

PCA: Overview

Introduction to Statistical Modelling

Prof. Joris Vankerschaver

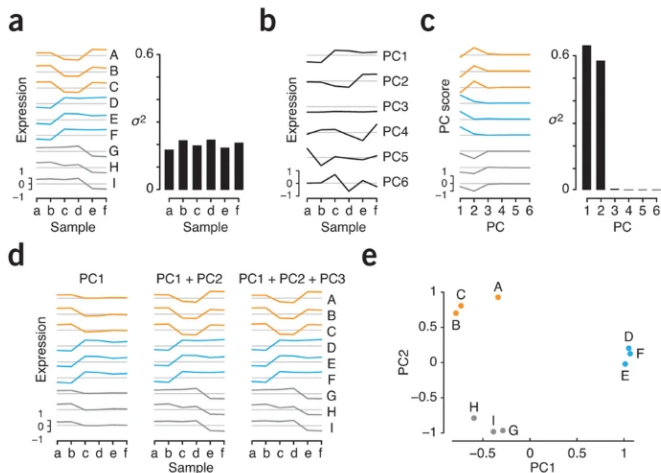
What is dimensionality reduction?

What and why

- Reduce the number of variables (“dimensionality”) in a dataset **in a principled way**.
- Useful for
 - Visualization
 - Data preprocessing
 - Computational efficiency
- Many different approaches
 - Principal component analysis (this course)
 - Multidimensional scaling
 - t-SNE, UMAP, ...

Visualization

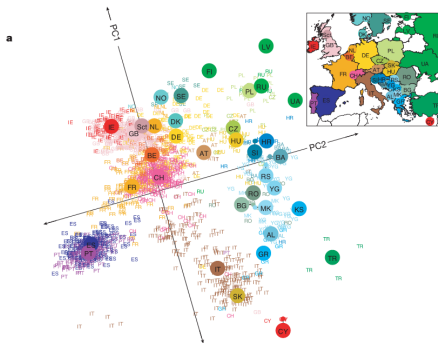
Figure 2: PCA reduction of nine expression profiles from six to two dimensions.



From: Lever et al., *Principal component analysis*, Nature Methods, Vol. 14, p. 641–642, 2017.

Visualization

Genotype data 197,146 loci in 1387 Europeans, summarized in two principal components (left) and compared to geographical origin (right).

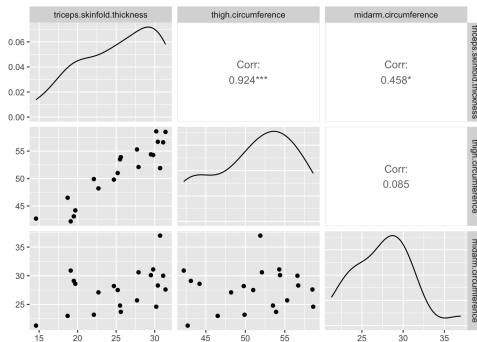


From: Novembre et al., *Genes mirror geography within Europe*, Nature, Vol. 456, 6 November 2008.

Data preprocessing

Bodyfat dataset:

- Suffered from high multicollinearity.
- Conclusions from regression model are doubtful.



Computational efficiency

- A 250×250 image consists of $250^2 = 62,500$ pixels.
- Not all pixels are equally informative.
- Extract signal that is maximally informative, discard rest.



Principal component analysis

- Covered in this course.
- Works by finding directions in which **variance is maximized**.
- Good first choice, not so good if patterns are highly nonlinear.

