NBE-4070 : Basics of Biomedical Data Analysis

Stéphane Deny
Prof. in Neuroscience and Biomedial Engineering and Computer Science
Aalto University

Lecture 5: Pearson Correlation, PCA and SVD

Quiz 4

Question 1

Edit question | Flag question | Marked out of 1.00 | Not yet answered

What is a correct definition of clustering?

- ☐ a. Clustering is grouping a collection of data points into subsets, such that the points within each cluster are further apart from one another than points assigned to different clusters.
- ☐ b. Clustering is grouping a collection of data points into subsets, such that the points within each cluster are closer to one another than points assigned to different clusters.

Question 2

What does 'K' stand for in K-means?

- ☐ a. 'K' specifies the number of clusters that the algorithm will find.
- ☐ b. 'K' specifies the number of iterations that the algorithm will go through before reaching convergence.

Question 3

Edit question | Flag question | Marked out of 1.00 | Not yet answered

How do clustering algorithms typically work on large datasets?

- ☐ a. They systematically explore all possible partitions of the data and select the one partition that optimizes the value of the criterion.
- ☐ b. They rely on iterative greedy descent: (1) an initial partition is specified; (2) at each iterative step, the cluster assignments are changed in such a way that the value of the criterion is improved from its previous value.

Question 4

Edit questic

How does hierarchical clustering work?

- ☐ a. Hierarchical clustering recursively merges the smallest cluster of points with the one that is closest to it.
- ☐ b. Hierarchical clustering recursively merges a selected pair of clusters into a single cluster. The pair chosen for merging consists of the two closest clusters.

Question 5

Edit question | Flag question | Marked out of 1.00 | Not yet answered

What are some risks when interpreting the results of cluster analysis?

Select one or more:

- ☐ a. Cluster analysis is far too risky and it is better never to use it.
- ☐ b. Clustering algorithms will typically always find clusters, whether they truly exist in the data or not. A visual inspection of the clusters might be helpful to decide whether the clusters found are truly homogeneous and distinct from one another.
- ☐ c. Classical clustering methods can be inadapted to the structure of the data. For example, if different groups of points are arranged in concentric circles, K-means clustering will fail to reveal this structure.

Question 6

Edit question | Flag que

What is a dendogram?

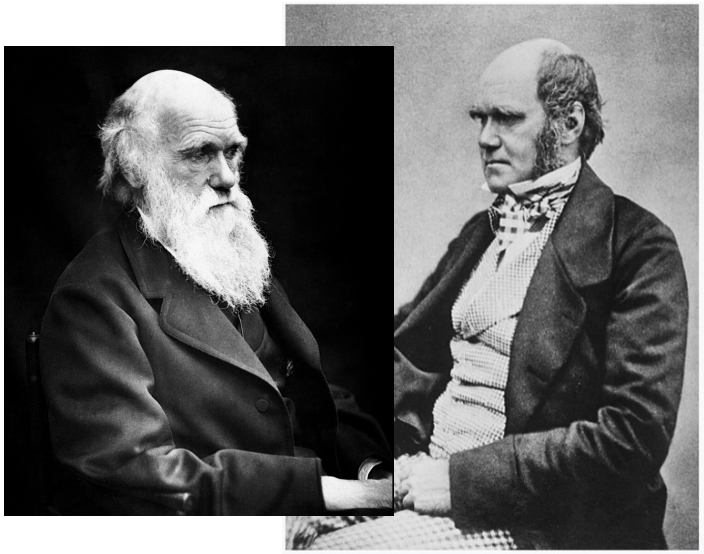
- ☐ a. It is binary tree representing the hierarchical clusters, such that the width of each node is proportional to the value of the intergroup dissimilarity between its two daughters.
- ☐ b. It is binary tree representing the hierarchical clusters, such that the height of each node is proportional to the value of the intergroup dissimilarity between its two daughters.

Outline of the course

1. Mean, Standard Deviation, Standard Error, Confidence Intervals, T-test
2. Fourier Transform, Wavelet Transforms, Spectrograms, High-pass, Low-pass filters
3. Covariance and Principal Component Analysis (PCA)
4. Clustering Methods
5. Pearson Correlation, PCA and SVD
6. Linear Regression / Logistic Regression
7. Non-linear Methods: Independant Component Analysis, t-Stochastic Neighbour Embedding, Random Forests, Deep Networks
8. Invited lectures from the biomedical industry
9. Solving oral exam problems of last year in class

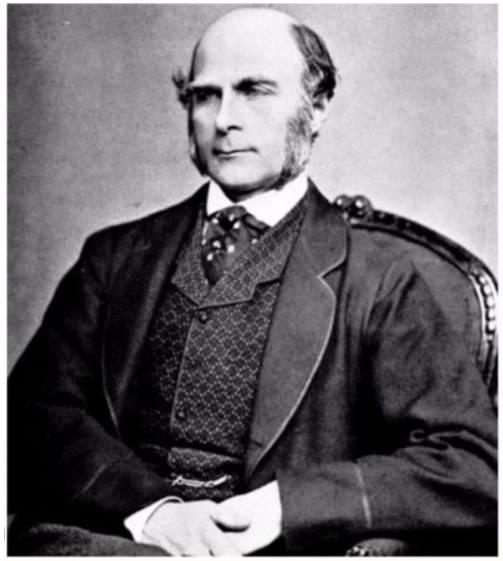
Do you know these persons?

Charles Darwin



1809-1982

Francis Galton



1822-1911

“the one grandson [of Erasmus Darwin], Charles Darwin, collected the facts which had to be dealt with and linked them together by wide-reaching hypotheses; the other grandson, Francis Galton, provided the methods by which they could be tested...”
— Karl Pearson

The troubled history of statistics

Francis Galton

From Wikipedia, the free encyclopedia

"Galton" redirects here. For other uses, see Galton (disambiguation).

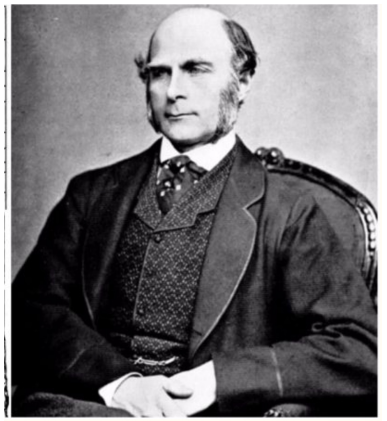
Sir Francis Galton, FRS FRAI (/ˈgɔːltən/; 16 February 1822 – 17 January 1911), was an English Victorian era polymath: a statistician, sociologist, psychologist,[1] anthropologist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician and a proponent of social Darwinism, eugenics, and scientific racism. He was knighted in 1909.

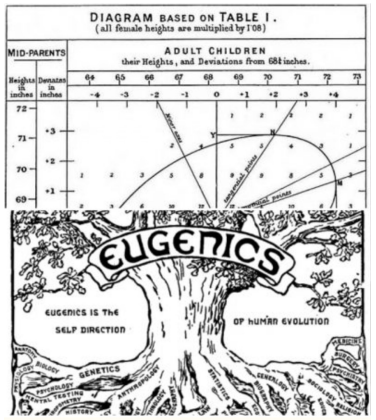
Galton produced over 340 papers and books. He also created the statistical concept of correlation and widely promoted regression toward the mean. He was the first to apply statistical methods to the study of human differences and inheritance of intelligence, and introduced the use of questionnaires and surveys for collecting data on human communities, which he needed for genealogical and biographical works and for his anthropometric studies. He was a pioneer of eugenics, coining the term itself in 1883, and also coined the phrase "nature versus nurture".[2] His book *Hereditary Genius* (1869) was the first social scientific attempt to study genius and greatness.[3]

As an investigator of the human mind, he founded psychometrics (the science of measuring mental faculties) and differential psychology, as well as the lexical hypothesis of personality. He devised a method for classifying fingerprints that proved useful in forensic science. He also conducted research on the power of prayer, concluding it had none due to its null effects on the longevity of those prayed for.[4] His quest for the scientific principles of diverse phenomena extended even to the optimal method for making tea.[5]

As the initiator of scientific meteorology, he devised the first weather map, proposed a theory of anticyclones, and was the first to establish a complete record of short-term climatic phenomena on a European scale.[6] He also invented the Galton Whistle for testing differential hearing ability.[7] He was Charles Darwin's half-cousin.[8]

Contents [hide]





The troubled history of statistics

Karl Pearson

From Wikipedia, the free encyclopedia

For the English cricketer, see Karl Pearson (cricketer).

Karl Pearson FRS FRSE[1] (/ˈpɪərsən/; born **Carl Pearson**; 27 March 1857 – 27 April 1936[2]) was an English mathematician and biostatistician. He has been credited with establishing the discipline of mathematical statistics.[3][4] He founded the world's first university statistics department at University College, London in 1911, and contributed significantly to the field of biometrics and meteorology. Pearson was also a proponent of social Darwinism, eugenics and scientific racism. Pearson was a protégé and biographer of Sir Francis Galton. He edited and completed both William Kingdon Clifford's *Common Sense of the Exact Sciences* (1885) and Isaac Todhunter's *History of the Theory of Elasticity*, Vol. 1 (1886–1893) and Vol. 2 (1893), following their deaths.

