Genomic Data Visualization

Feb 2, 2022

Homework 0 -> Blackboard
-> grades will be uploaded to Blackboard
-> questions drop by office, reach out via Slack

Quiz
- everyone is understanding MERFISH and Visium
- helped identify some common points of confusion in how we describe and communicate about data visualization
-> how to make a data visualization, how to make to make a good, how to distinguish and describe a good vs bad one

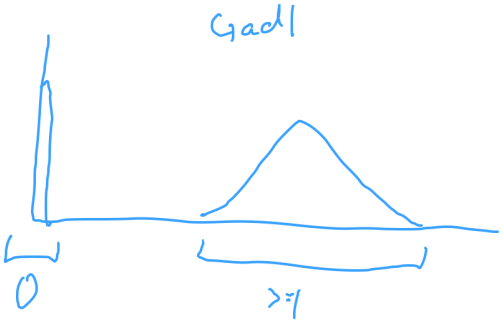
Homework (due Saturday Midnight)
- making a data visualization and describing it

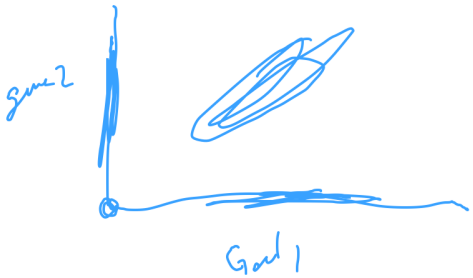
-> go through this together

Principle components analysis (PCA)
- dimensionality reduction

Learning objectives:
- understand how PCA works intuitive
-> giving a more intuitive explanation of PCA rather than regoing through the linear algebra
- apply PCA to our data

Motivate





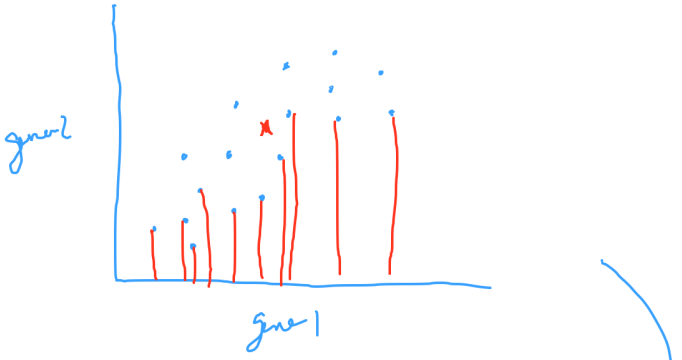
$\mathbb{R}^N \rightarrow \mathbb{R}^M \qquad M < N$

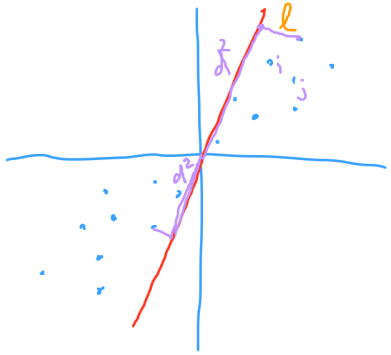
$N = 1600$
$N = 500$

$M = 2, 3$

PCA on 2 genes $\rightarrow$ 1 D

$\mathbb{R}^2 \rightarrow \mathbb{R}^1$



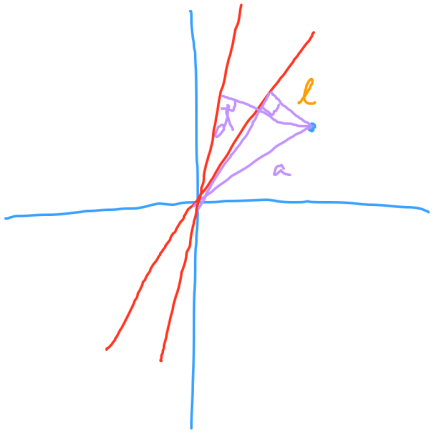


PCA starts with centering the data

Fit a line that maximizes the distance from the projected points to the origin d

$\sum_i d_i^2$

Fit a line that minimizes the distance of the points to the line l

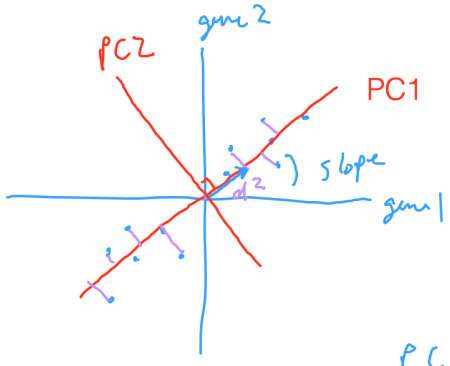


$a^2 = d^2 + l^2$

$d \uparrow \quad l \downarrow$

$d \downarrow \quad l \uparrow$

maximizing d is equivalent to minimizing l -> computationally one optimization is easier than the other



