Principal component analysis

BDA 2025

Exploratory data analysis

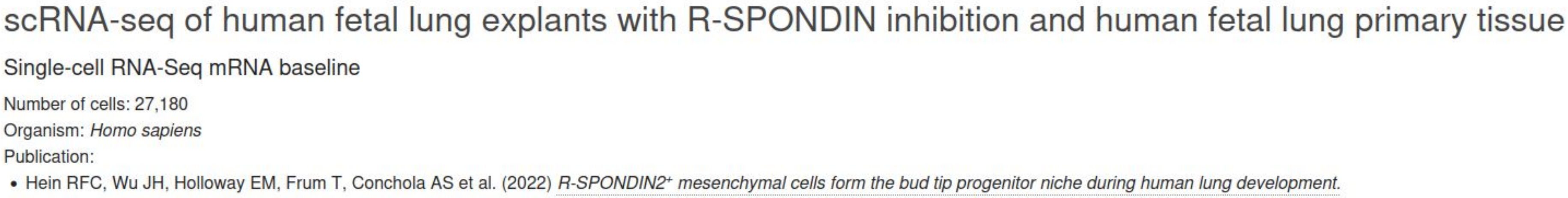
Data is often high dimensional.

It is very hard to observe the structure of high dimensional data at a glance.

1. Plotting is not possible for more than 3 dimensions.
2. Displaying the data as a table is ineffective.

We need methods to summarize the data and get hints of its structure.

Example: single cell transcriptomics



scRNA-seq of human fetal lung explants with R-SPONDIN inhibition and human fetal lung primary tissue

Single-cell RNA-Seq mRNA baseline

Number of cells: 27,180

Organism: *Homo sapiens*

Publication:

- Hein RFC, Wu JH, Holloway EM, Frum T, Conchola AS et al. (2022) *R-SPONDIN2^{+} mesenchymal cells form the bud tip progenitor niche during human lung development.*

- 27 thousand observations of about 20 thousand genes.
- There is complex structure in the form of gene regulation, active biological processes, etc.
- Most of them are not obvious or even known.

PCA motivation

The main question motivating PCA:

In which ways are the data points different from each other?

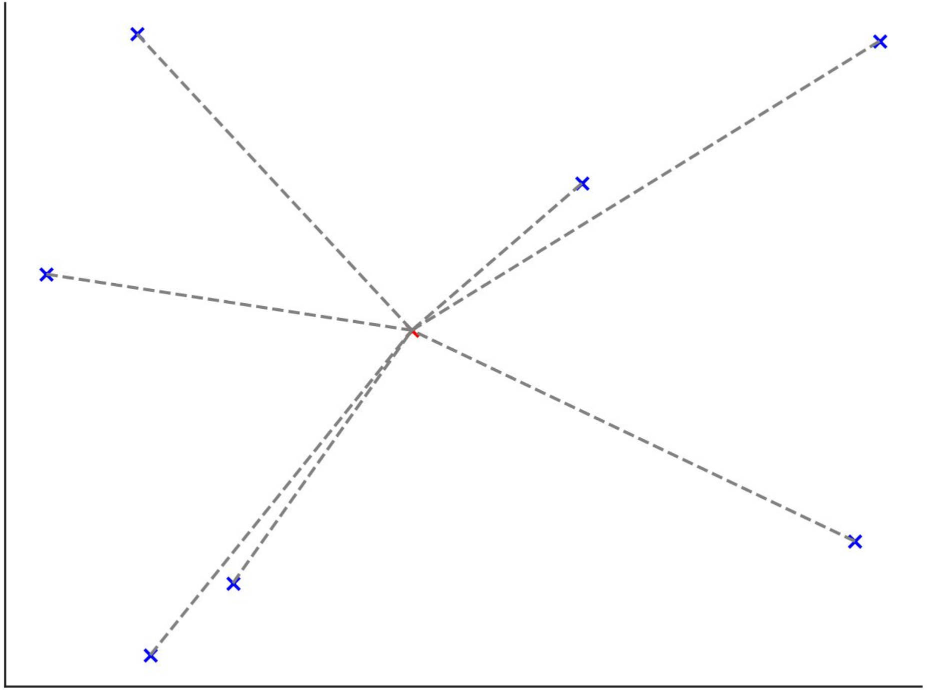
And in PCA, we use a very simple statistical measure for difference: the **variance**