Known for

- Principal component analysis
- Pearson distribution
- Pearson's chi-squared test
- Pearson's r
- Phi coefficient
- Chi-square distribution
- Contingency table
- Histogram
- Kurtosis
- Mode
- Random walk
- *The Grammar of Science*

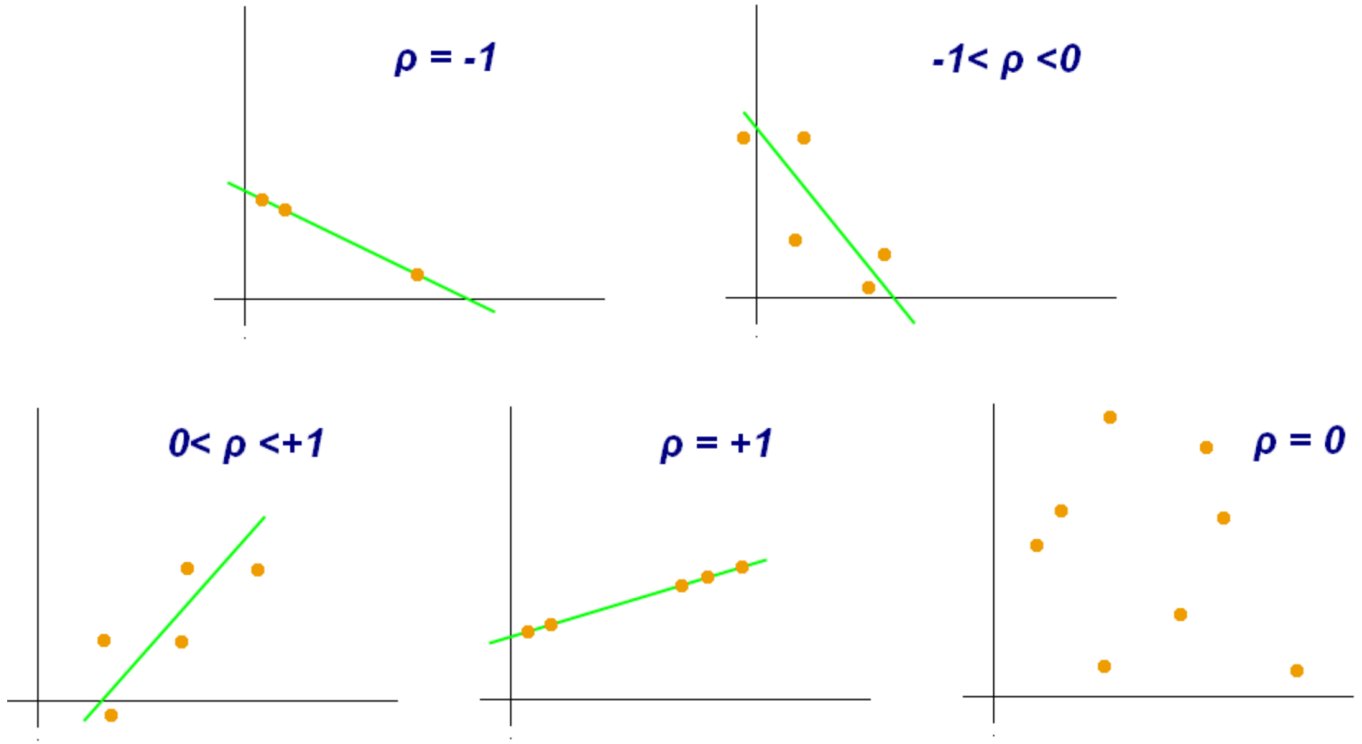
Definition: Pearson correlation

- Given two paired variables $X = \{x_1, x_2, ..., x_N\}$
  $Y = \{y_1, y_2, ..., y_N\}$
- Their *(Pearson) correlation coefficient* is given by

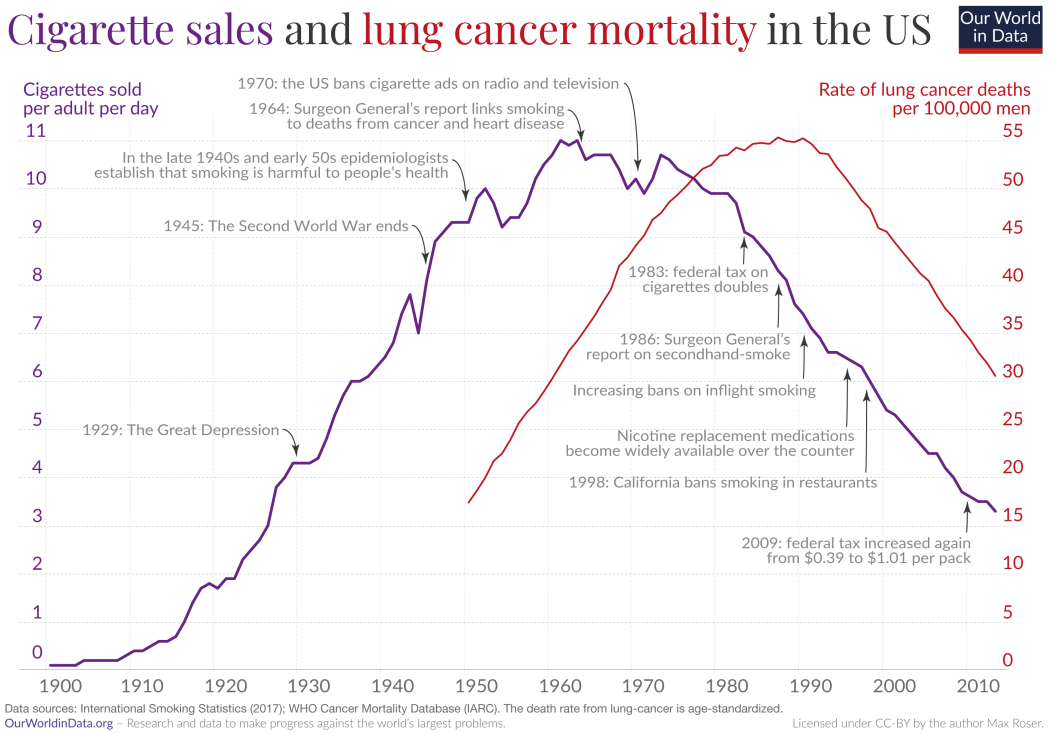
$$R = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{(\sum_{i=1}^{N}(x_i - \mu_X)^2)}\sqrt{(\sum_{i=1}^{N}(y_i - \mu_Y)^2)}}$$

- It is a normalized version of the covariance, such that the result always has a value between −1 and 1. There exists tests to decide whether the correlation is statistically significant or not.

Examples of Pearson correlation coefficients

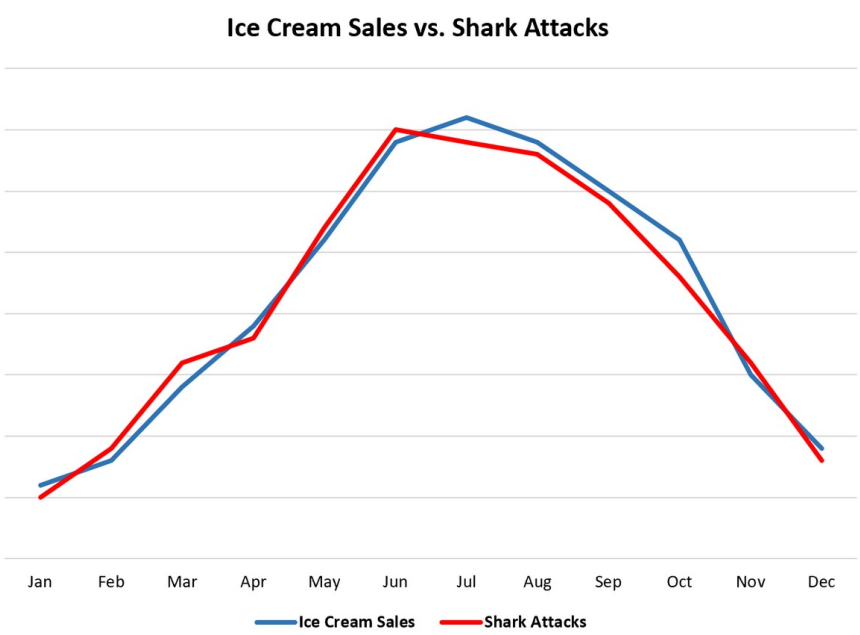


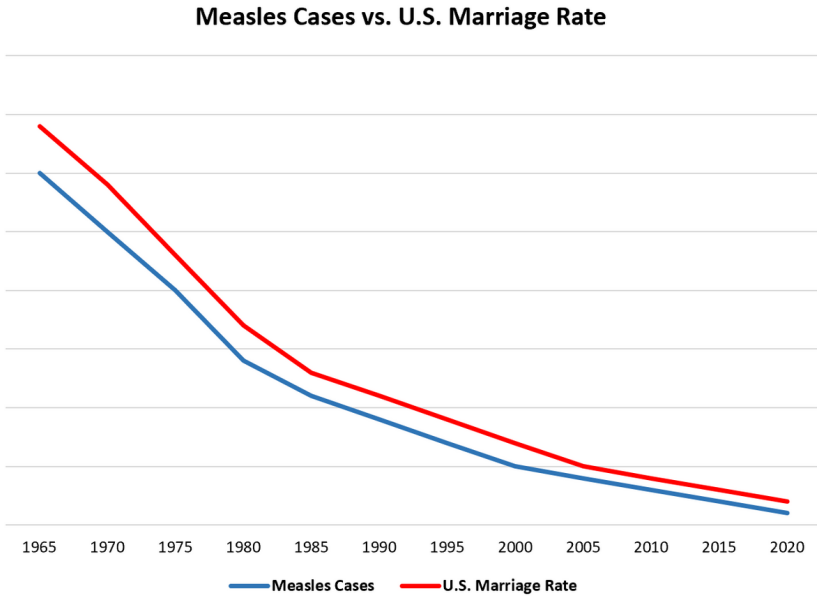
Correlations can be useful to observe...



