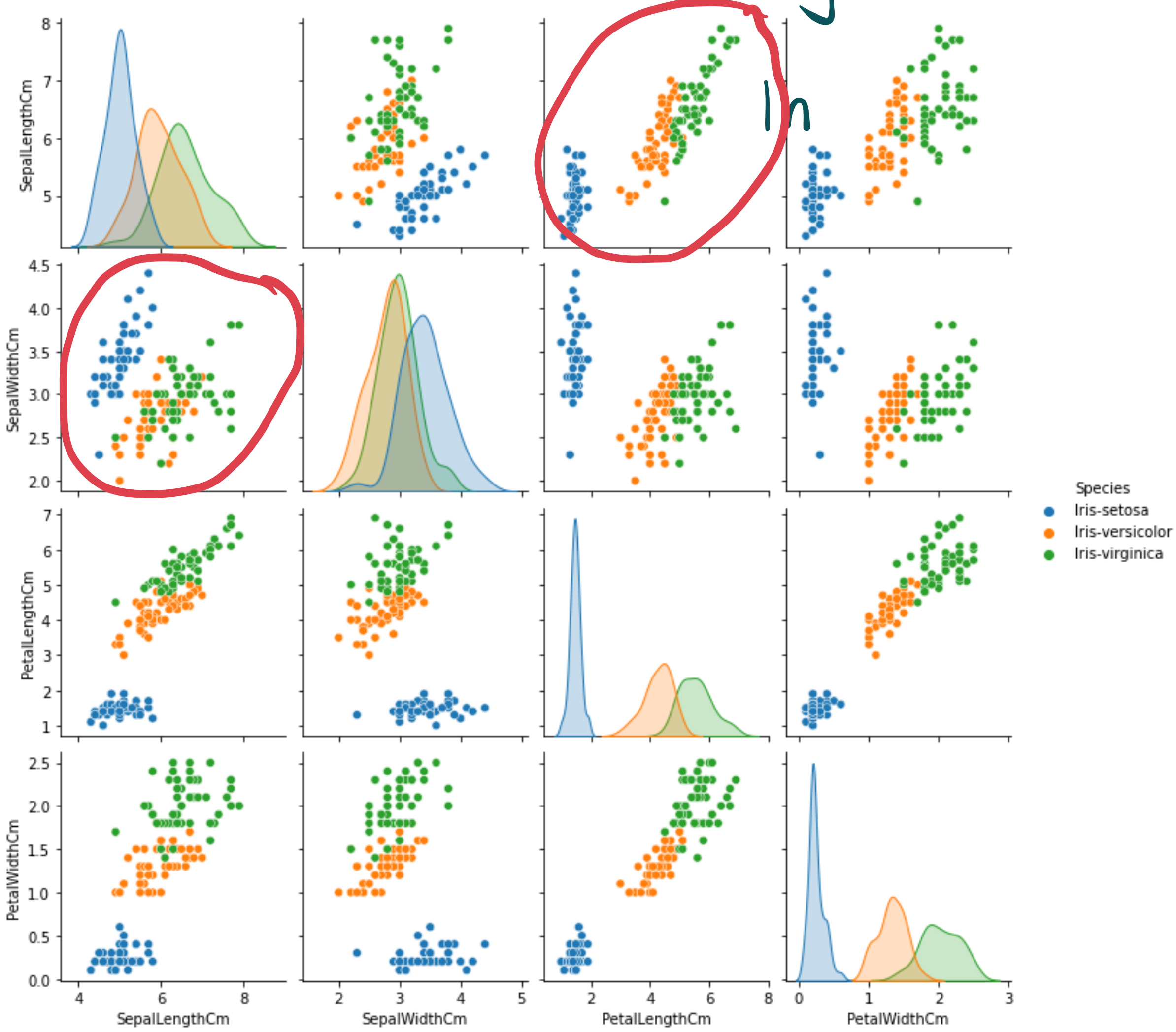
What good does this 2D visualization do us?