Variance

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu})^2$$

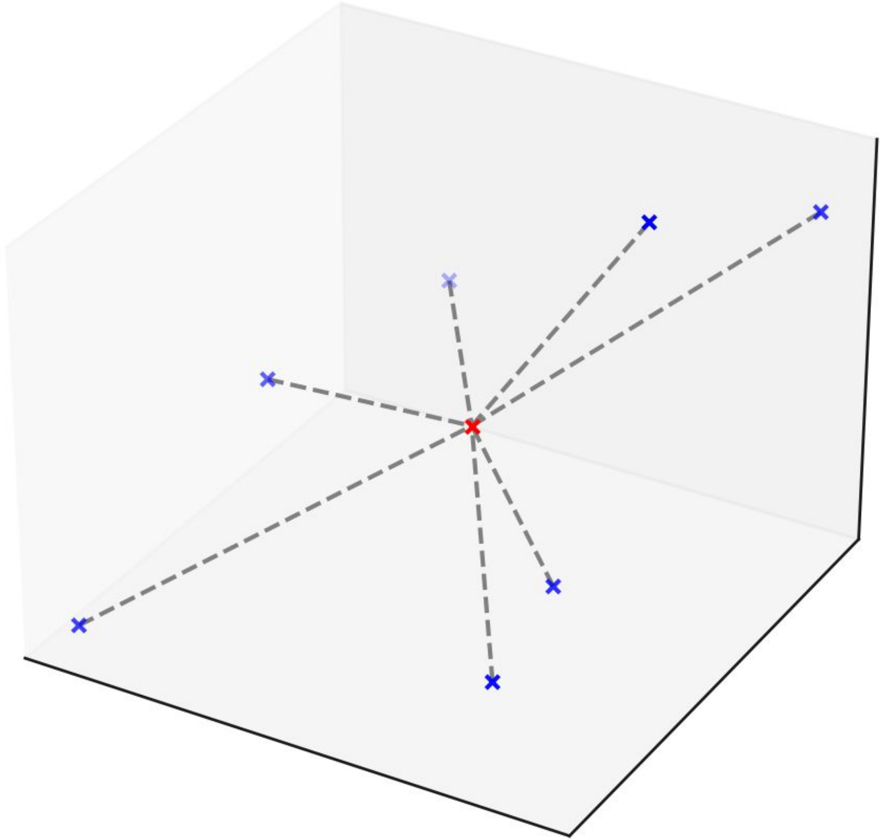
Variance

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu})^2$$

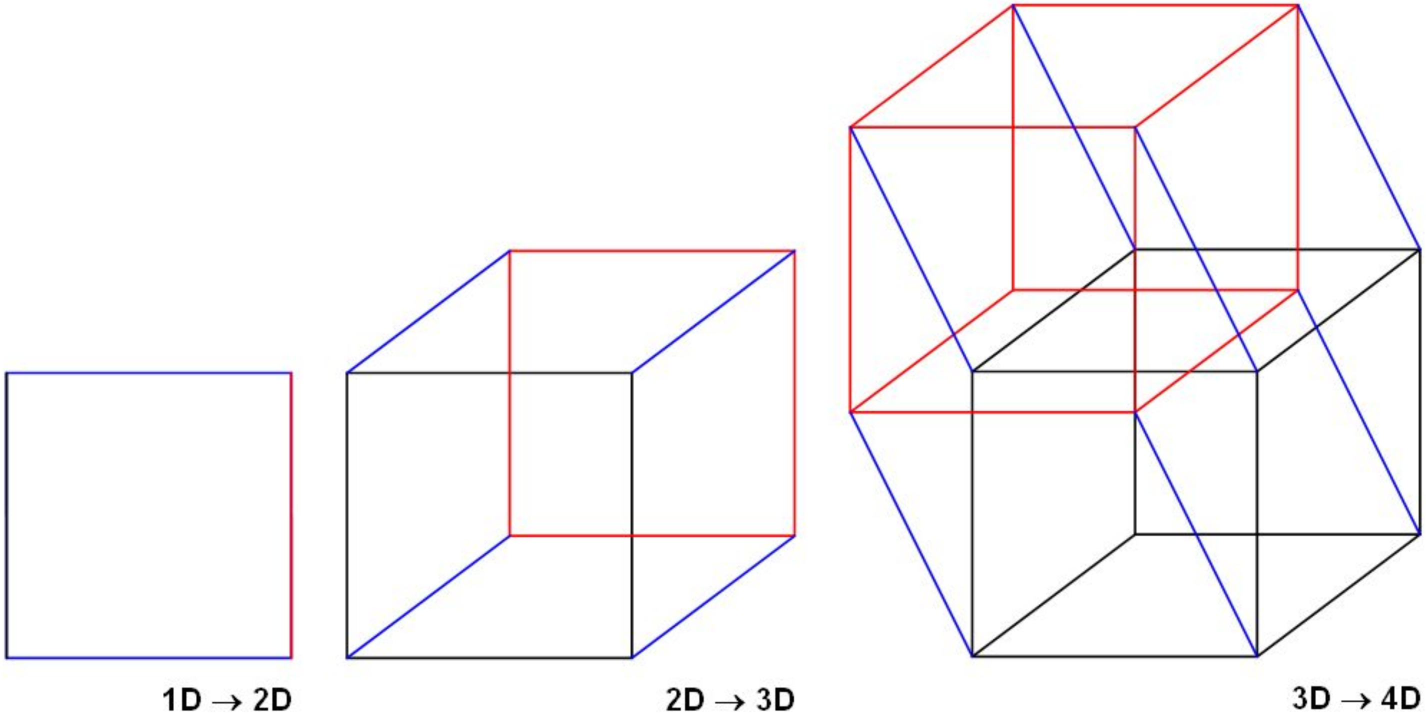


Variance

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu})^2$$

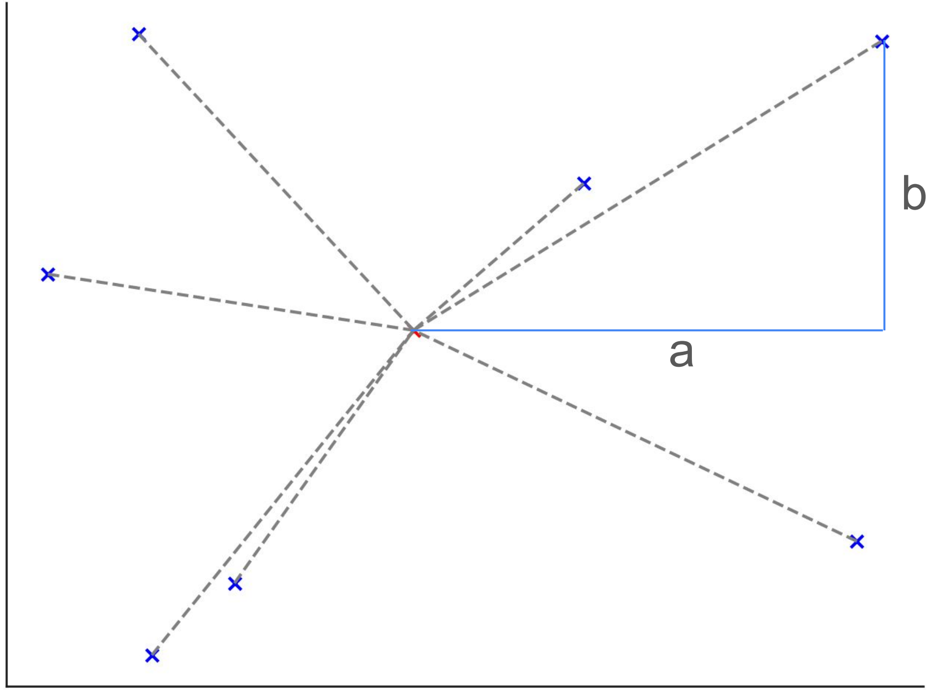


In 4D?



Variance can be decomposed along its dimensions

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu})^2$$

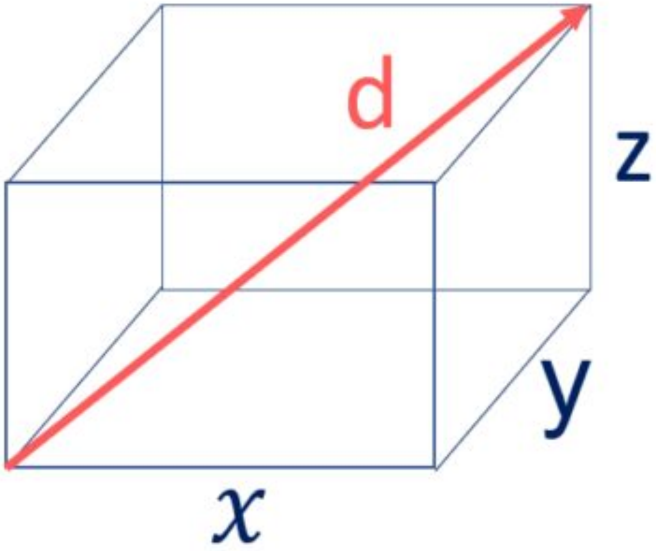


Variance can be decomposed along its dimensions

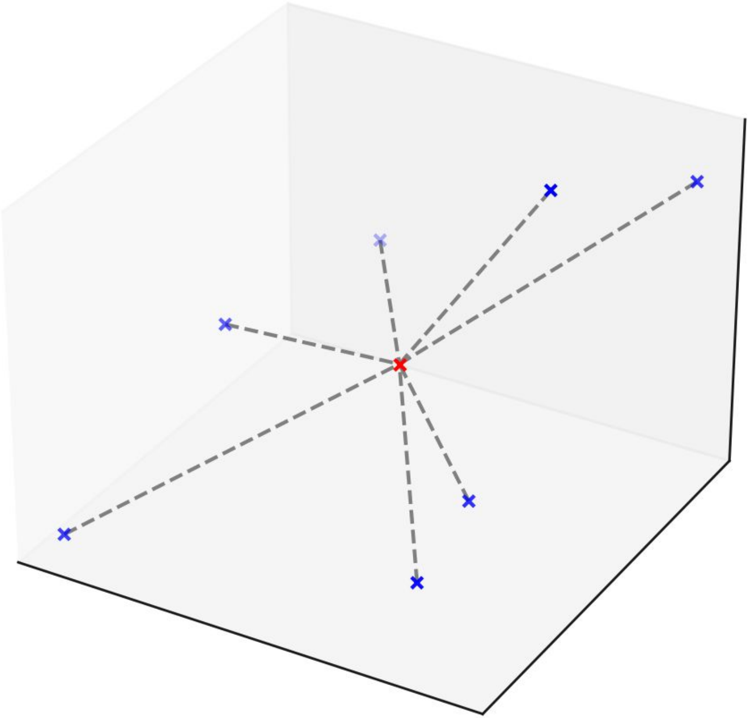
$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(a_i - \mu_a)^2 + (b_i - \mu_b)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_a)^2 + \frac{1}{n}\sum_{i=1}^{n}(b_i - \mu_b)^2$$