https://ourworldindata.org/smoking-big-problem-in-brief

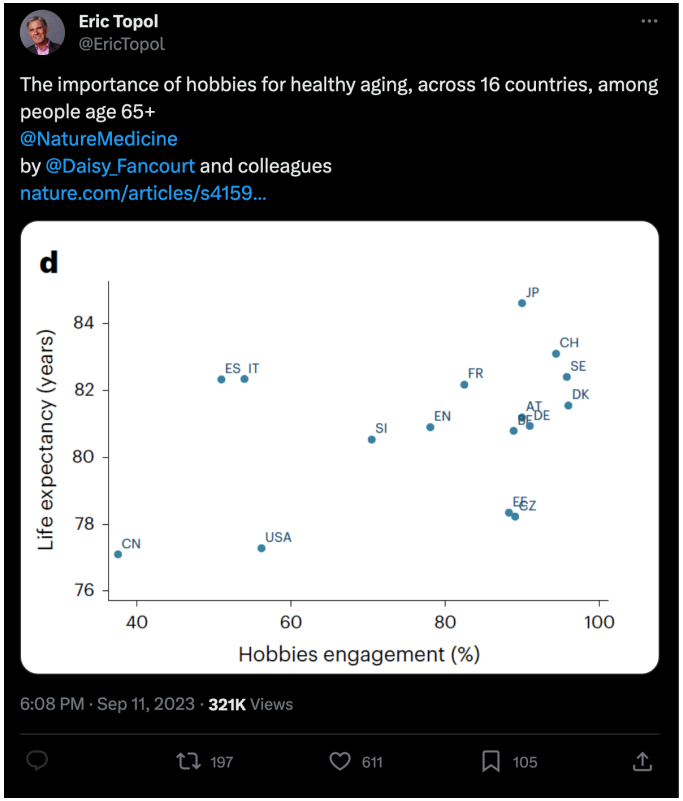
... but correlation ≠ causation

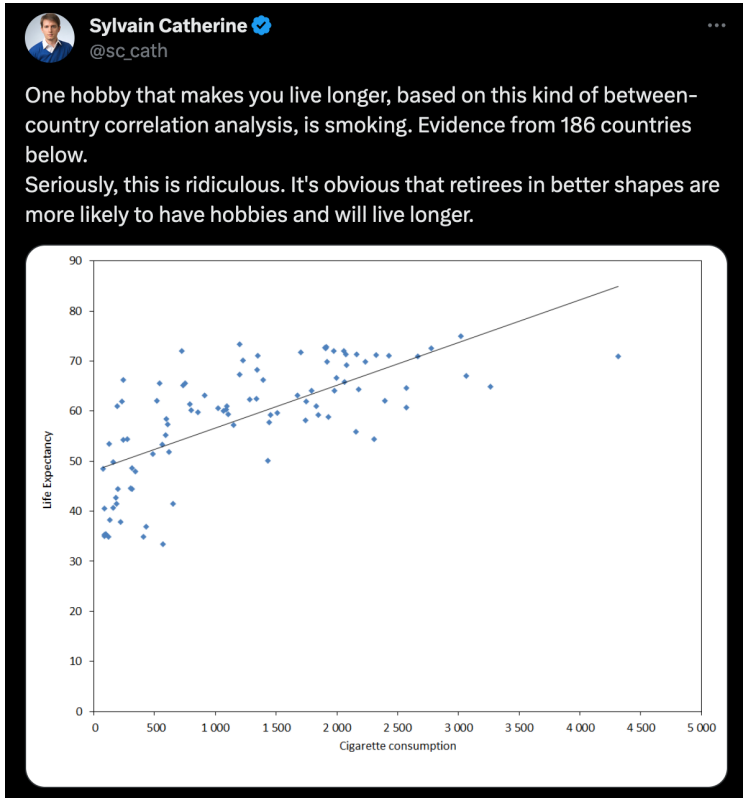




source: https://www.statology.org/correlation-does-not-imply-causation-examples/

… but correlation ≠ causation



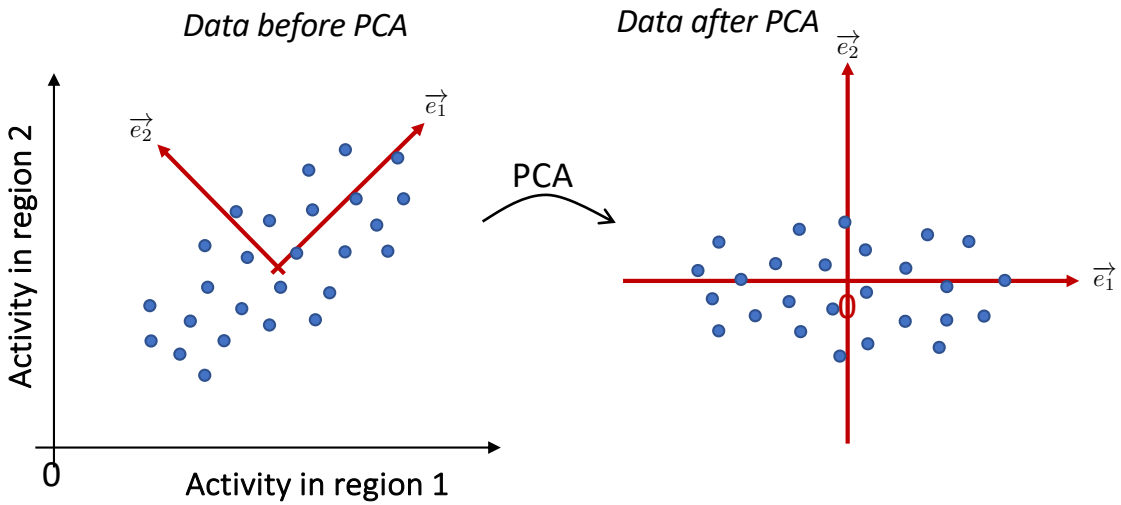


https://twitter.com/sc_cath/status/1701275606296519002
(Sylvain Catherine, Economist. Assistant Professor of Finance at Wharton)

Reminder about PCA

PCA finds an orthonormal basis ‘E’ such that ‘D’, the covariance of ‘X’, is diagonal in that basis:

$$Cov(X) = E^t D E$$



Subtle question: in PCA, can we replace the covariance matrix with the correlation matrix of the data?

1. Covariance matrix of X:

$$X_\mu = X - \overrightarrow{x_\mu} \qquad Cov(X) = X_\mu X_\mu^t$$

2. Correlation matrix of X:

$$\widetilde{X} = \frac{X - \overrightarrow{x_\mu}}{\overrightarrow{\sigma}} \qquad Corr(X) = \widetilde{X}\widetilde{X}^t$$

- Yes! If the scale of individual dimensions is not meaningful, it can be a good idea to z-score the data before doing PCA

Now, we will see 3 case studies of PCA

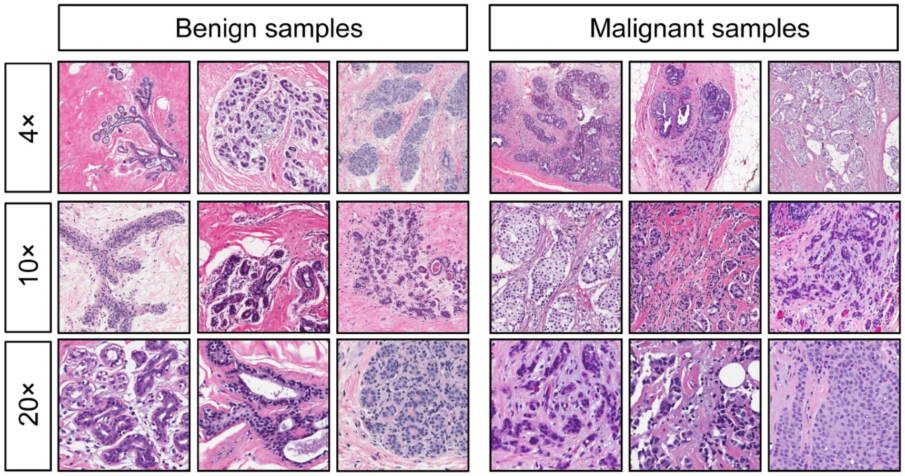
1. Visualizing a breast cancer data set
   source: https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff

2. Denoising an electrocardiogram
   source: https://medium.com/@andrewtan_36013/principal-components-of-electrocardiograms-14874b3a96b1

3. Modelling the dynamics of C. elegans
   source: Stephens at al. 2008, PloS Computational Biology

1. Visualizing a breast cancer dataset

Breast Cancer Wisconsin Diagnostic Dataset



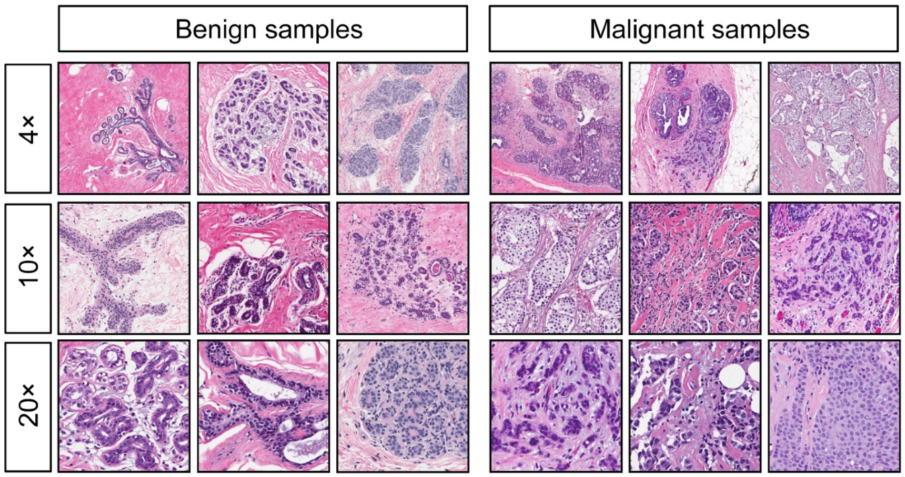
Description:

- Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

- 569 samples including their diagnostic

1. Visualizing a breast cancer dataset

Breast Cancer Wisconsin Diagnostic Dataset



Description:

- Features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image.

