Module 5C: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

Principal **C**omponent **A**nalysis (PCA)

Identifies the major drivers of variation

Why PCA:

- Very few assumptions
- Non-parametric
- It **reduces** the dimensionality of your data
  - It may be surprising to you that you can reduce the dimensionality of your data without losing much information.
  - This occurs when the **variables are highly correlated**.
    - If you have included the following variables in your data set: arm length, leg length, height, you probably don't need them all – a linear combination of the three of them would capture the variation.
    - You can then use a smaller dataset of uncorrelated characteristics (or a smaller set of linear combinations of characteristics)
- Pearson, 1901 (yes, it is > 100 years old).

Principal **C**omponent **A**nalysis (PCA)

Identifies the major drivers of variation

Why PCA:

- Very few assumptions
- Non-parametric
- It **reduces** the dimensionality of your data

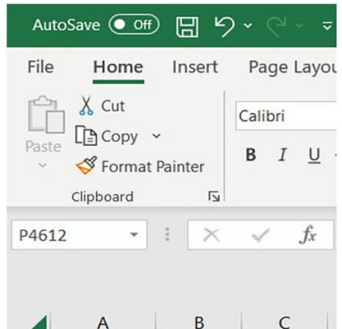
How PCA:

- It shifts the axes of your data from standard Cartesian axes (x, y, z) to a new set of axes, the PC axes, calculated from the largest Eigenvalue, and the second Eigenvalue (which is uncorrelated to the linear combination of the first Eigenvalue via….algebra. Magic, too, but mostly algebra).
- The largest Eigenvalue (the most variation) is a linear combination of the variables in your dataset.
- Therefore, these new axes demonstrate the linear characteristics that are most important in driving the variation of the characteristic under investigation.

- This is A LOT, but I found some material that helps walk us through PCA principles without algebra

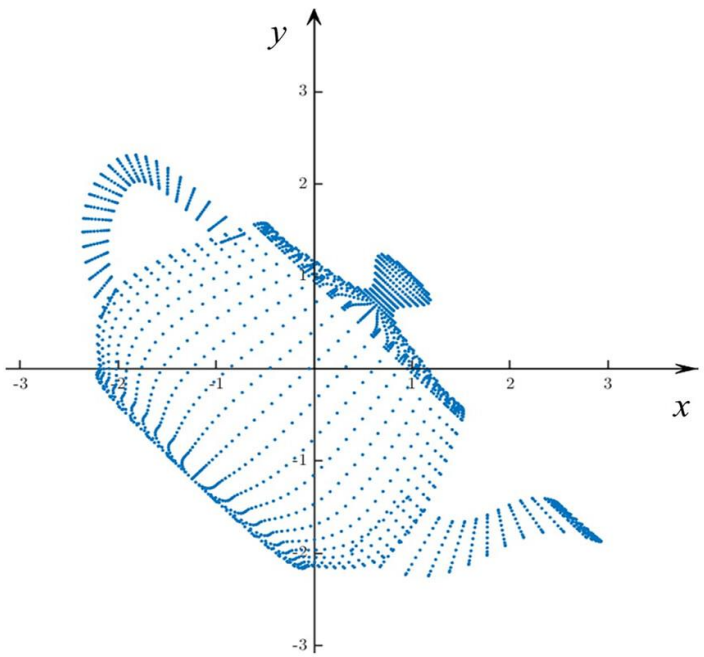
Principal **C**omponent **A**nalysis (PCA)

(A)