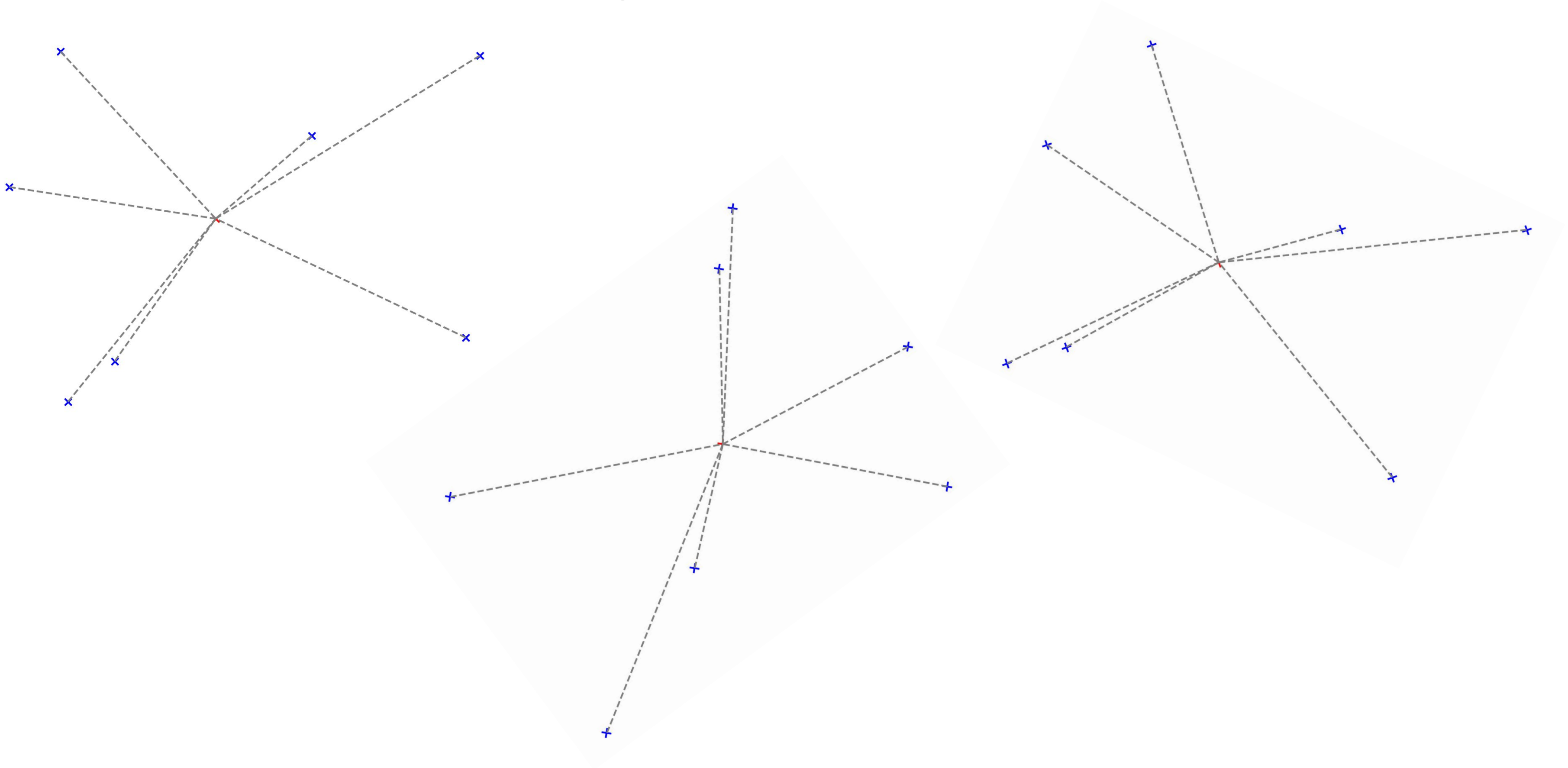
This also works for 3 or more dimensions



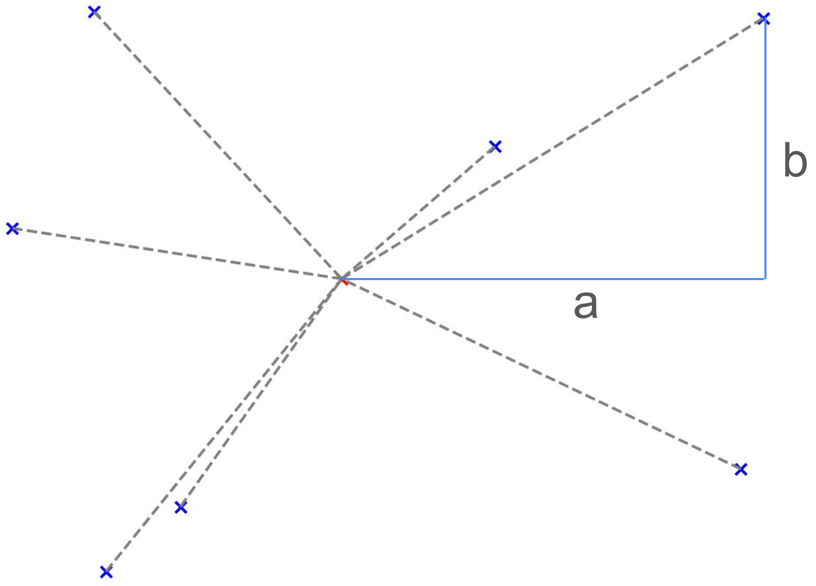
$$d^2 = x^2 + y^2 + z^2$$

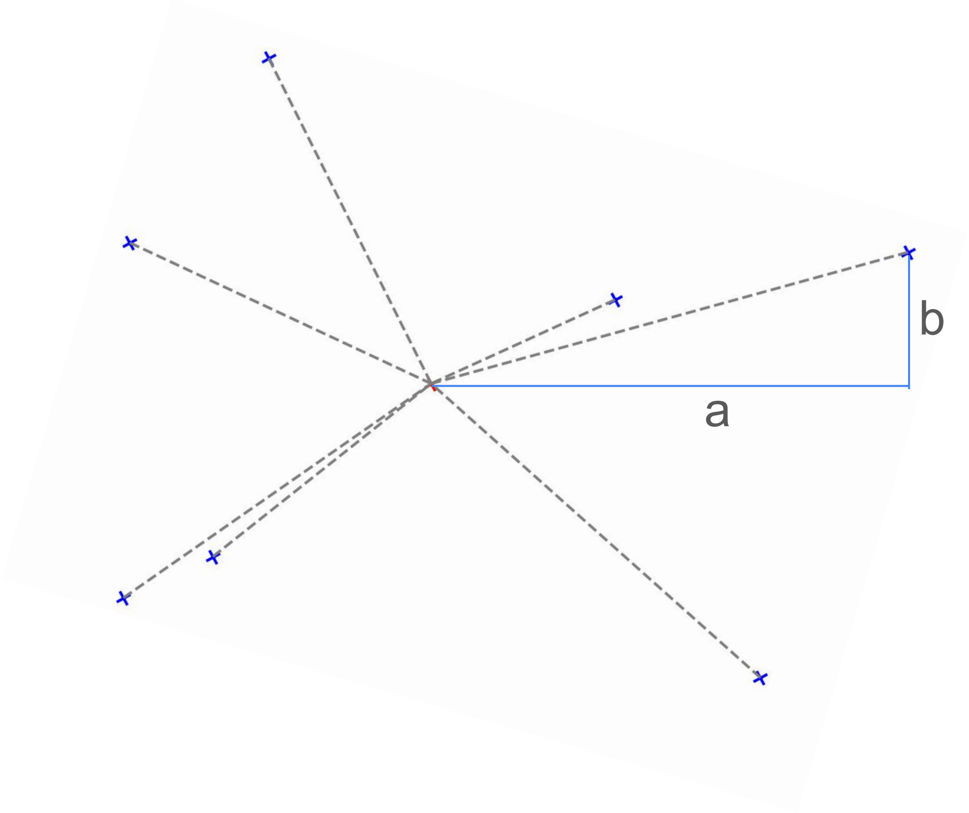


Variance does not change with translation or rotation



But when I rotate, the portion of variance in each dimension changes





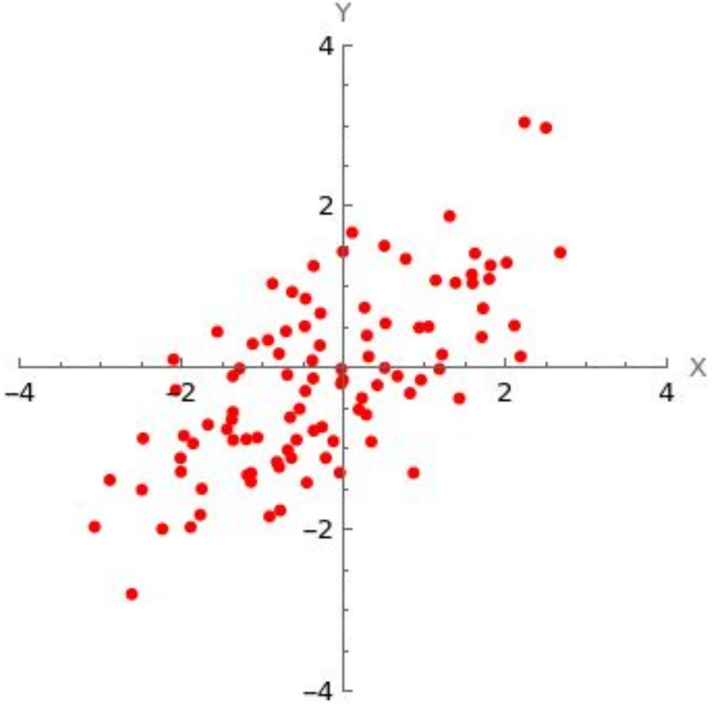
The intuition

In PCA, we want **most of the variance** to be present in **as few dimensions as possible**.