- 569 samples including their diagnostic

Features:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

*The **mean**, **standard error**, and "**worst**" or largest (mean of the three largest values) of these features were computed for each image, resulting in **30 features***

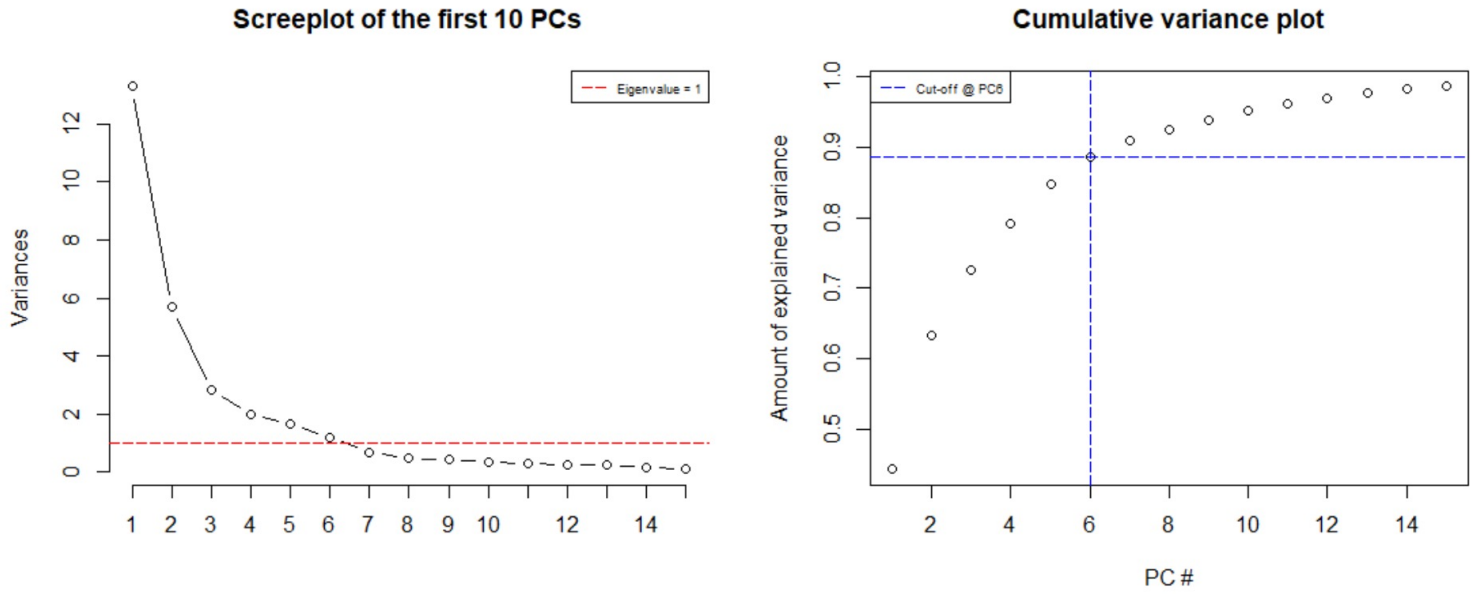
1. Visualizing a breast cancer dataset

Applying PCA:

- **Standardize the data** (Center and scale).
- **Calculate the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix** (One could also use Singular Vector Decomposition).
- **Sort the Eigenvalues in descending order and choose the *K* largest Eigenvectors** (Where *K* is the desired number of dimensions of the new feature subspace $k \leq d$).
- **Construct the projection matrix W from the selected *K* Eigenvectors.**
- **Transform the original dataset X via W to obtain a *K*-dimensional feature subspace Y.**

1. Visualizing a breast cancer dataset

Plotting the eigenvalues:



Screeplot of the Eigenvalues of the first 15 PCs (left) & **Cumulative variance plot** (right)

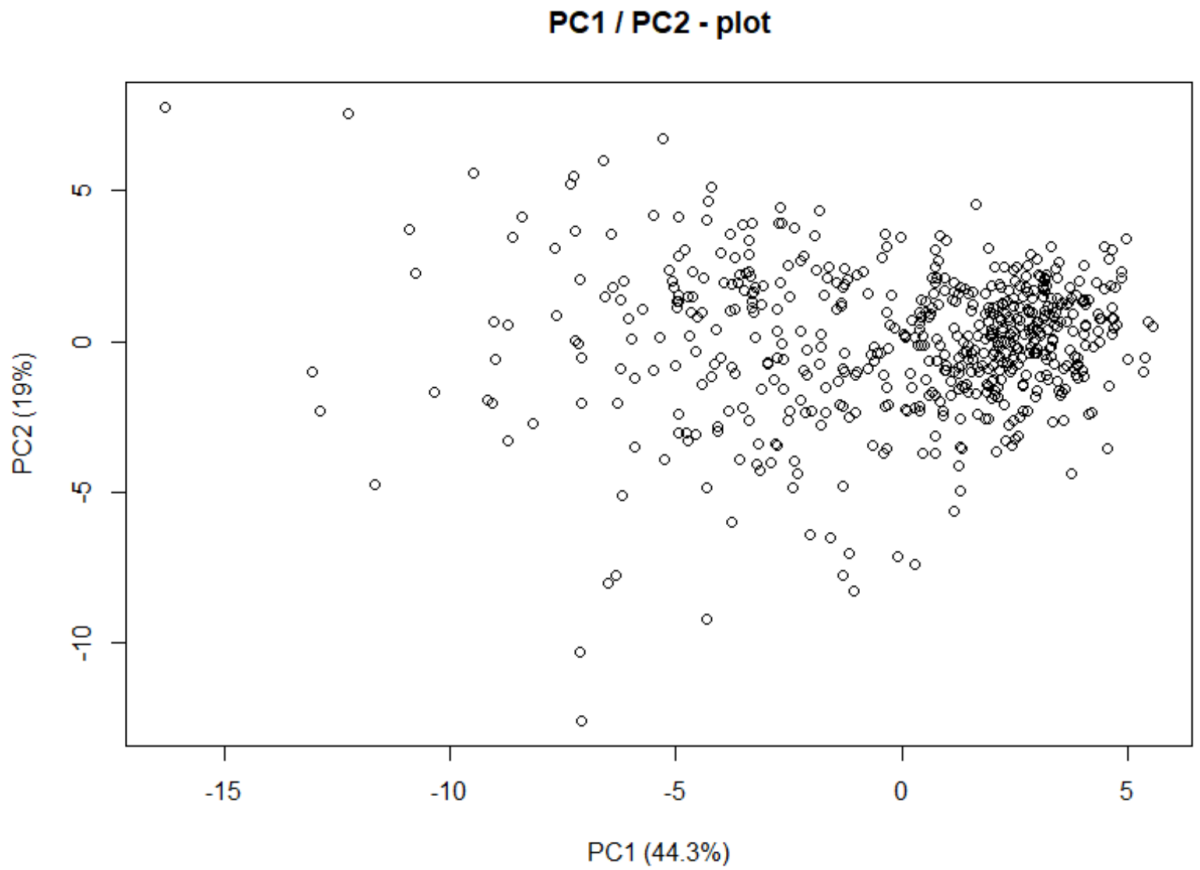
The 6 first components can explain almost 90% of the variance of the data