eigenvalue for PC1 = $\sum_i d_i^2$

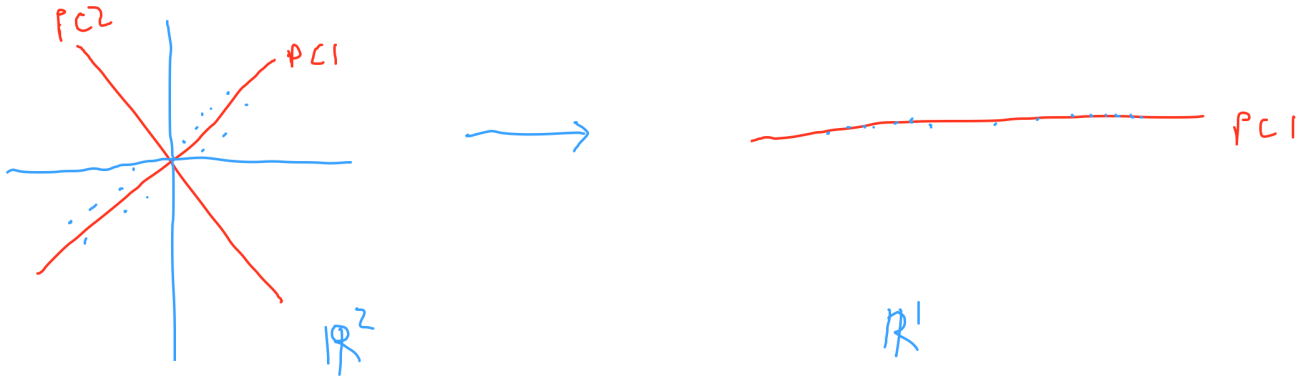
PC1 is maximizing ds, has the largest eigenvalue

unit vector along this fitted line = eigenvector

$PC1 = A\,\text{gene 1} + B\,\text{gene 2}$

linear combination of genes

genes as having loading values on PCs

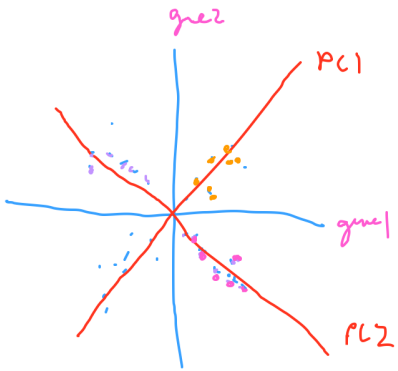


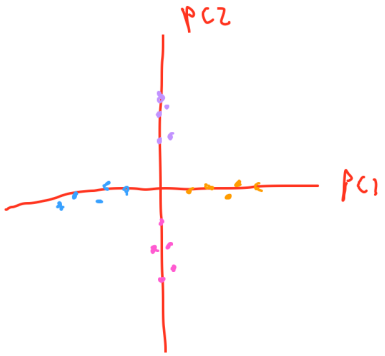
How much variance in our data is being captured by the PCs

$$Var(PC) = \frac{\sum d^2}{n-1}$$

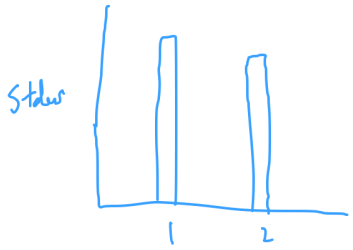
$$\text{Stdev} = \sqrt{\frac{\sum d^2}{n-1}}$$

Scree plot

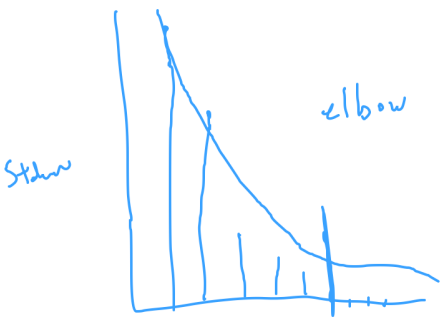


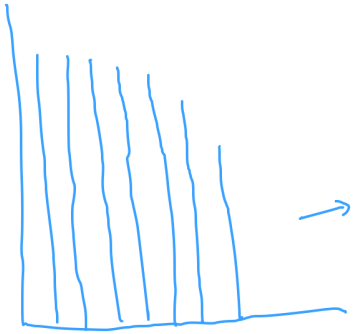


Purple and pink populations are similar to each other



How to use the Scree plots to evaluate how many PCs to look at





$\mathbb{R}^N \rightarrow \mathbb{R}^M \quad M < N$

$N = 1600$ 500

↳ M?