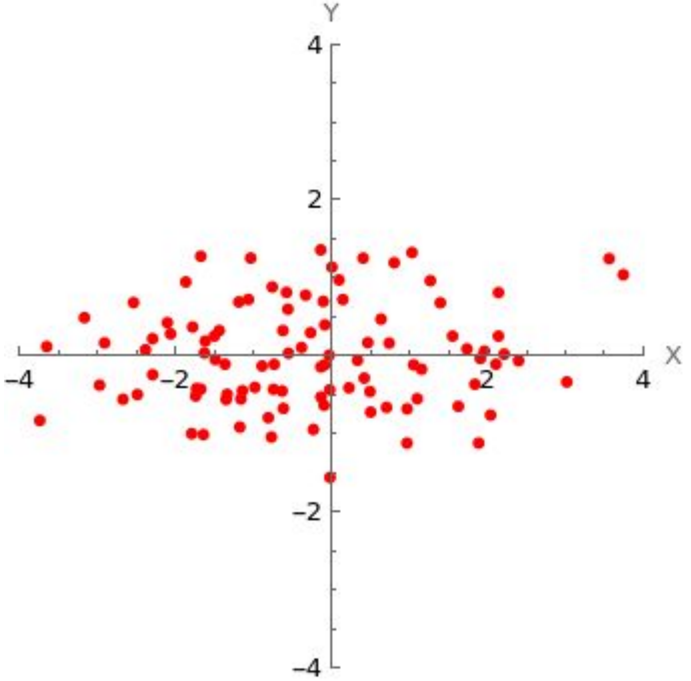
We do so by rotating the data.

Example

Var X: 1.75449, Var Y: 1.1902,
Sum: 2.9447



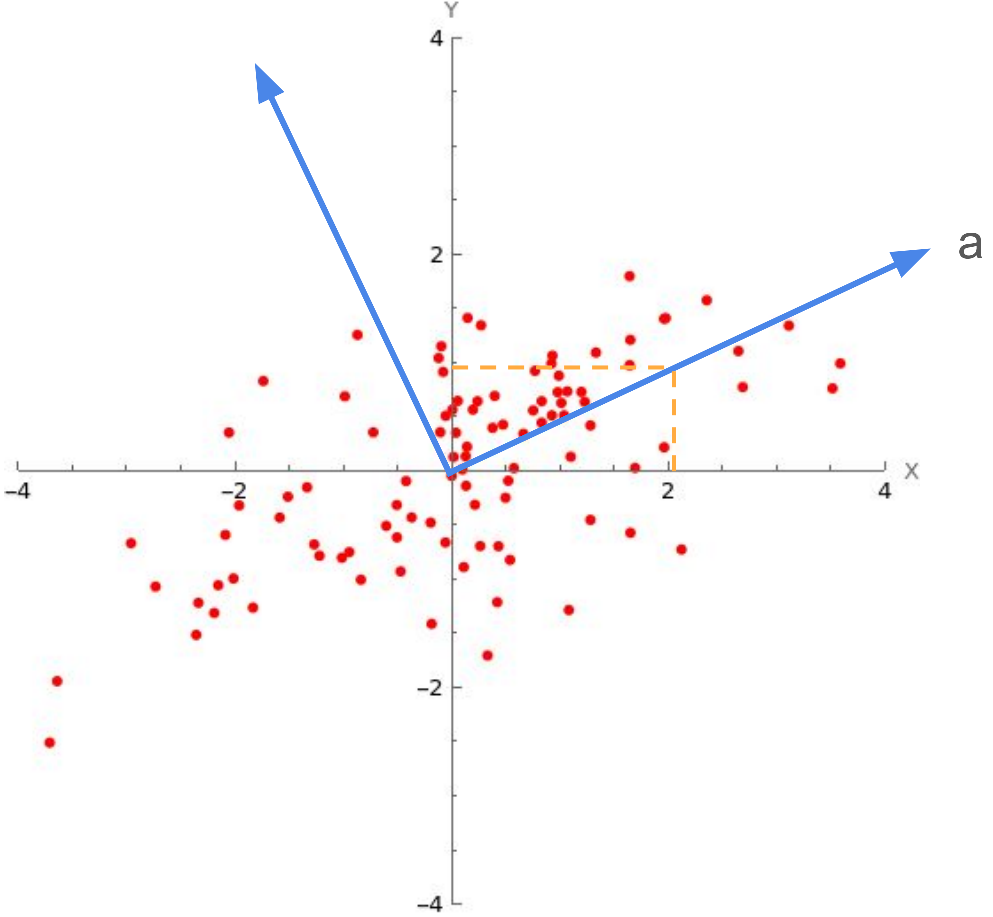
Var X: 2.50881, Var Y: 0.435888,
Sum: 2.9447



Principal components

These new coordinates are called the **principal components**, and they are combinations of the original data dimensions.

$$a = 2.1x + 0.9y$$



Principal component analysis

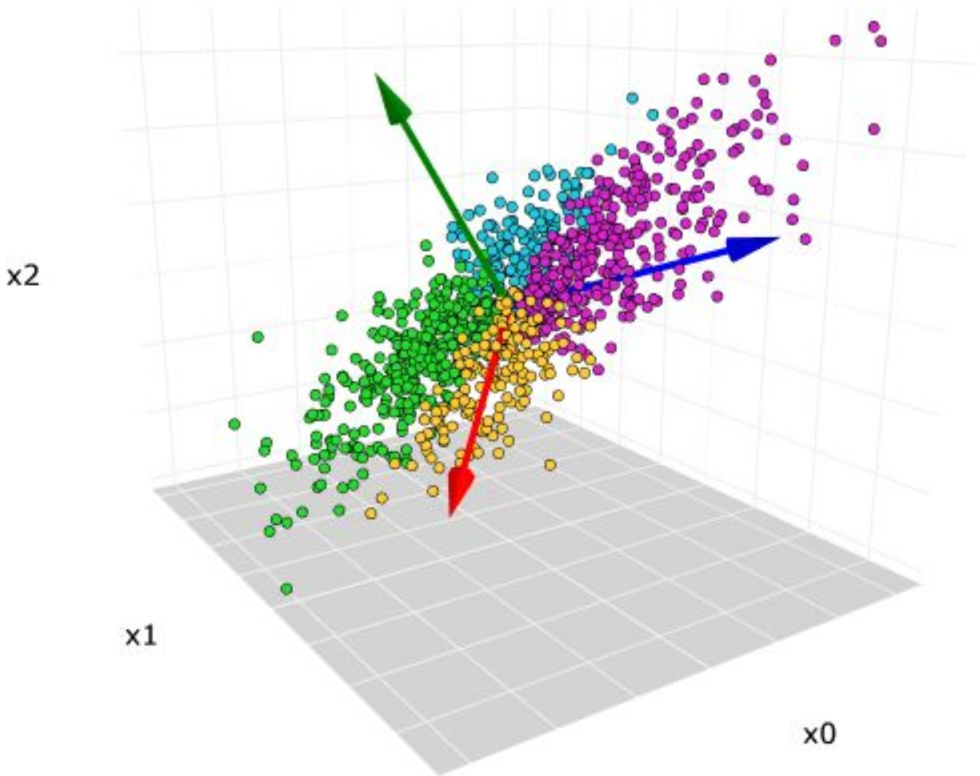
PCA is a rotation of the data into a new coordinate system, such as:

- The greatest variance lies in the first coordinate
    - This is the first principal component
- The second-greatest variance in the second coordinate
- And so on

There are as many principal components as there are dimensions

The general process

- Find the first principal component, the general coordinate on which the variance in the data is highest.
- With the first component fixed, find the next coordinate that contains the most variance and is perpendicular to first.
- Repeat this process until all components are found.


x2
x1
x0

How to find them?

These are mathematical ways to find such components, the most common being though Singular Value Decomposition.

If you are interested on the details, see this video.

But for now, let's focus on the results of such algorithms.