- this contrasts "setosa" with "versicolor" and "virginica"

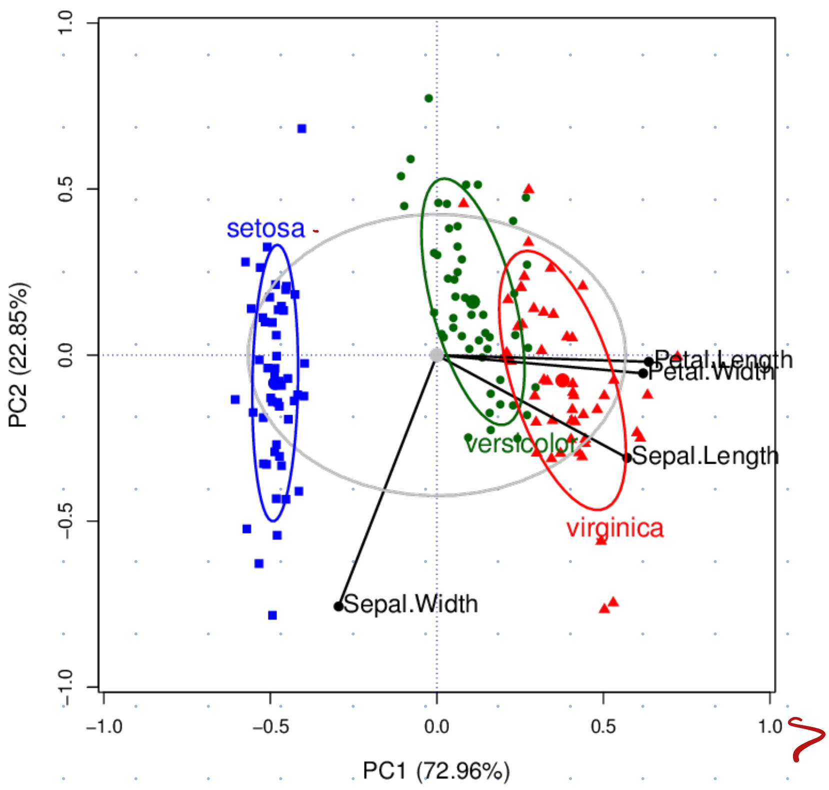




• but what if we didn't know the groups ahead of time, and we had many measurements?



• What low dimensional representation captures the most variance in the data?



1.0
0.5
0.0
−0.5
−1.0
PC2 (22.85%)
setosa
versicolor
virginica
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
−1.0
−0.5
0.0
0.5
1.0
PC1 (72.96%)

How to compute the direction of maximum variance.

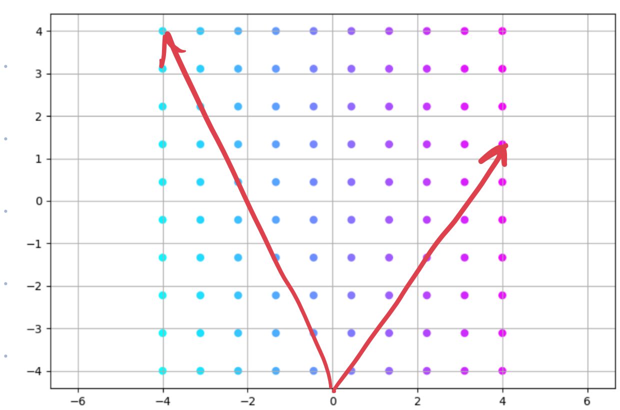
1. Ye Olde Matrix Multiplication
2. Matrices Do Things
3. Eigenvectors + Eigenvalues
4. Correlation Matrix as a Matrix that Does Something
5. Eigen Vectors of Correlation Matrix