1. Visualizing a breast cancer dataset

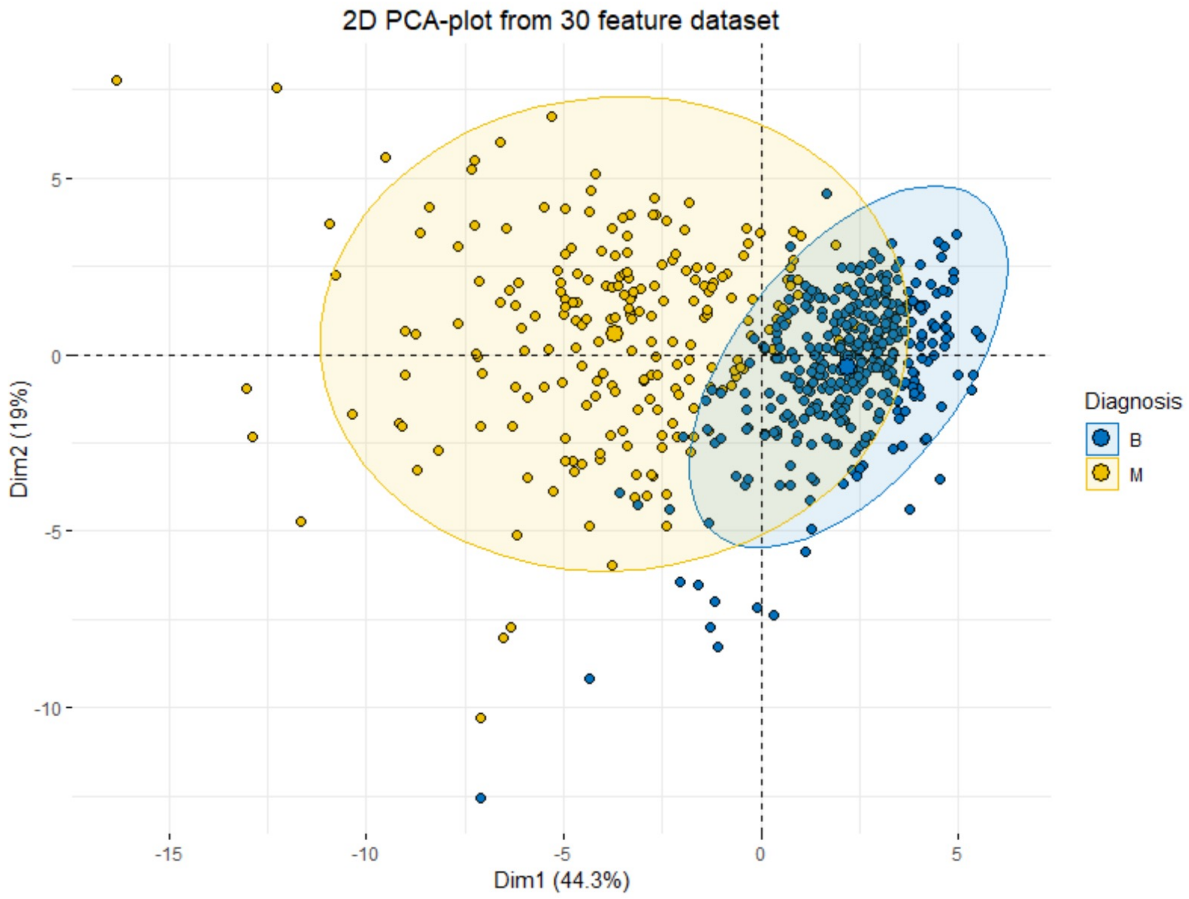
Applying PCA:

- **Standardize the data** (Center and scale).
- **Calculate the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix** (One could also use Singular Vector Decomposition).
- **Sort the Eigenvalues in descending order and choose the *K* largest Eigenvectors** (Where *K* is the desired number of dimensions of the new feature subspace k ≤ d).
- **Construct the projection matrix W from the selected *K* Eigenvectors.**
- **Transform the original dataset X via W to obtain a *K*-dimensional feature subspace Y.**

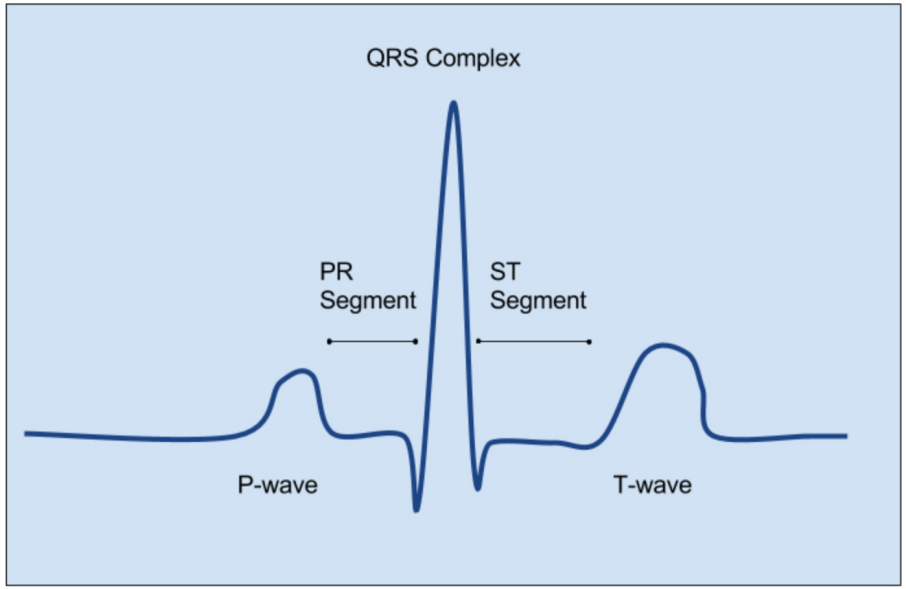
1. Visualizing a breast cancer dataset



1. Visualizing a breast cancer dataset



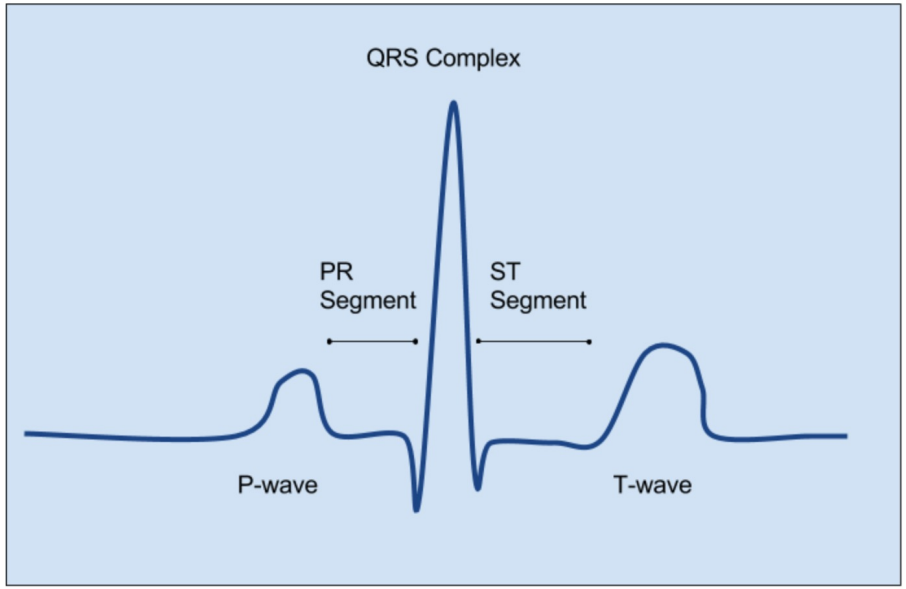
2. Denoising an electrocardiogram



Typical waveform of a heartbeat from an ECG. The vertical axis represents voltage and the horizontal axis represents time.

An electrocardiogram or ECG is a signal representation of the heart. It is a recording of the heart's electrical activity providing useful information about its overall functioning behaviour. Medical devices called Holter monitors are placed on patients to provide a continual ECG. They can be used to diagnose various cardiac diseases and some are able to sense the failure of a heart and provide an electrical jolt to defibrillate a dying patient.

2. Denoising an electrocardiogram



Typical waveform of a heartbeat from an ECG. The vertical axis represents voltage and the horizontal axis represents time.

A typical waveform can be broken down into several different waves and segments:

P-wave: The depolarization of the atria, the upper chambers of the heart. Blood now begins to flow from the atria into the ventricles, the lower chambers of the heart.

PR Segment: Represents the time delay to allow the ventricles to fill.

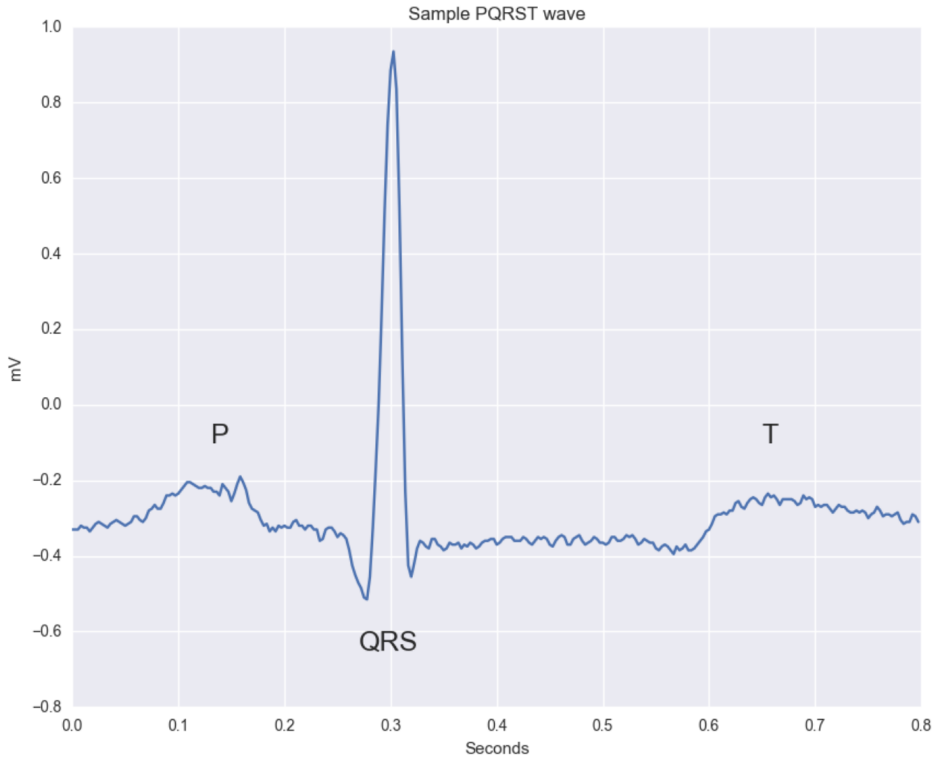
QRS Complex: The main contraction of the heart where the ventricles depolarize.

ST Segment: The time delay between depolarization and repolarization of the ventricles.

T-wave: Repolarization of the ventricles.

Abnormalities in the function of the heart can be found by closely examining the appearance of these features.

2. Denoising an electrocardiogram



Waveform from an actual ECG recording.

