



# Principle Component Analysis

- Reducing the complexity of data

Simon Debes, Andreas M. Jørgensen, Jorge Prado and Peter Andresen



## High dimensional data can be difficult to work with and visualize

- In some fields, 10s or 100s of parameters are measure for each object
- Correlated or abundant variables
- Expensive computations
- Impossible to visualize



# Karl Pearson introduced a way of simplifying things in 1901

Not exactly a new method.

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, F.R.S., University College, London\*.

(1) **I**N many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the “best-fitting” straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1x, \quad \text{or} \quad z = a_0 + a_1x + b_1y,$$

$$\text{or} \quad z = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n,$$

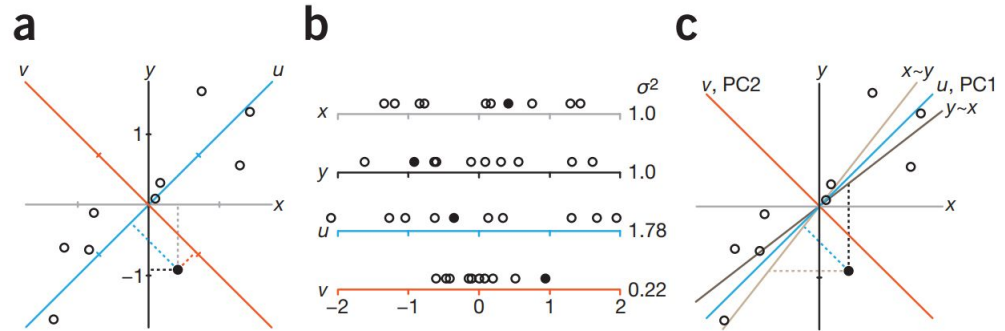
where  $y, x, z, x_1, x_2, \dots, x_n$  are variables, and determining the “best” values for the constants  $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \dots, a_n$

[2] Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. <https://doi.org/10.1080/14786440109462720>

# Principal component analysis simplifies the data by reducing the dimensionality

Data is projected onto lines or hyperplanes that are uncorrelated and maximise the variance of the data projections.

Not the same as a linear regression

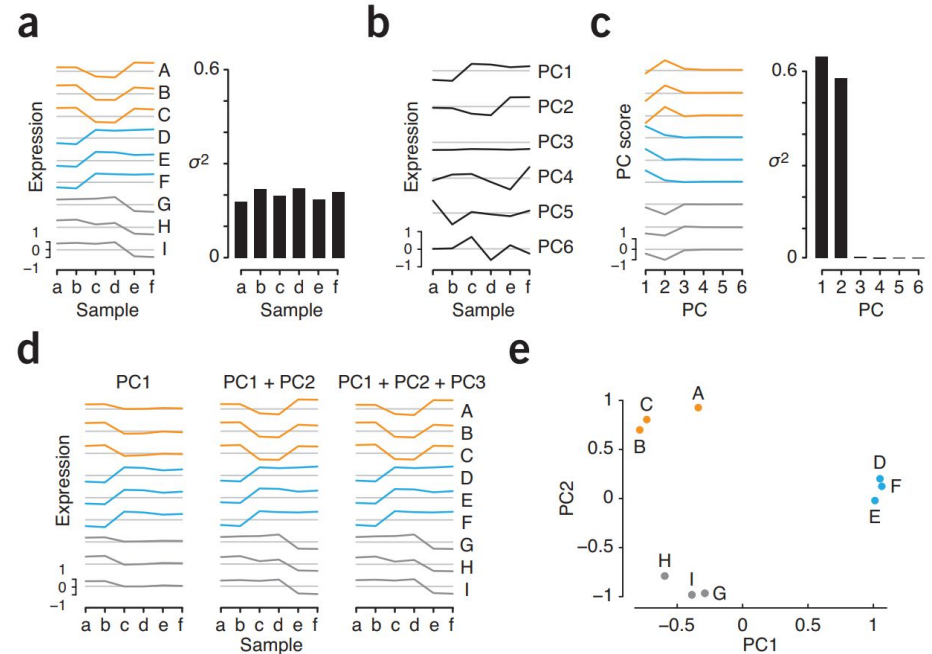


[1] Lever, J., Krzywinski, M. Altman, N. Principal component analysis. Nat Methods 14, 641–642 (2017). <https://doi.org/10.1038/nmeth.4346>