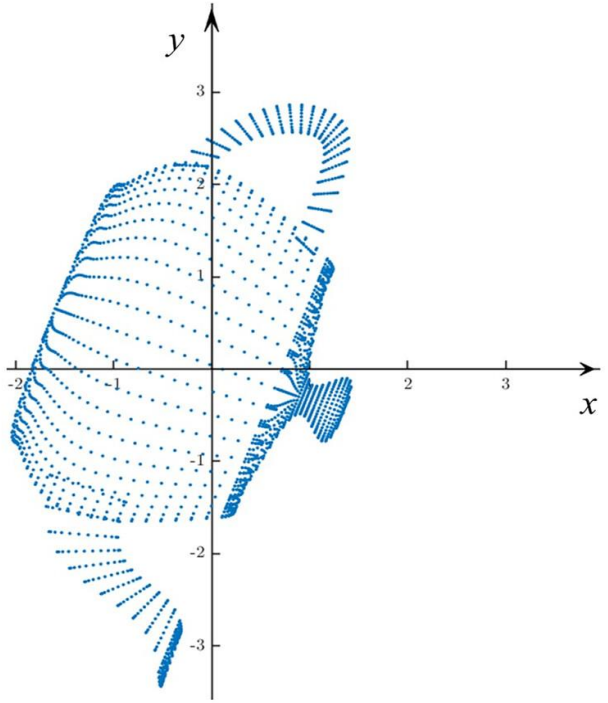


	A	B	C
1		VARIABLES	
2	Observation	x	y
3	1	0.64	-0.21
4	2	0.67	-0.22
5	3	0.70	-0.23
6	4	0.72	-0.23
7	5	0.73	-0.24
8	6	0.73	-0.24
9	7	0.73	-0.24
10	8	0.73	-0.24
11	9	0.72	-0.23
12	10	0.70	-0.23
13	11	0.67	-0.22
14	12	0.64	-0.21
15	13	0.56	-0.42
16	14	0.59	-0.43
17	15	0.62	-0.43
18	16	0.64	-0.44
19	17	0.65	-0.44
20	18	0.66	-0.44
21	19	0.66	-0.44
22	20	0.65	-0.43
23	...	...	...
4610	4608	-1.60	0.65
4611			

(B)

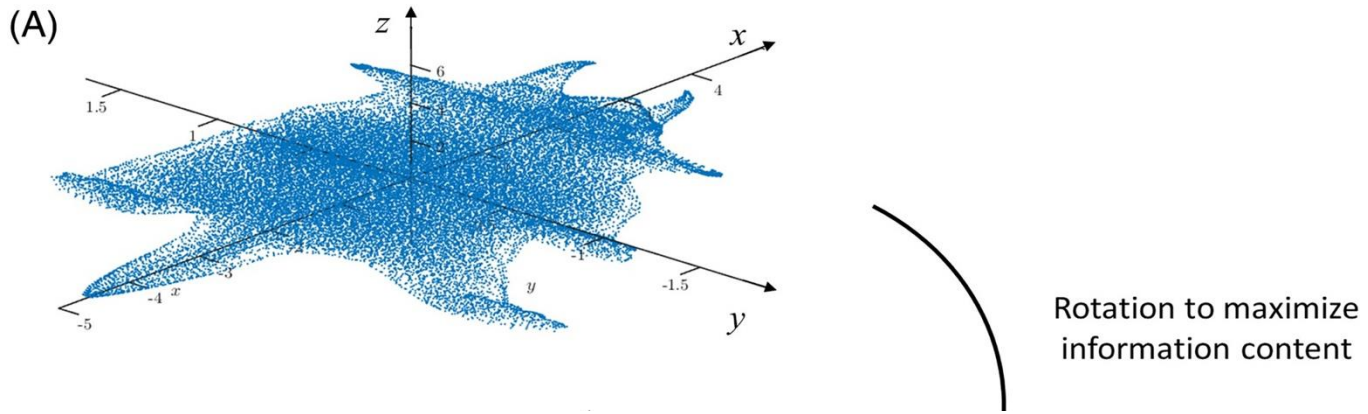


(C)