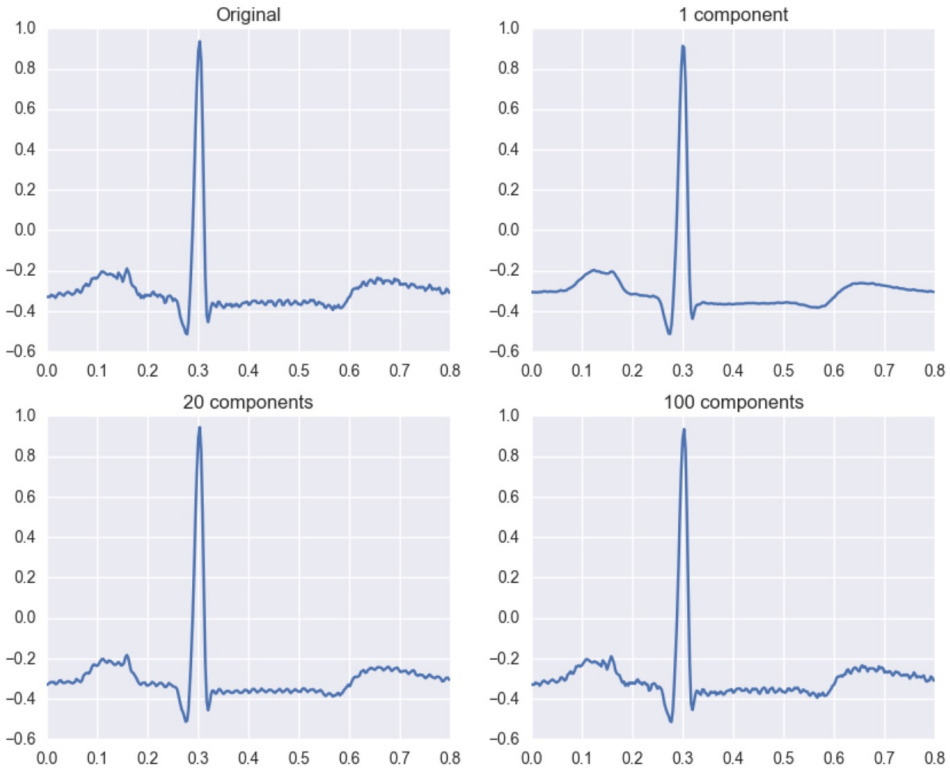
"However there are problems in practice with gathering sufficient clean ECG recordings to properly view these features. Noise filtering is a must in any ECG setup as a patient's breathing, muscle movement, perspiration and nearby transmission lines all contribute to noise in the signal. The above figure shows an actual ECG recording, albeit cleaner than most signals there are still some small artifacts present."

2. Denoising an electrocardiogram

The PCA algorithm can be summarized in the following steps:

1. Construct the covariance matrix of your n-dimensional dataset $\mathbf{X}$.
2. Find the eigen-decomposition of the covariance matrix.
3. Select k eigenvectors that correspond to the k largest eigenvalues, k will now be the new dimensionality of the transformed dataset ($k \leq n$).
4. Construct a projection matrix $\mathbf{W}$ from the top k eigenvectors.
5. Transform the n-dimensional input dataset $\mathbf{X}$ using the projection matrix $\mathbf{W}$ to obtain the new transformation.

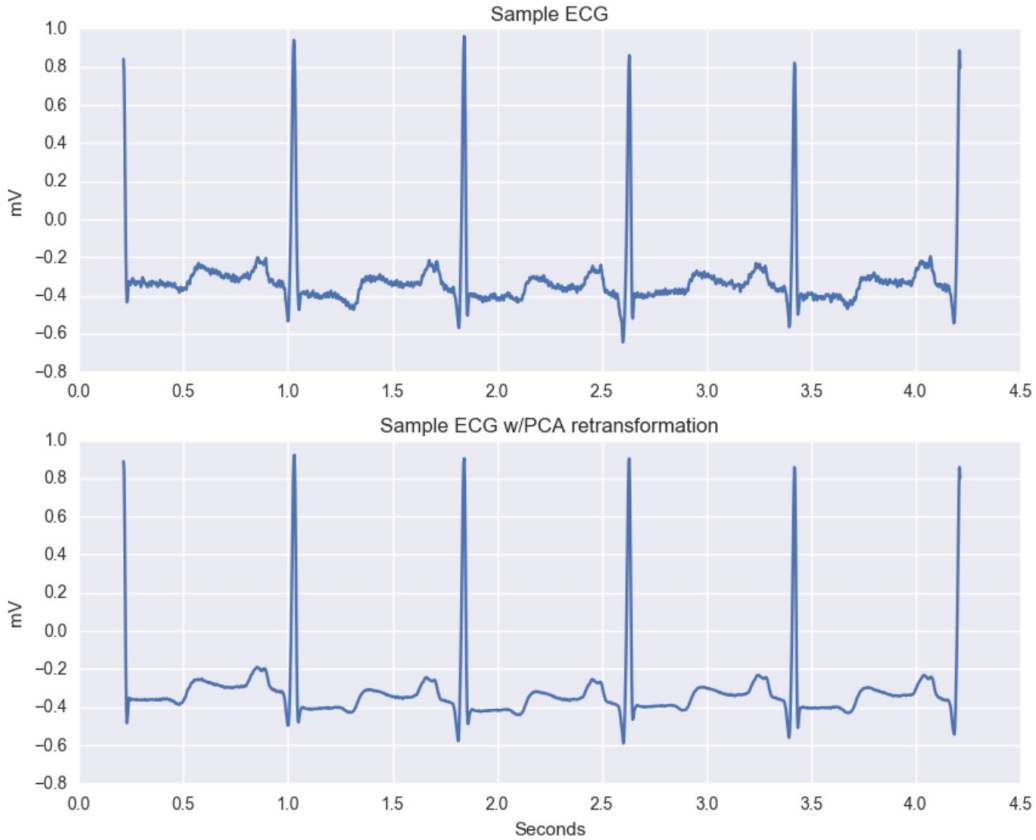
2. Denoising an electrocardiogram



Various PCA transformations using k = 1, 20 and 100.

"In the example above, I generated a dataset of PQRST waveforms from a 30-minute ECG record and ran the PCA algorithm with 1, 20 and 100 components. The example shows the original signal alongside several augmented signals where the original was reduced to a smaller number of components and then re-projected back onto the original signal space."

2. Denoising an electrocardiogram



"The final processing step is to restitch each interval together."

3. Modelling the dynamics of C. elegans

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Dimensionality and Dynamics in the Behavior of *C. elegans*

Greg J. Stephens[1,2,3]*, Bethany Johnson-Kerner[1], William Bialek[1,2], William S. Ryu[1]*

3. Modelling the dynamics of C. elegans

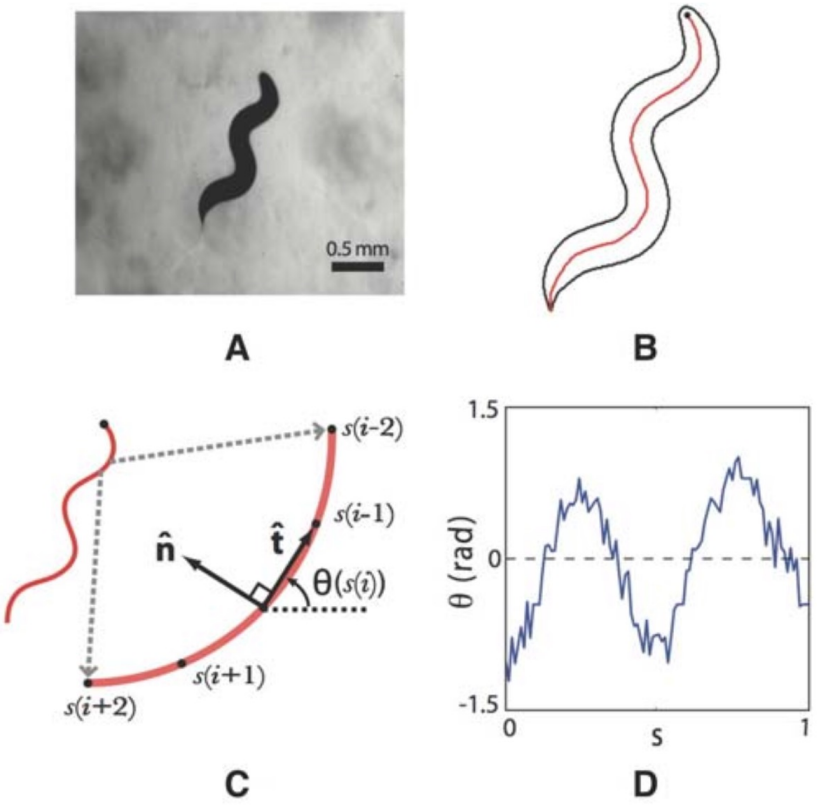


Figure 1. Describing the shapes of worms. (A) Raw image in the tracking microscope. (B) The curve through the center of the body. The black circle marks the head. (C) Distances along the curve (arclength s) are measured in normalized units, and we define the tangent $\hat{\mathbf{t}}(s)$ and normal $\hat{\mathbf{n}}(s)$ to the curve at each point. The tangent points in a direction $\theta(s)$, and variations in this angle correspond to the curvature $\kappa(s) = d\theta(s)/ds$. (D) All images are rotated so that $\langle\theta\rangle = 0$; therefore θ (s) provides a description of the worm's shape that is independent of our coordinate system, and intrinsic to the worm itself.