# This helps extract the important features and group data points

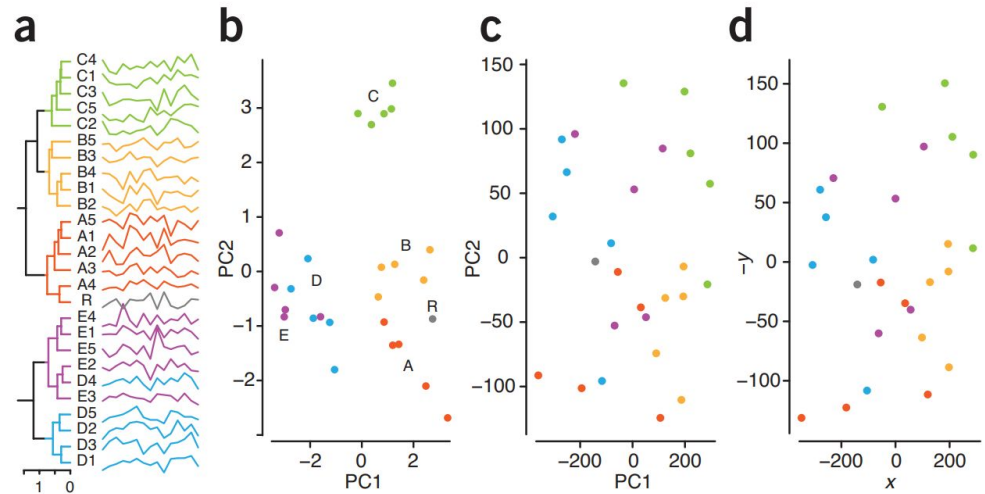
Often the important information can be summarized using only a few principal components, allowing for simple visualization and categorizing of the data.

However, this is not always the case, and one should look at the variance of the subsequent PC's to determine how many is required



# It is important to normalize the data, since PCA depends on the scale of the variables

Data should be scaled, since the PC's are influenced more by high valued variables

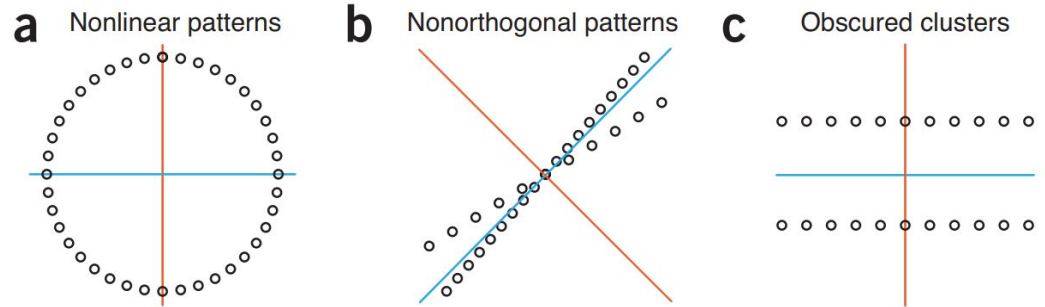


[1] Lever, J., Krzywinski, M. Altman, N. Principal component analysis. Nat Methods 14, 641–642 (2017). <https://doi.org/10.1038/nmeth.4346>


## The PCA method also has drawbacks, since it relies on linear patterns

The method might miss nonlinear and non-orthogonal patterns

It is also not guaranteed to capture all clusters



[1] Lever, J., Krzywinski, M. Altman, N. Principal component analysis. Nat Methods 14, 641–642 (2017). <https://doi.org/10.1038/nmeth.4346>



## Therefore more advanced methods has been developed to try to capture these effects

- Nonlinear PCA
  - Able to detect more complex patterns
- Sparse PCA
  - Reduce number of input variables included
  - Reduce computing time
- Robust PCA
  - Remove outliers
  - Assigning different weights to data