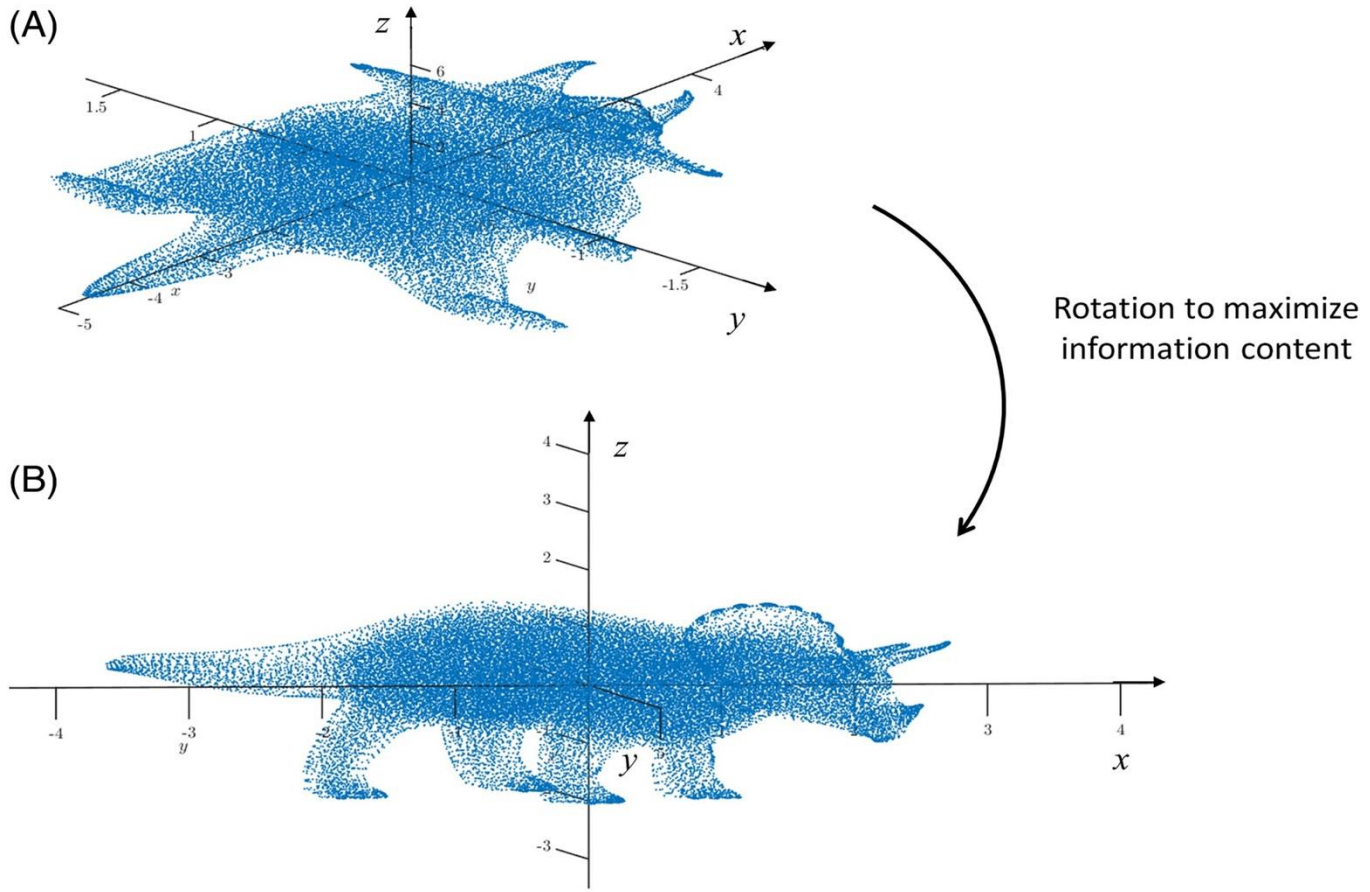
Principle **C**omponent **A**nalysis (PCA)

- Now we increase data points
- 3-D (x, y, z) for 36,876 points



Principle **C**omponent **A**nalysis (PCA)