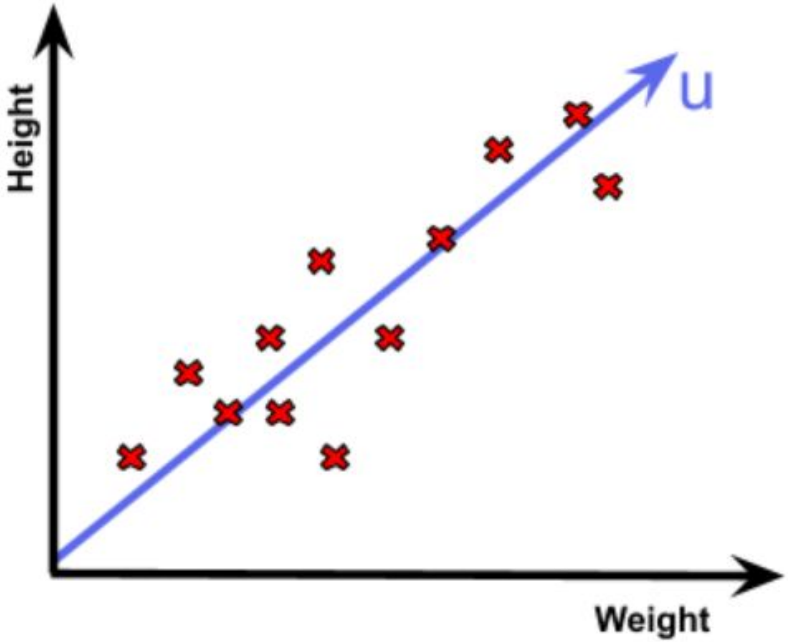
After a PCA is performed, it will output **loadings** and **score** vectors.

Principal component loadings

The loadings of a principal component are a vector describe its direction.

They are a linear combination of the original dimensions of the data.

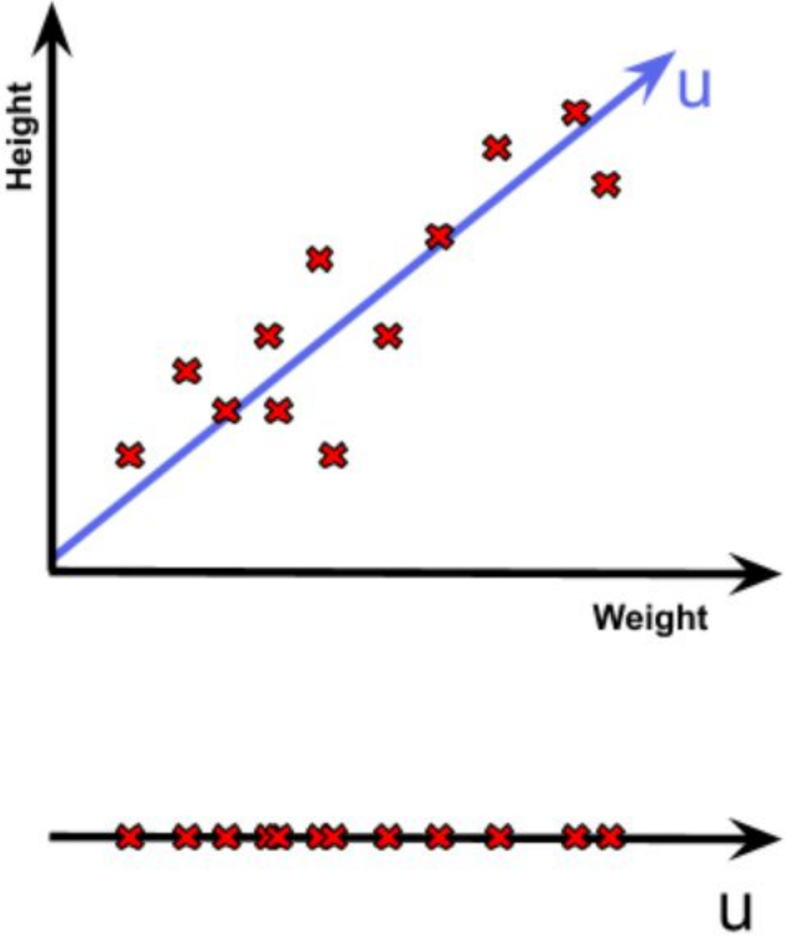
Q: What is the size of each loading vector?



Scores

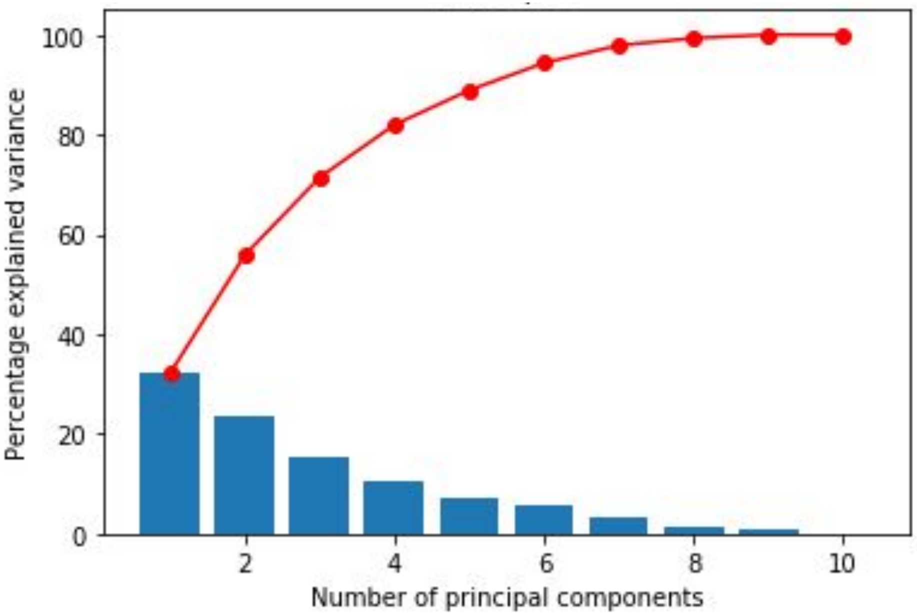
Scores are the value of the projection of each data point into the principal component.

Q: What is the size of each score vector?



Variance explained

How much of the variance in the data is present in each Principal Component



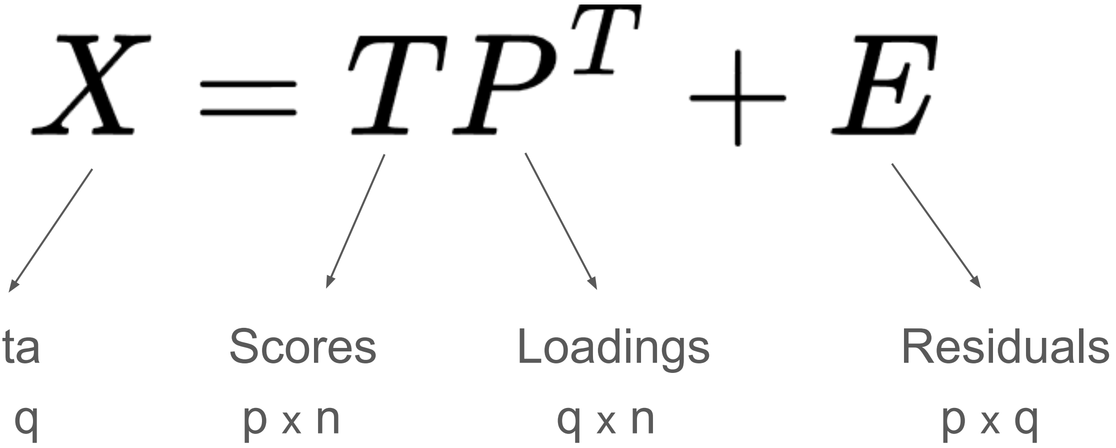
A matrix interpretation

PCA can be seen as a matrix decomposition.

$$\underset{\text{Data}}{\underset{\swarrow}{X}} = \underset{\text{Scores}}{\underset{\swarrow}{T}}\,\underset{\text{Loadings}}{\underset{\searrow}{P^T}} + \underset{\text{Residuals}}{\underset{\searrow}{E}}$$

Using *n* principal components

The dimensions of the matrices are



$$X = TP^T + E$$

Data

p x q

Scores

p x n

Loadings

q x n

Residuals

p x q

p observations
q variables
n principal components

PCA as a model of the data

This decomposition can be seen as a model of the data.

The main assumption of this model is the effective rank of the data.