3. Modelling the dynamics of C. elegans

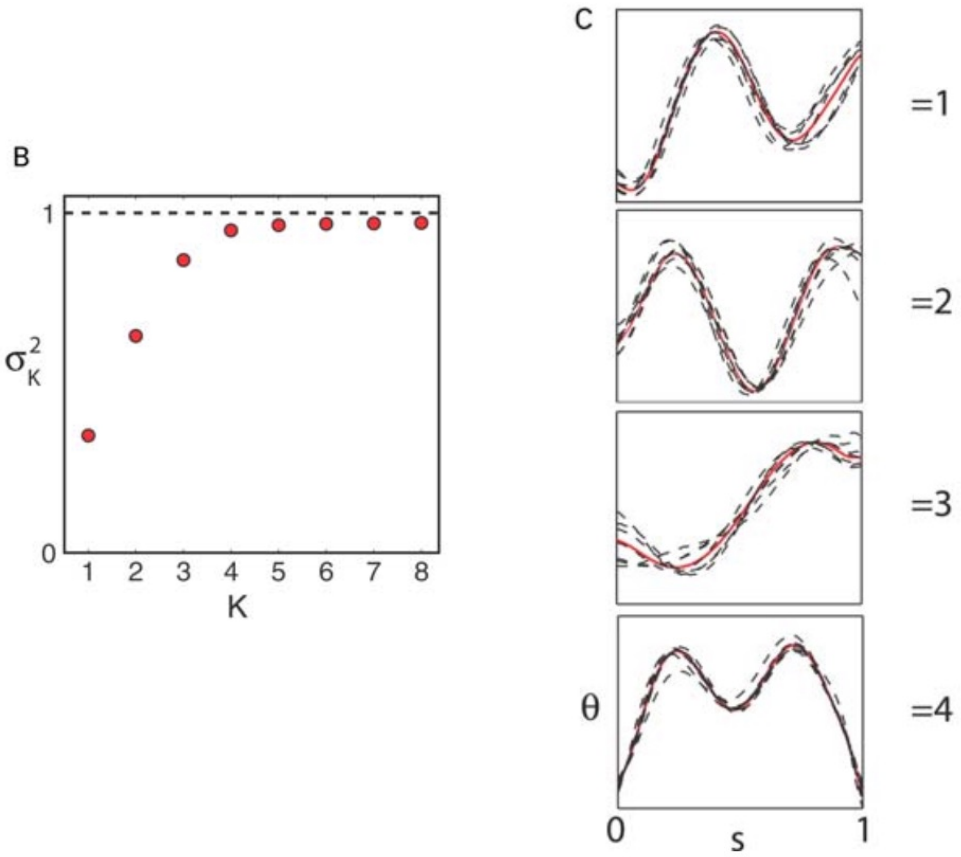


Figure 2

(B) We find the eigenvalues of and compute the fraction of the total variance captured by keeping K modes

(C) Associated with each dominant mode is an eigenvector and we refer to these as eigenworms. The population-mean eigenworms (red) are highly reproducible across individual worms (black).

3. Modelling the dynamics of C. elegans

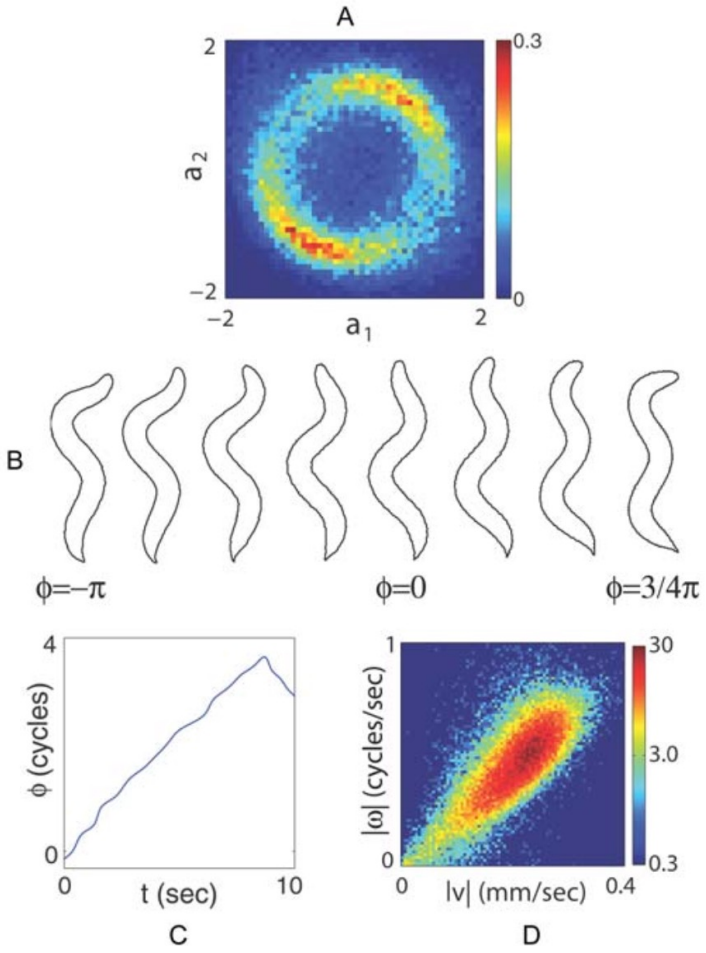


Figure 3. Motions along the first two eigenworms. (A) The joint probability density of the first two amplitudes, $\rho(a_1, a_2)$, with units such that $\langle a_1^2 \rangle = \langle a_2^2 \rangle = 1$. The ring structure suggests that these modes form an oscillator with approximately fixed amplitude and varying phase $\phi = \tan^{-1}(-a_2/a_1)$. (B) Images of worms with different values of ϕ show that variation in phase corresponds to propagating a wave of bending along the worm's body. (C) Dynamics of the phase $\phi(t)$ shows long periods of linear growth, corresponding to a steady rotation in the $\{a_1, a_2\}$ plane, with occasional, abrupt reversals. (D) The joint density $\rho(|v|,|\omega|)$. The phase velocity $\omega = d\phi/dt$ in shape space predicts worm's crawling speed.

3. Modelling the dynamics of C. elegans

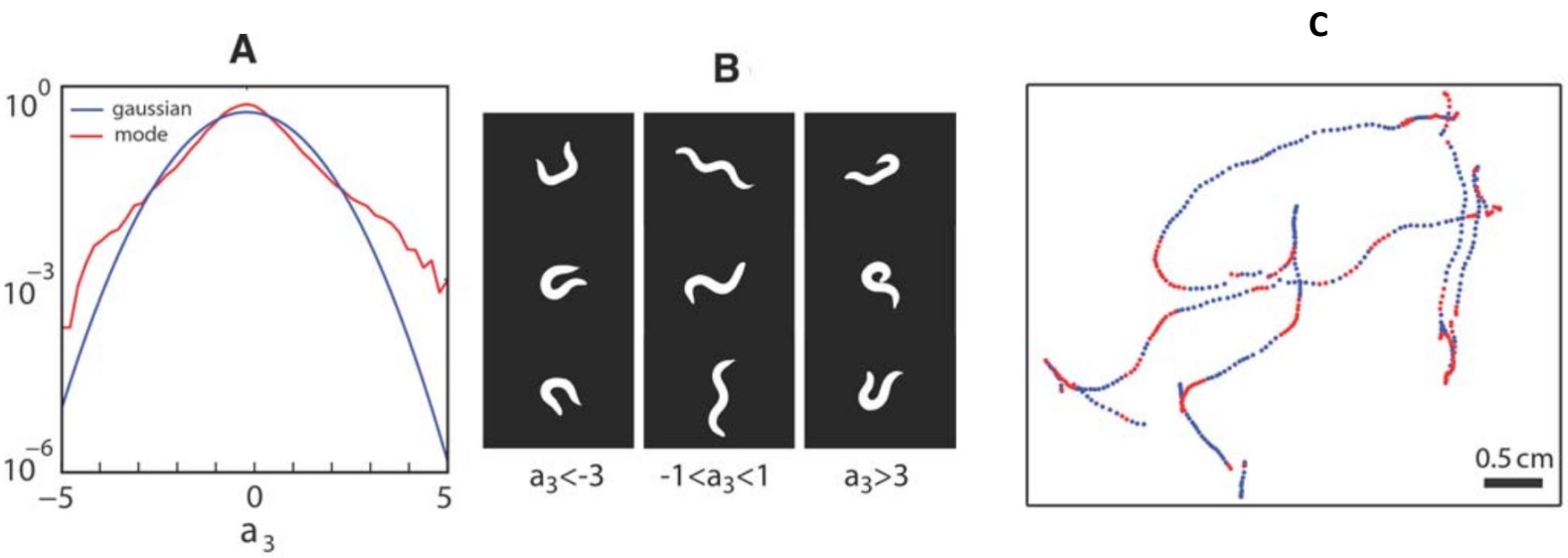


Figure 4. Motions along the third eigenworm. (A) The distribution of amplitudes $\rho(a_3)$, shown on a logarithmic scale. Units are such that $\langle a_3^2 \rangle = 1$, and for comparison we show the Gaussian distribution; note the longer tails in $\rho(a_3)$. (B) Images of worms with values of a_3 in the negative tail (left), the middle (center) and positive tail (right). Large negative and positive amplitudes of a_3 correspond to bends in the dorsal and ventral direction, respectively. (C) A two minute trajectory of the center of mass sampled at 4 Hz. Periods where $|a_3|>1$ are colored red, illustrating the association between turning and large displacements along this mode.