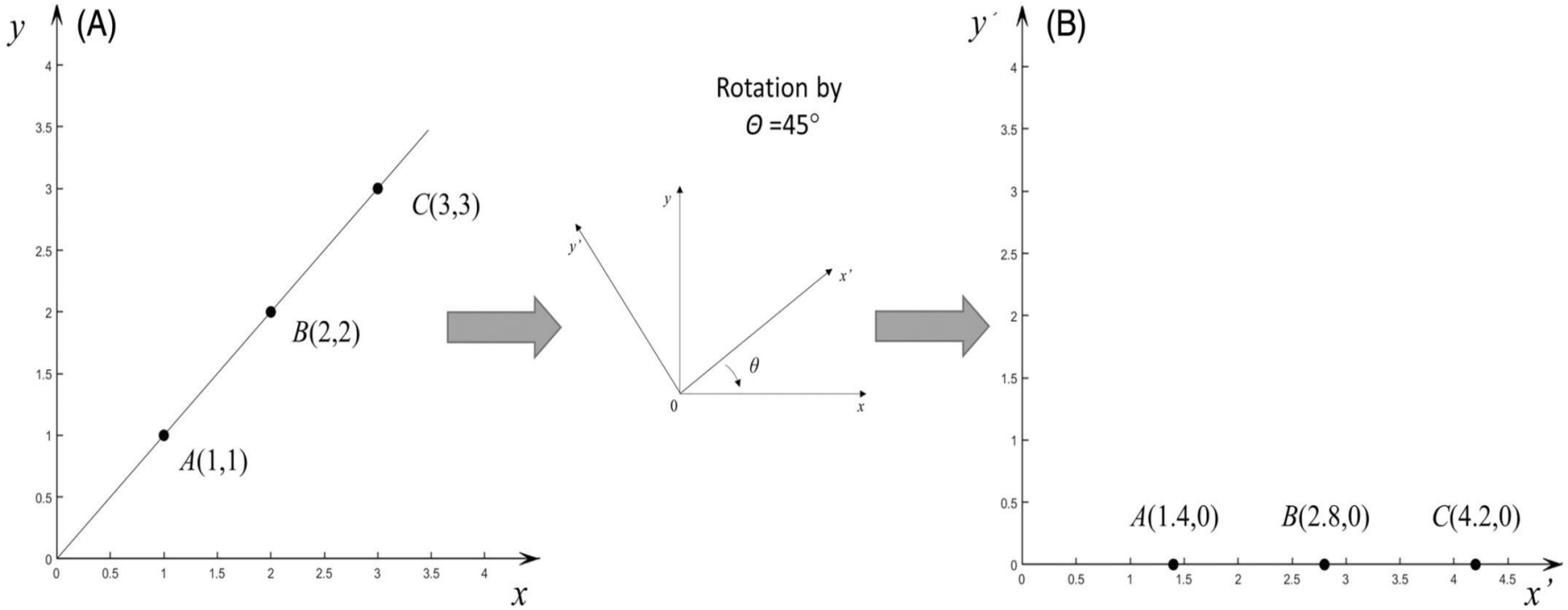
- Now we increase data points
- 3-D (x, y, z) for 36,876 points



Principle **C**omponent **A**nalysis (PCA)

- What happens if there are ≥ 4 variables? How do we uncover the structure of a high-dimensional data set?
- **Variation is information**; the more dispersed along an axis, the greater the information content is along that axis.
- You can use matrix decomposition:
  - N predictor variables → n x n covariance (or dispersion) matrix
  - Eigenvectors from this matrix gives us the direction of maximum variation
  - Eigenvalues weights the importance of the new axes

$$\mathbf{X} = \begin{bmatrix} & x & y \\ A| & 1 & 1 \\ B| & 2 & 2 \\ C| & 3 & 3 \end{bmatrix}.$$



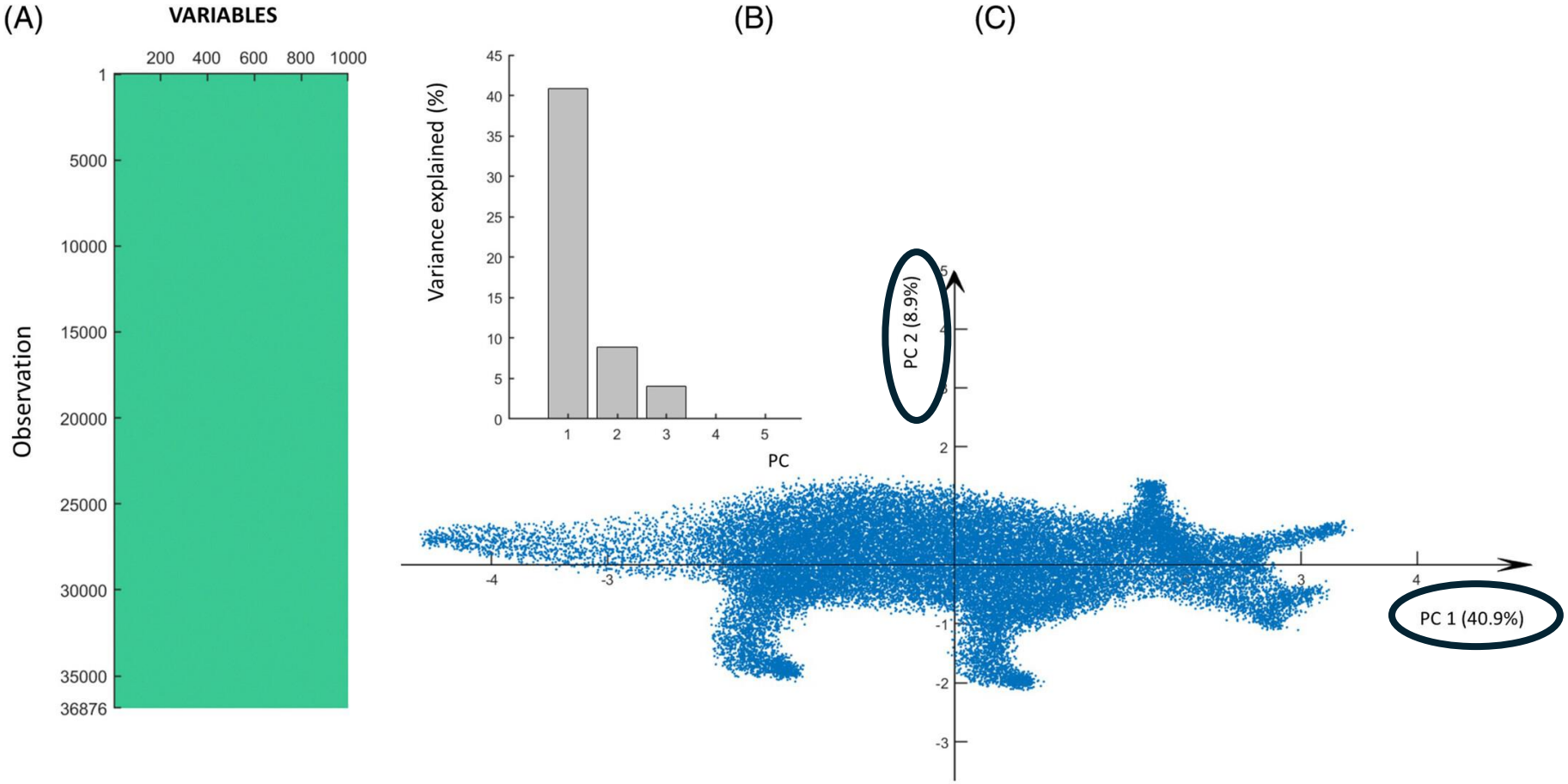
This becomes too complex as data become high dimensional.