The data is inherently low rank, but added noise makes it full rank

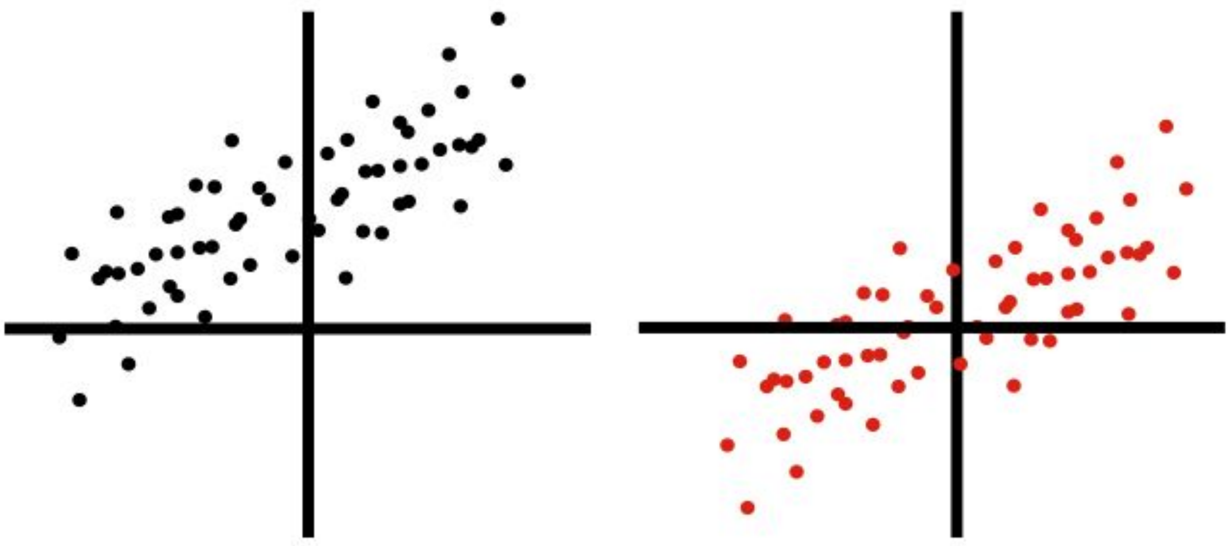
PCA as a model of the data

$$\underset{\substack{\downarrow \\ \text{Data} \\ \text{p x q}}}{X} = \overbrace{\underset{\substack{\downarrow \\ \text{Scores} \\ \text{p x n}}}{T}\underset{\substack{\downarrow \\ \text{Loadings} \\ \text{q x n}}}{P^T}}^{\text{Low rank approximation}} + \overbrace{\underset{\substack{\downarrow \\ \text{Residuals} \\ \text{p x q}}}{E}}^{\text{Noise (full rank)}}$$

Best practices

Mean centering

When performing PCA, it is important to mean center the data



Mean centering

When performing PCA, it is important to mean center the data.

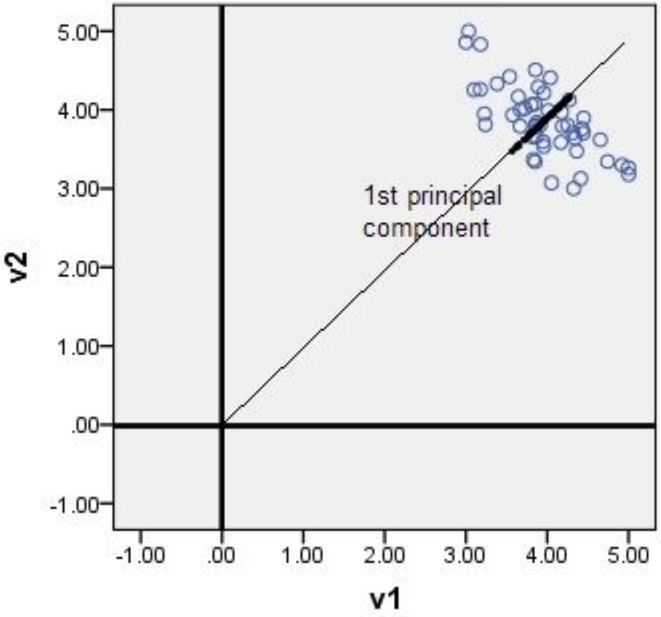
The main reason is linear algebra: through matrix multiplication we rotate the data around the origin.

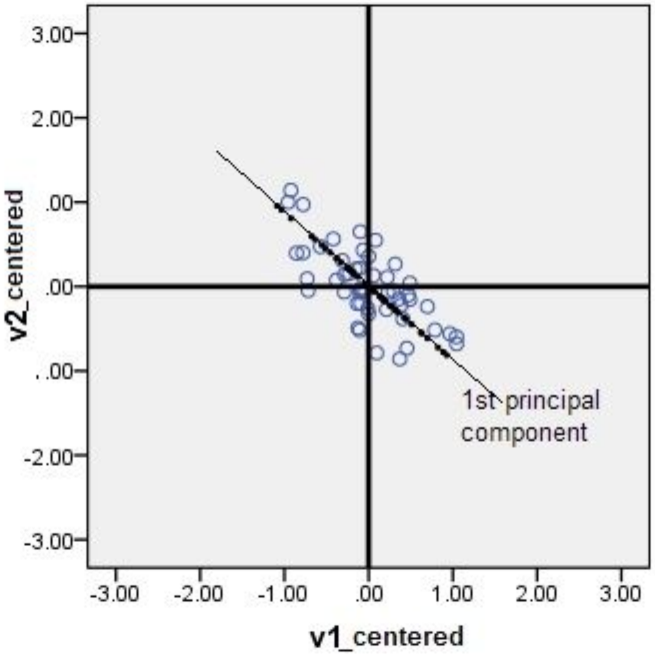
So PCA assumes the origin to be the mean of the data.

Centering is done my subtracting the mean of each column.

Mean centering

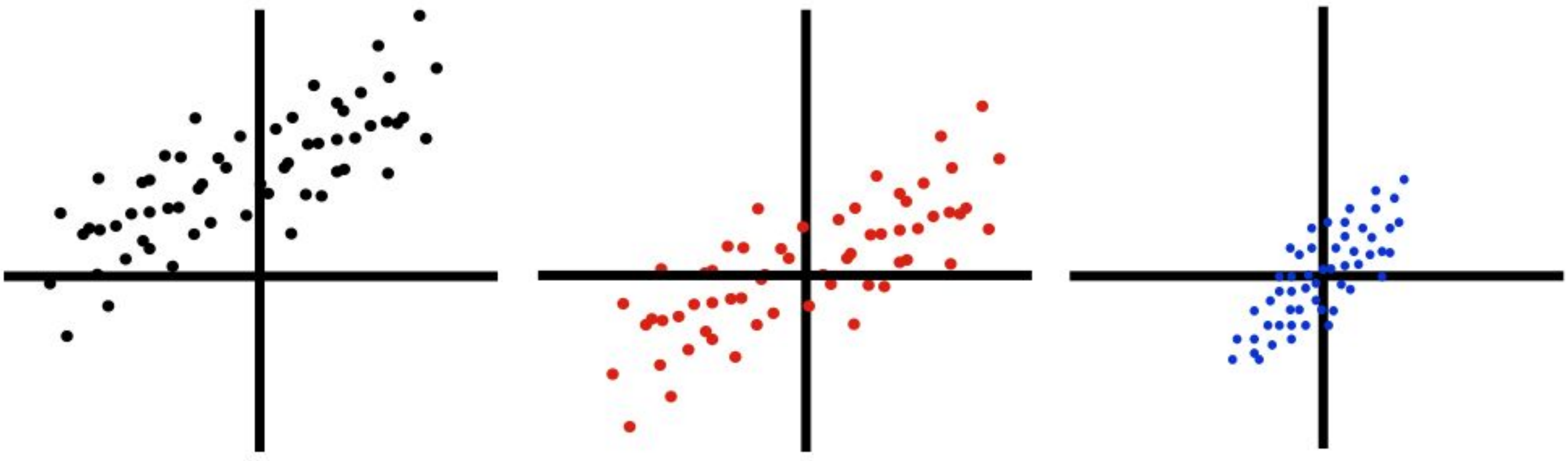
So without centering, the first principal component will only point to the center of the data





Scaling

After centering, it is important to scale the data so the variance is the same in all dimensions



Scaling

After centering, it is important to scale the data so the variance is the same in all dimensions.