A Timeless Passtime

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

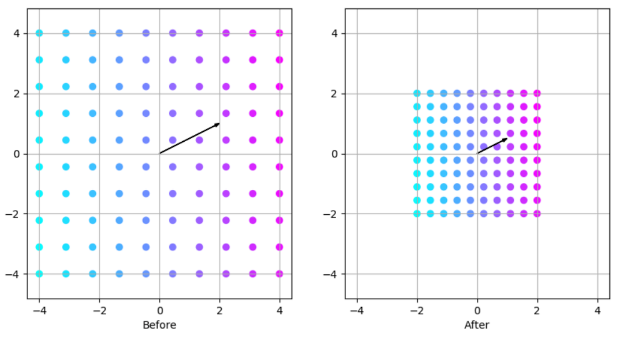
just a new vector
in a new direction

Matrices Do Things



SCALING

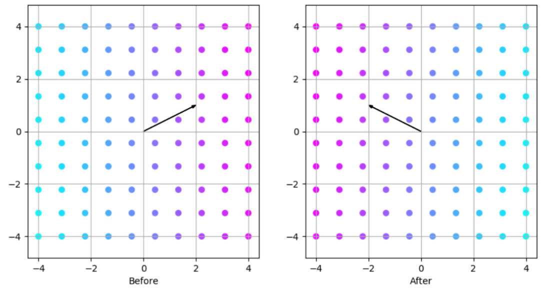
$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



REFLECTION

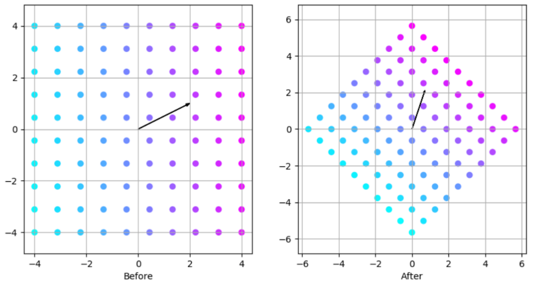
+Eigenvectors

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$



ROTATION

$$A = \begin{bmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{bmatrix}$$

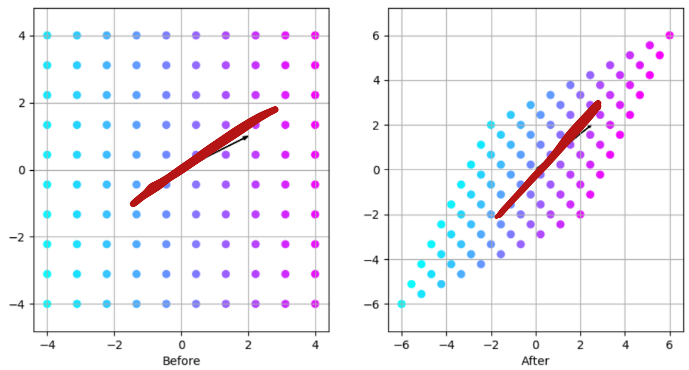


Eigenvectors

DIAGONAL SQUEEZE

$$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

(Looks kind of like our "covariance" matrix in PCA)

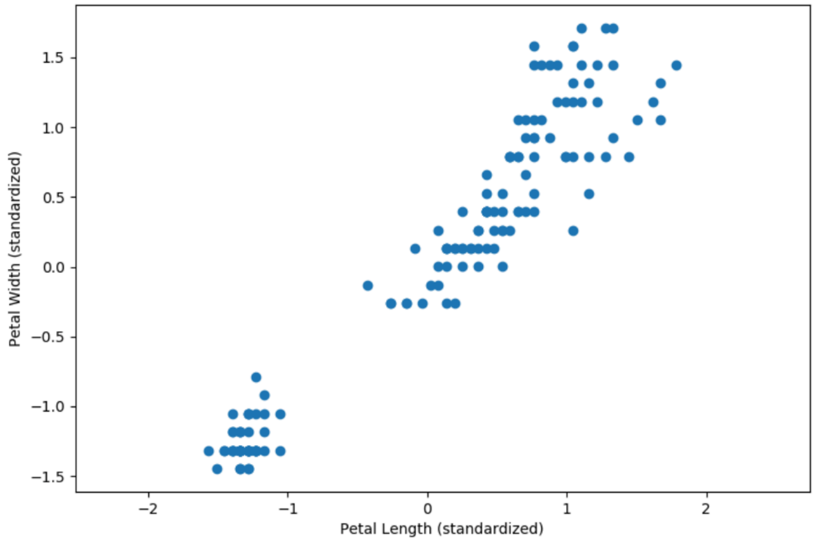


- What's the one vector that would never get squeezed?