Use Principal Components which are linear combinations of the original variables:

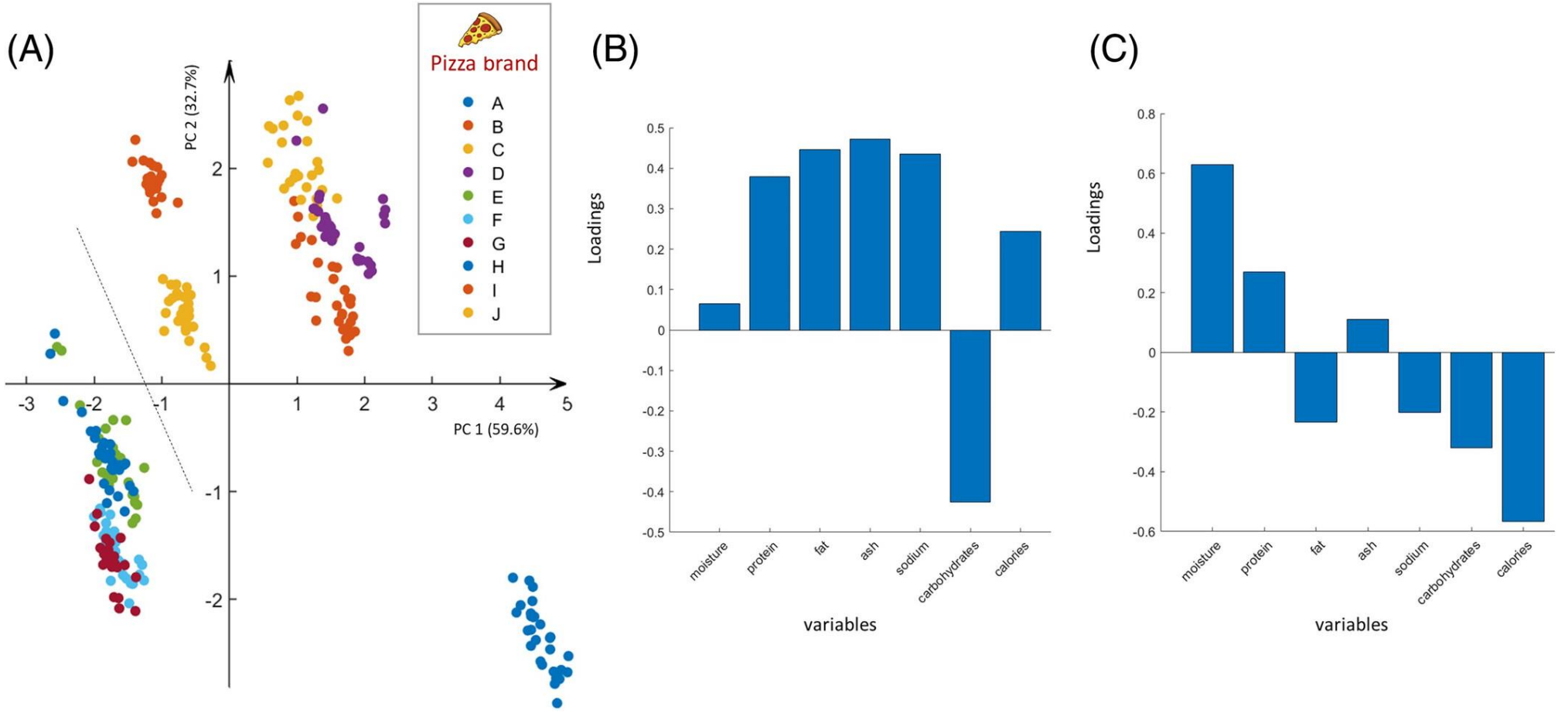
1. linear combination of data points are ordered by their variance
2. linear combination of data points are uncorrelated