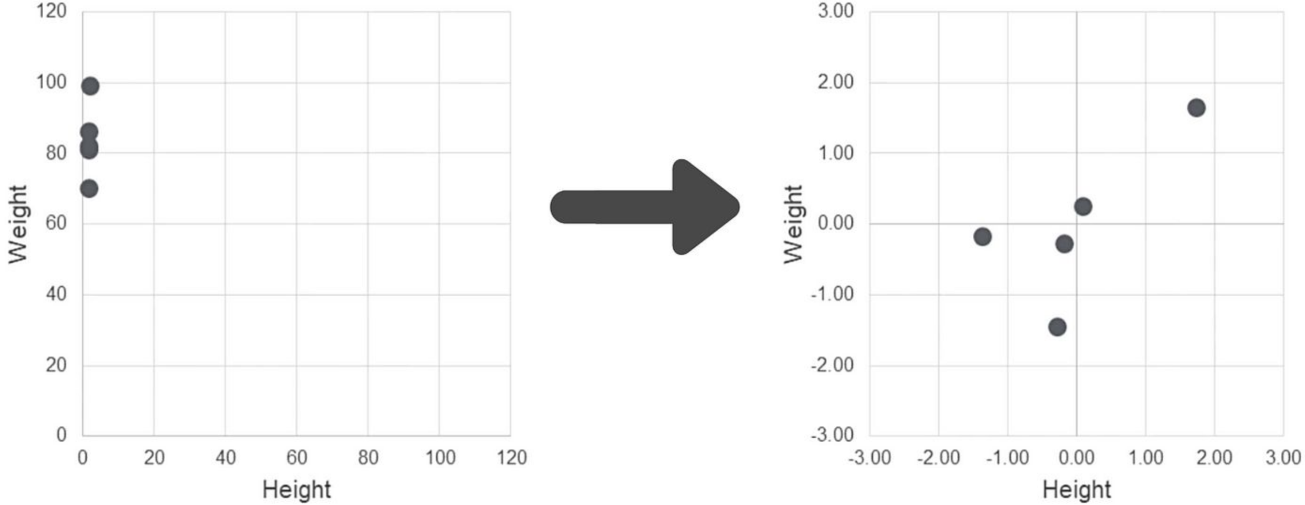
This is done by dividing each column by its standard deviation.

Q: Why could this be important? Why do we change the variance?

Scaling

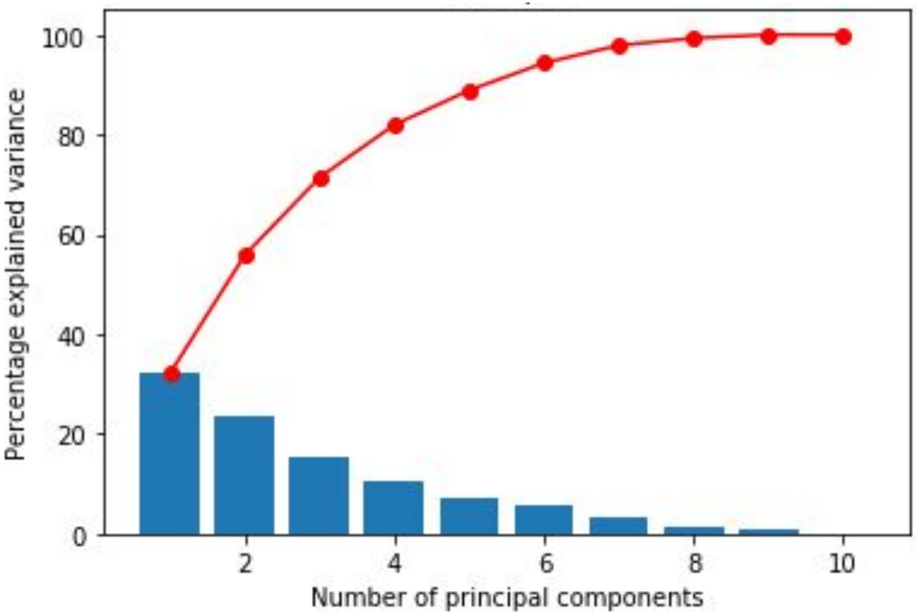
Different variables are measured in different units.

If not scaled, the principal components will align with variables whose absolute numbers are bigger.



Variance explained

Whenever you choose to use a specific number of principal components, it is always important to report how much of the variance they represent.

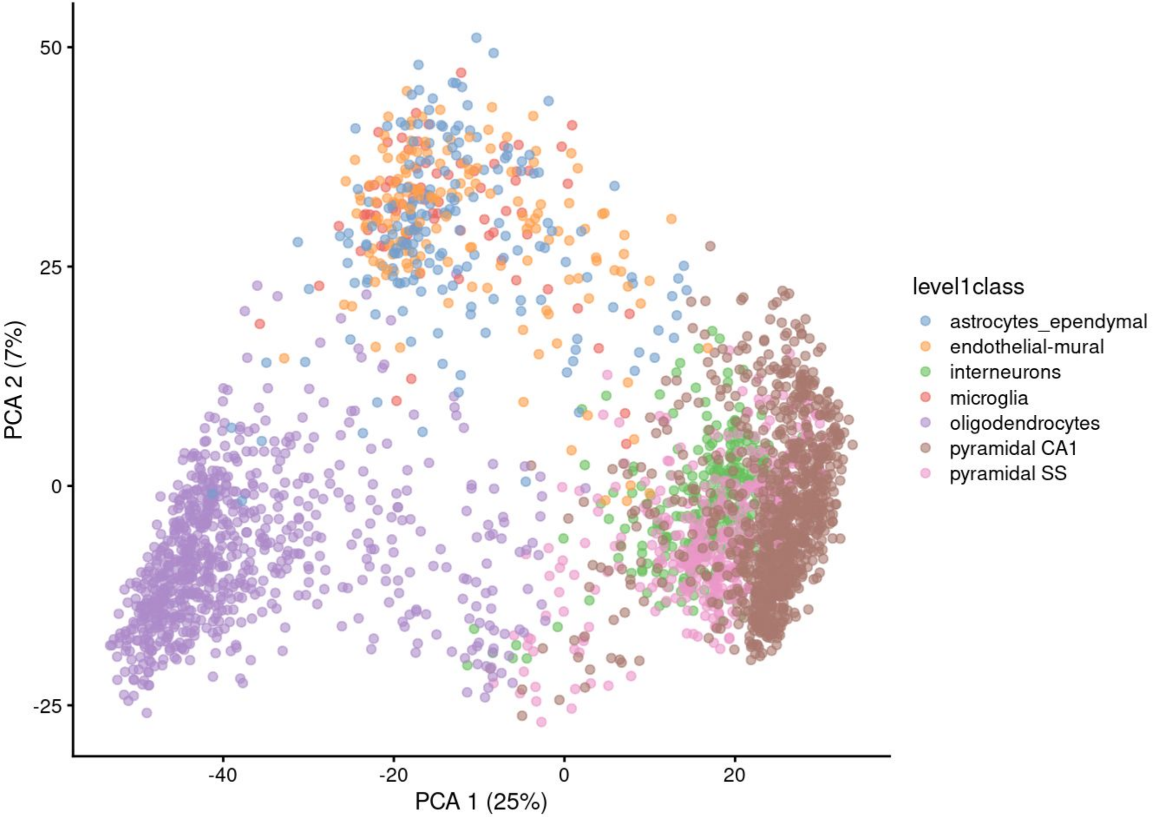


PCA in practice

In practice, PCA is used as:

- A data visualization technique
- A dimensionality reduction method
- An analysis method to understand the main sources of variation in the data

As a vizualization technique

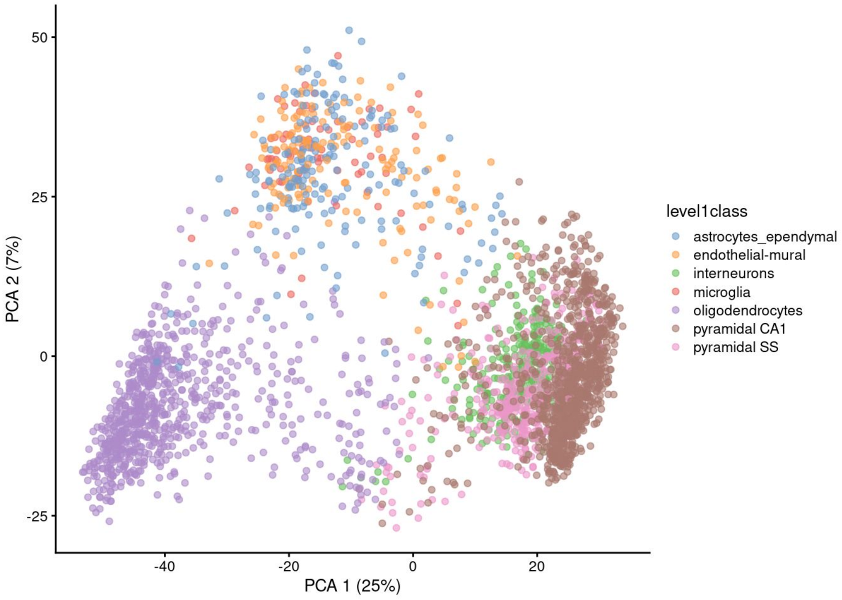


As a vizualization technique

Vizualizing data is an easy way to analyze different measurements simultaneously.