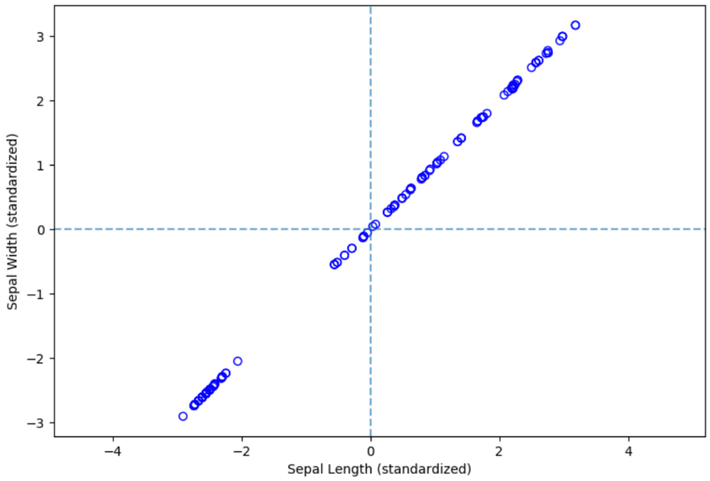
$$\underset{\text{matrix}}{M}\vec{v} = \underset{\text{some number}}{\lambda}\vec{v}$$

A haunting discovery

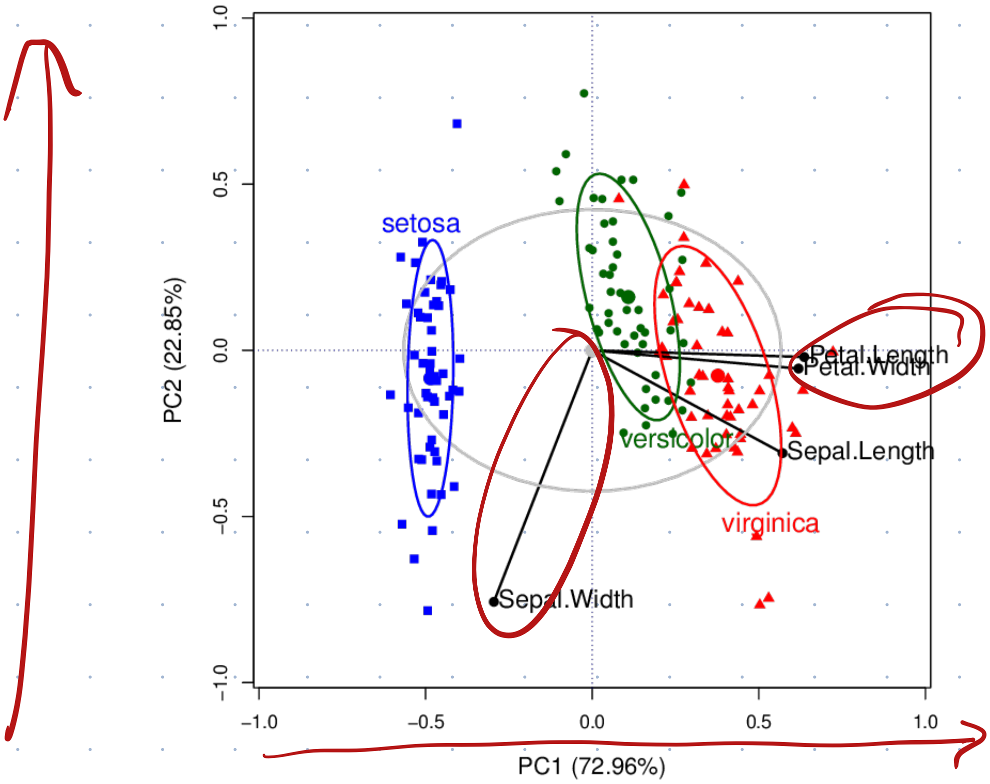
- The correlation matrix is a matrix and matrices do things. What does the correlation matrix do!



$$\Sigma = \begin{bmatrix} 1 & 0.96 \\ 0.96 & 1 \end{bmatrix}$$



The correlation matrix is a squeeze in the direction of the most variance!



setosa
versicolor
virginica
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
PC1 (72.96%)
PC2 (22.85%)
-1.0
-0.5
0.0
0.5
1.0