3. Modelling the dynamics of C. elegans

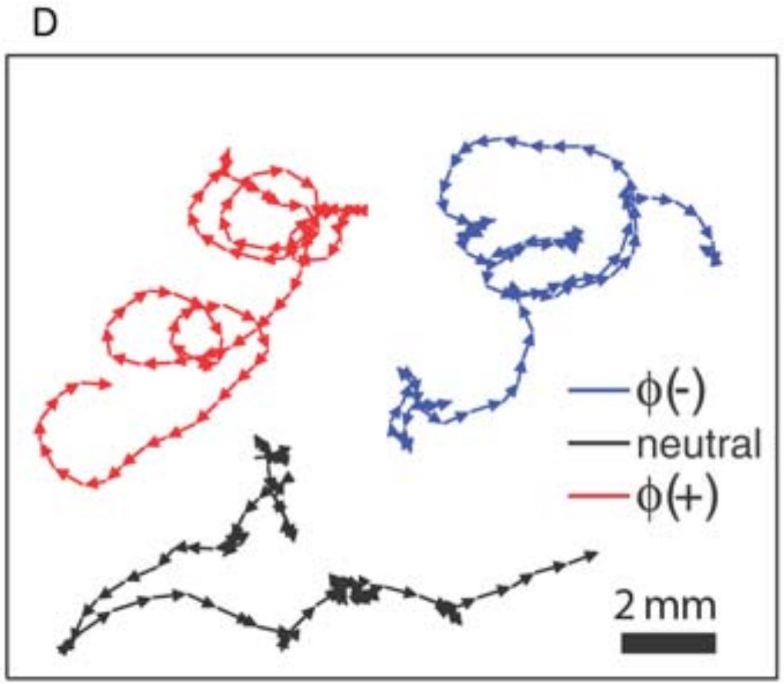


Figure 5
(D) Worm ‘‘steering.’’ A thermal impulse conditioned on the instantaneous phase was delivered automatically and repeatedly, causing an orientation change in the worm’s trajectory.

PCA vs Singular Value Decomposition (SVD)

- PCA finds an orthonormal basis ‘E’ such that ‘D’, the covariance of ‘X’, is diagonal in that basis:

$$Cov(X) = E^t D E$$

- SVD directly decomposes X_μ into a product of matrices:

$$X_\mu = U S V^t$$

where ‘U, V’ are orthornormal matrices and ‘S’ is a diagonal matrix

source: https://en.wikipedia.org/wiki/Singular_value_decomposition#Based_on_the_spectral_theorem

Mathematical connection between SVD and PCA

PCA

Definition

$$Cov(X) = E^t D E$$

$$Z = E X_\mu \quad \text{scores}$$

$$D = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \end{bmatrix}$$ eigenvalues

SVD

Definition

$$X_\mu = U S V^t$$

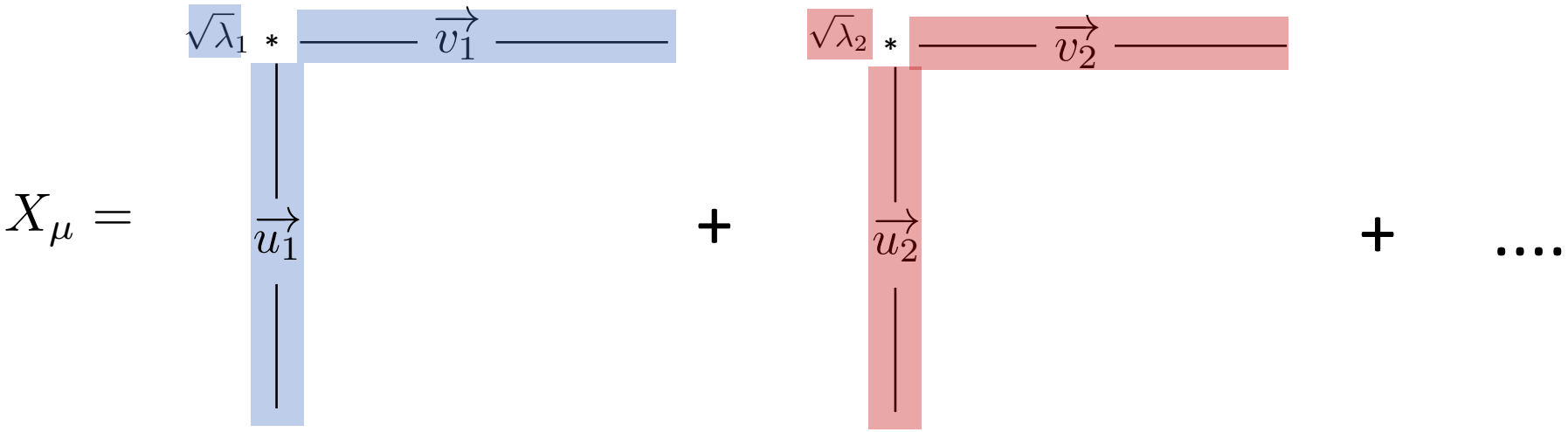
Connections with PCA:

- $U = E^t$
- $SV^t = Z$
- $S = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \sqrt{\lambda_2} & \\ 0 & & \ddots \end{bmatrix}$

=> PCA and SVD are closely connected and can be used interchangibly for all our purposes!

Outer product view of SVD

SVD can be seen as decomposing the data into a sum of rank-one matrices, each defined by an outer-product of the corresponding left and right singular vectors and weighted by the corresponding singular value:



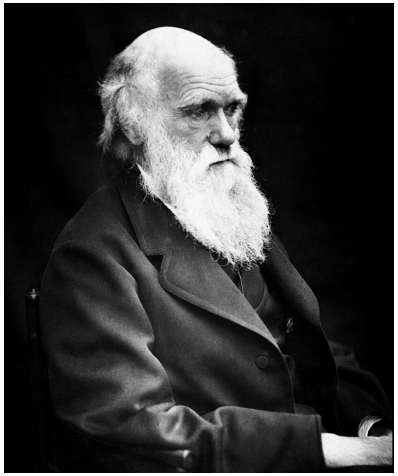
It can sometime be useful to think of the dataset as such a ranked sum of outer-products, where each term accounts for a decreasing fraction of the variance.

Next lecture

- Linear regression / logistic regression

Feedback / suggestions always welcome at:
https://presemo.aalto.fi/bda2023

Supplementary material



Charles Darwin

Galton and Pearson

Charles Darwin

195 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

For other people named Charles Darwin, see Charles Darwin (disambiguation).

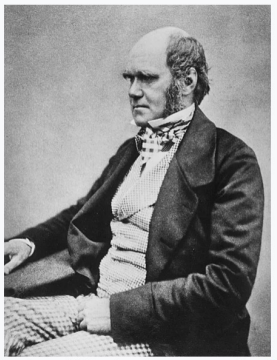
Charles Robert Darwin FRS FRGS FLS FZS JP[6] (/ˈdɑːrwɪn/[7] *DAR-win*; 12 February 1809 – 19 April 1882) was an English naturalist, geologist, and biologist,[8] widely known for his contributions to evolutionary biology. His proposition that all species of life have descended from a common ancestor is now generally accepted and considered a fundamental concept in science.[9] In a joint publication with Alfred Russel Wallace, he introduced his scientific theory that this branching pattern of evolution resulted from a process he called natural selection, in which the struggle for existence has a similar effect to the artificial selection involved in selective breeding.[10] Darwin has been described as one of the most influential figures in human history and was honoured by burial in Westminster Abbey.[11][12]

Darwin's early interest in nature led him to neglect his medical education at the University of Edinburgh; instead, he helped to investigate marine invertebrates. His studies at the University of Cambridge's Christ's College from 1828 to 1831 encouraged his passion for natural science.[13] His five-year voyage on HMS *Beagle* from 1831 to 1836 established Darwin as an eminent geologist, whose observations and theories supported Charles Lyell's concept of gradual geological change. Publication of his journal of the voyage made Darwin famous as a popular author.[14]

Puzzled by the geographical distribution of wildlife and fossils he collected on the voyage, Darwin began detailed investigations and, in 1838, devised his theory of natural selection.[15] Although he discussed his ideas with several naturalists, he needed time for extensive research and his geological work had priority.[16] He was writing up his theory in 1858 when Alfred Russel Wallace sent him an essay that described the same idea, prompting immediate joint submission of both their theories to the Linnean Society of London.[17] Darwin's work established evolutionary descent with modification as the dominant scientific explanation of diversification in nature.[18] In 1871, he examined human evolution and sexual selection in *The Descent of Man, and Selection in Relation to Sex*, followed by *The Expression of the Emotions in Man and Animals* (1872). His research on plants was published in a series of books, and in his final book, *The Formation of Vegetable Mould, through the Actions of Worms* (1881), he examined earthworms and their effect on soil.

Darwin published his theory of evolution with compelling evidence in his 1859 book *On the Origin of Species*.[19][20] By the 1870s, the scientific community and a majority of the educated public had accepted evolution as a fact. However, many favoured competing explanations that gave only a minor role to natural selection, and it was not until the emergence of the modern evolutionary synthesis from the 1930s to the 1950s that a broad consensus developed in which natural selection was the basic mechanism of evolution.[18][21] Darwin's scientific discovery is the unifying theory of the life sciences, explaining the diversity of life.

Charles Darwin
FRS FRGS FLS FZS JP



Darwin, c. 1854, when he was preparing *On the Origin of Species*[4]

Born	Charles Robert Darwin 12 February 1809 Shrewsbury, England
Died	19 April 1882 (aged 73) Down, Kent, England
Resting place	Westminster Abbey
Alma mater	University of Edinburgh Christ's College, Cambridge (BA, 1831; MA, 1836)[5]
Known for	*The Voyage of the Beagle* *On the Origin of Species* *The Descent of Man*
Spouse	Emma Wedgwood (m. 1839)
Children	10
Parents	Robert Darwin Susannah Wedgwood