But, we are limited to two dimensions on paper and screen.

Always report the explained variance!

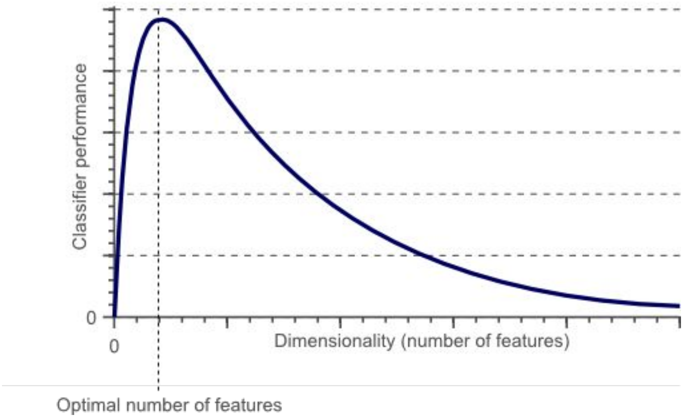


As a dimensionality reduction method

The performance of many statistical and machine learning methods will go down with an increase in number of dimensions.

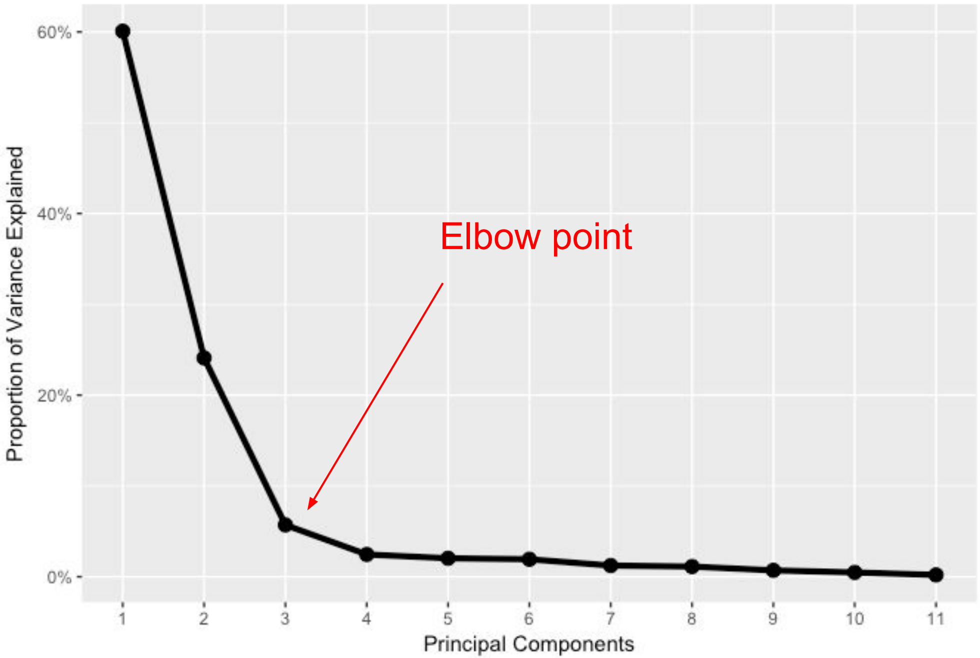
This is known as the **curse of dimensionality**.

PCA can be used to reduce the number of dimensions while keeping most of the important information.



How many principal components to use?

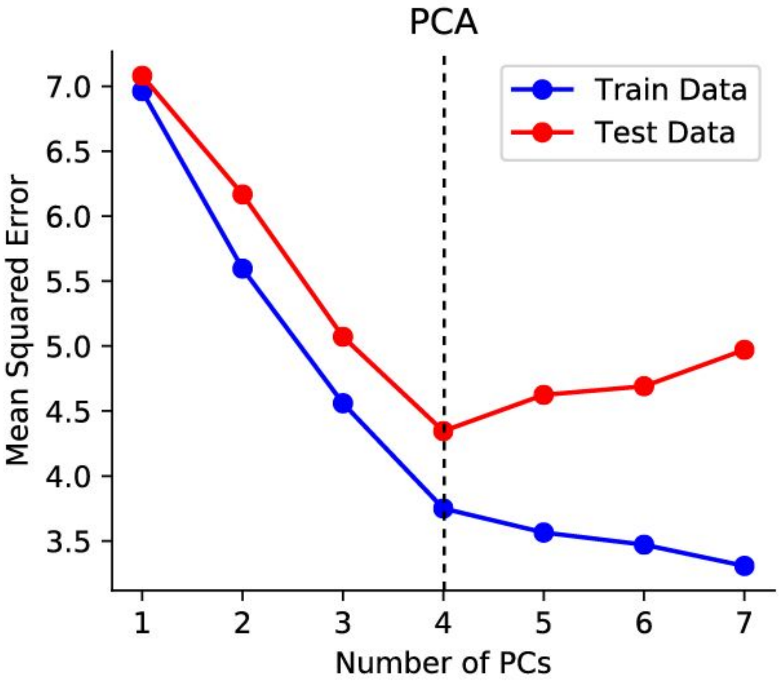
The elbow method



How many principal components to use?

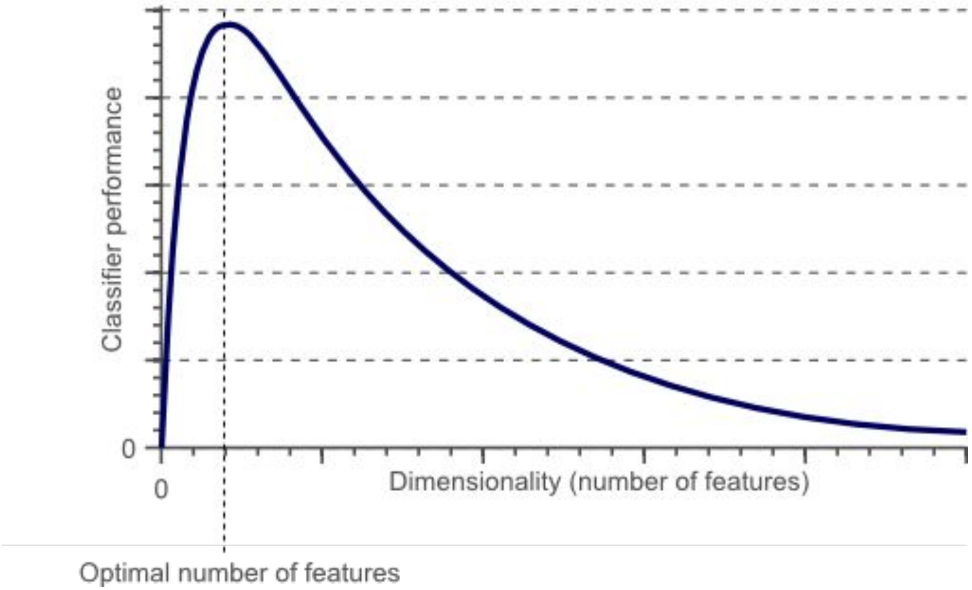
Using cross validation

$$X = TP^T + E$$



How many principal components to use?

Or simply directly calculating performance



To understand sources of variation in the data

The first principal components represent the main direction of variation in the data.

Their loadings can be interpreted to understand where this variation is coming from.

To understand sources of variation in the data

Example: houses

Variable	PC1 Loading	PC2 Loading
Area	0.42	0.01
Bedrooms	0.38	0.02
Bathrooms	0.35	0.03
Age	-0.12	0.40
Distance to public transportation	-0.10	0.45

To understand sources of variation in the data

Gene regulation patterns and biological processes.

Gene	Loading
Gene1	0.12
Gene2	0.25
Gene3	-0.31
Gene4	0.18
...	...