Principal **C**omponent **A**nalysis (PCA)

Revisiting example, but with PCA:



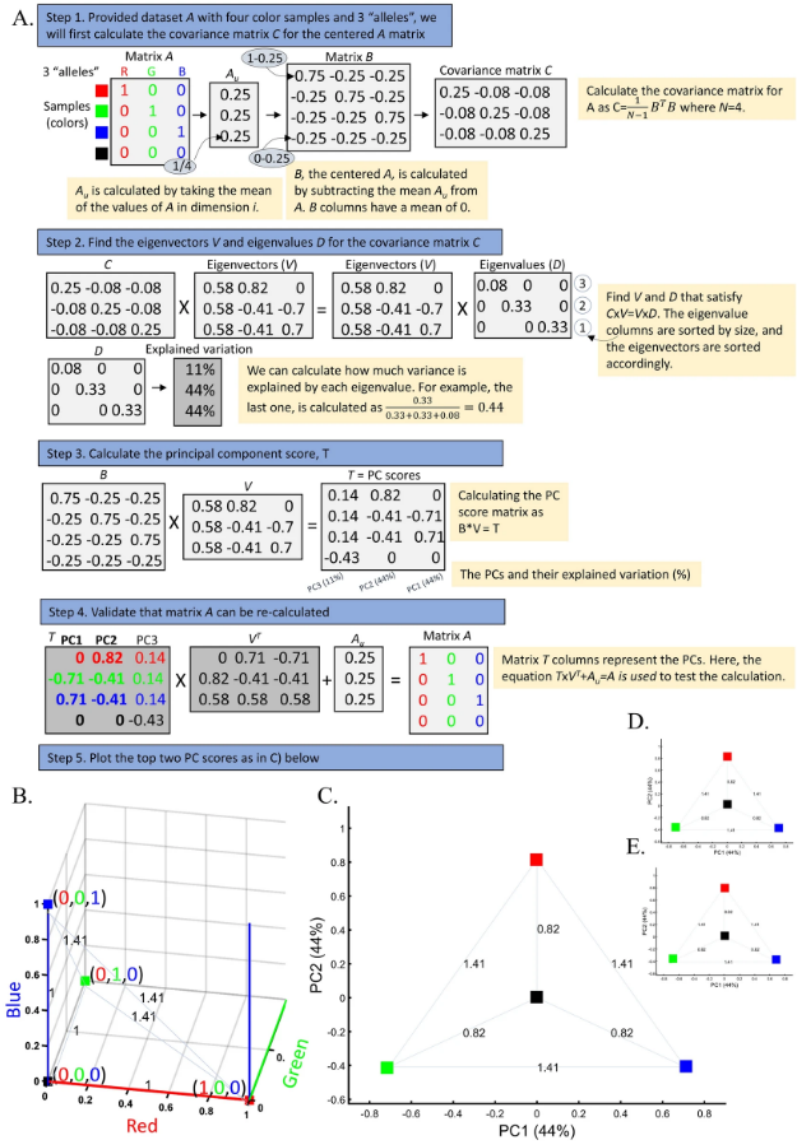
Principal **C**omponent **A**nalysis (PCA)



Elhaik E. doi: 10.1038/s41598-022-14395-4. PMID: 36038559; PMCID: PMC9424212.

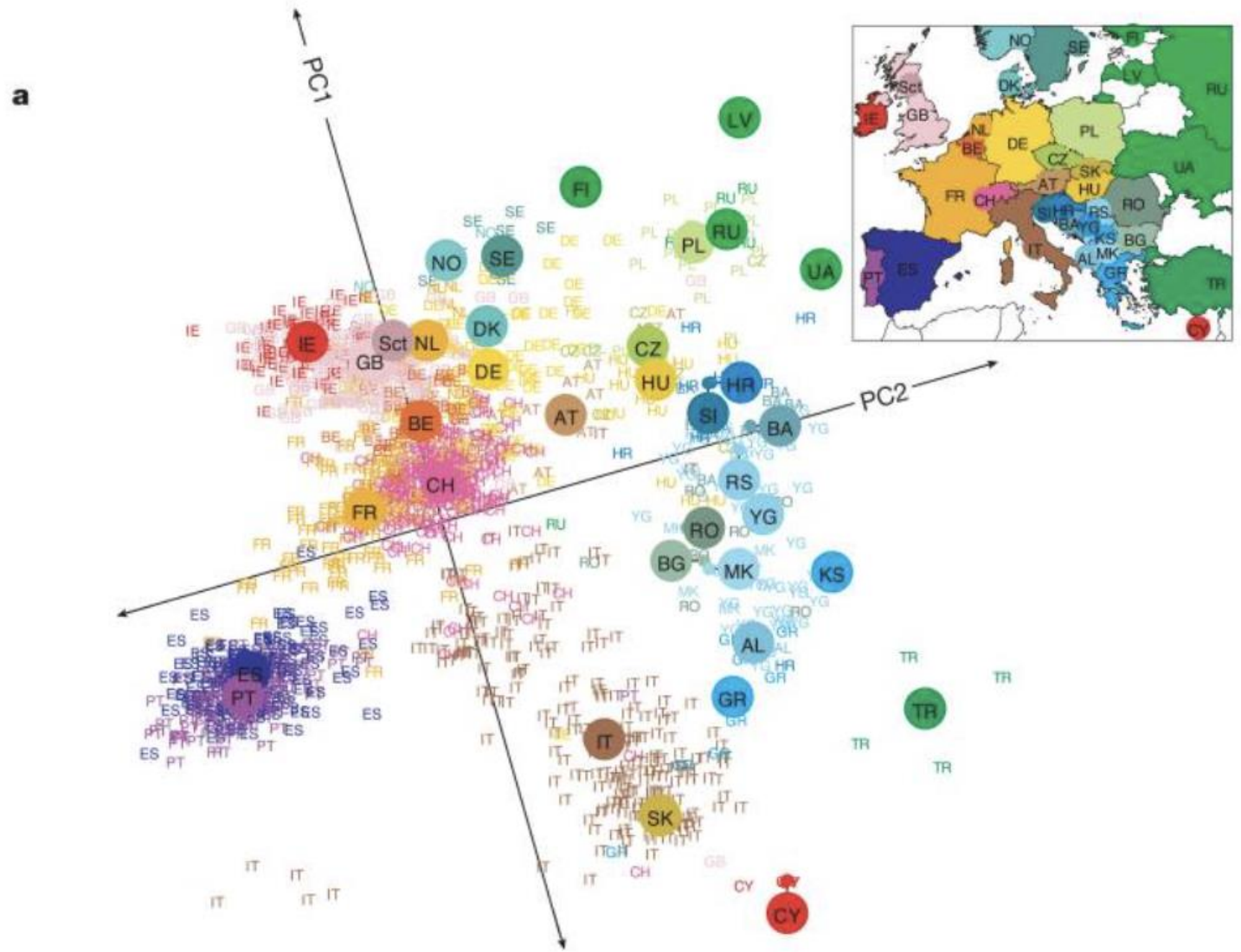
Figure 1

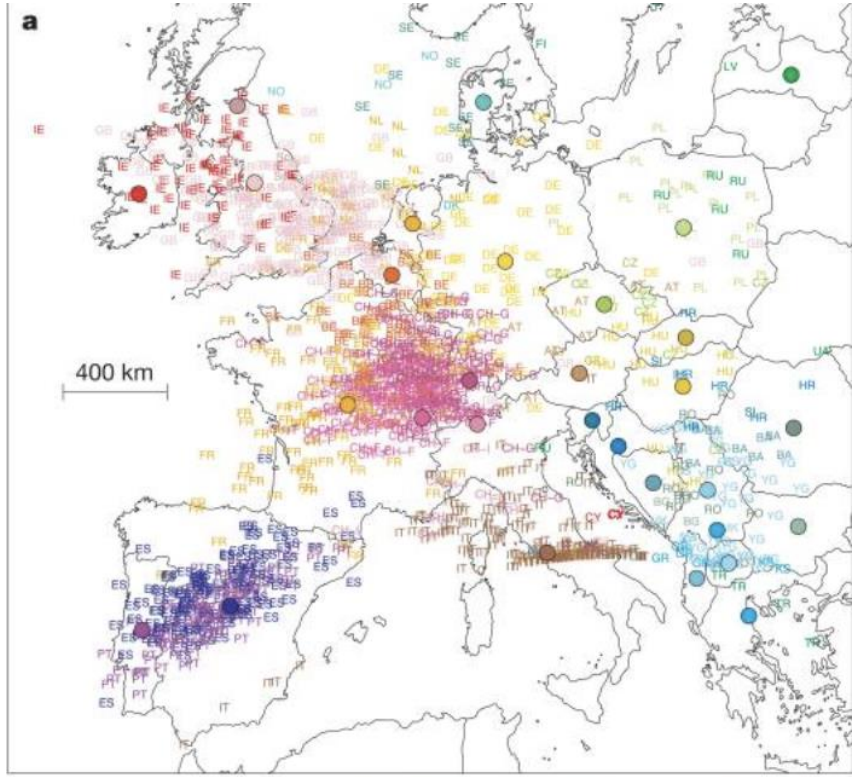
From: Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated

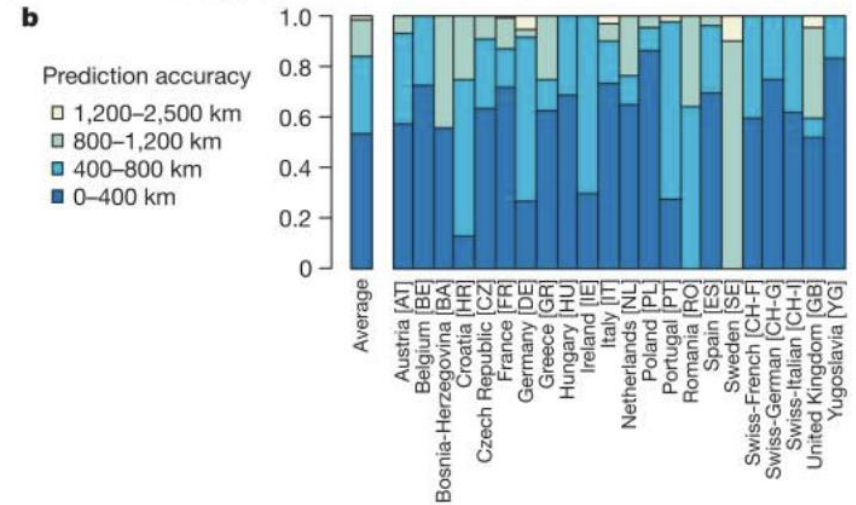


Applying PCA to four color populations. (**A**) An illustration of the PCA procedure (using the singular value decomposition (SVD) approach) applied to a color dataset consisting of four colors (n_{All}=1). (**B**) A 3D plot of the original color dataset with the axes representing the primary colors, each color is represented by three numbers ("SNPs"). After PCA is applied to this dataset, the projections of color samples or populations (in their original color) are plotted along their first two eigenvectors (or principal components [PCs]) with (**C**) n_{All}=1, (**D**) n_{All}=100, and (**E**) n_{All}=10,000. The latter two results are identical to those of (**C**). Grey lines and labels mark the Euclidean distances between the color populations calculated across all three PCs.

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000682







Novembre et al (2008). Genes mirror geography within Europe.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/

- By removing the redundant variables, PCA captures which variables are actually foundational and are responsible for the phenomenon that we are measuring
- Bumpus Dataset.

Figure 11.1: This graph shows an "elbow" which gives a graphic reminder of how many variables you need to include to account for most of the variation in your data