Statistics and ML – 2

A. K-means and B. Principal Component Analysis
(aka PCA, Hotelling transform, Karhunen-Loève expansion)

Edoardo Milotti

Advanced Statistics for Physics

A.Y. 2023-24

A. The K-means algorithm

History of the K-means algorithm

The K-means problem belongs to the class of NP problems. However, K-means is an efficient heuristic solution and a very popular unsupervised algorithm used for clustering tasks.

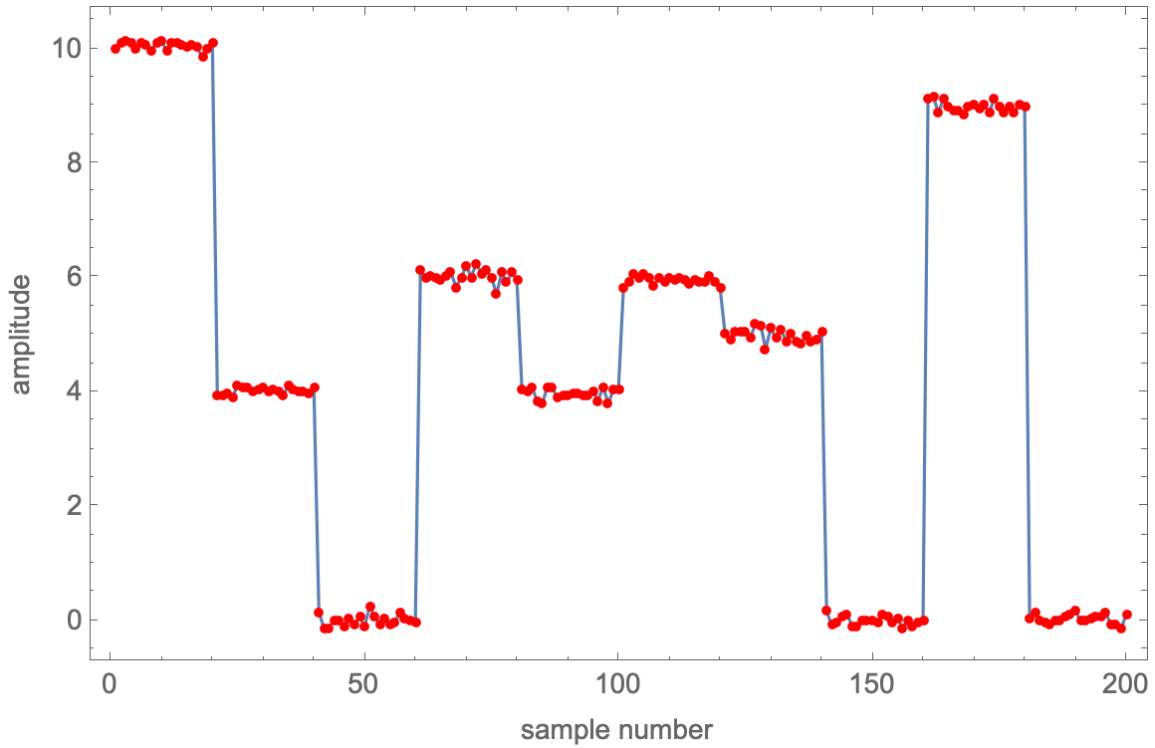
The algorithm was originally proposed by the Polish Mathematician Hugo Steinhaus in 1956.
Hugo Steinhaus made amazing contributions to functional analysis (see, e.g, the Banach-Steinhaus Theorem). He also played a key role in the reconstruction of mathematics in Poland after WWII.

Later on, in 1967, James MacQueen introduced the term K-Means in his article “Some Methods for Classification and Analysis of Multivariate Observations” published by University of California.

The standard algorithm was first proposed by Stuart Lloyd of Bell Labs in 1957 as a technique for pulse-code modulation, although it was not published as a journal article until 1982.
In 1965, Edward W. Forgy published essentially the same method, which is why it is sometimes referred to as the Lloyd–Forgy algorithm.

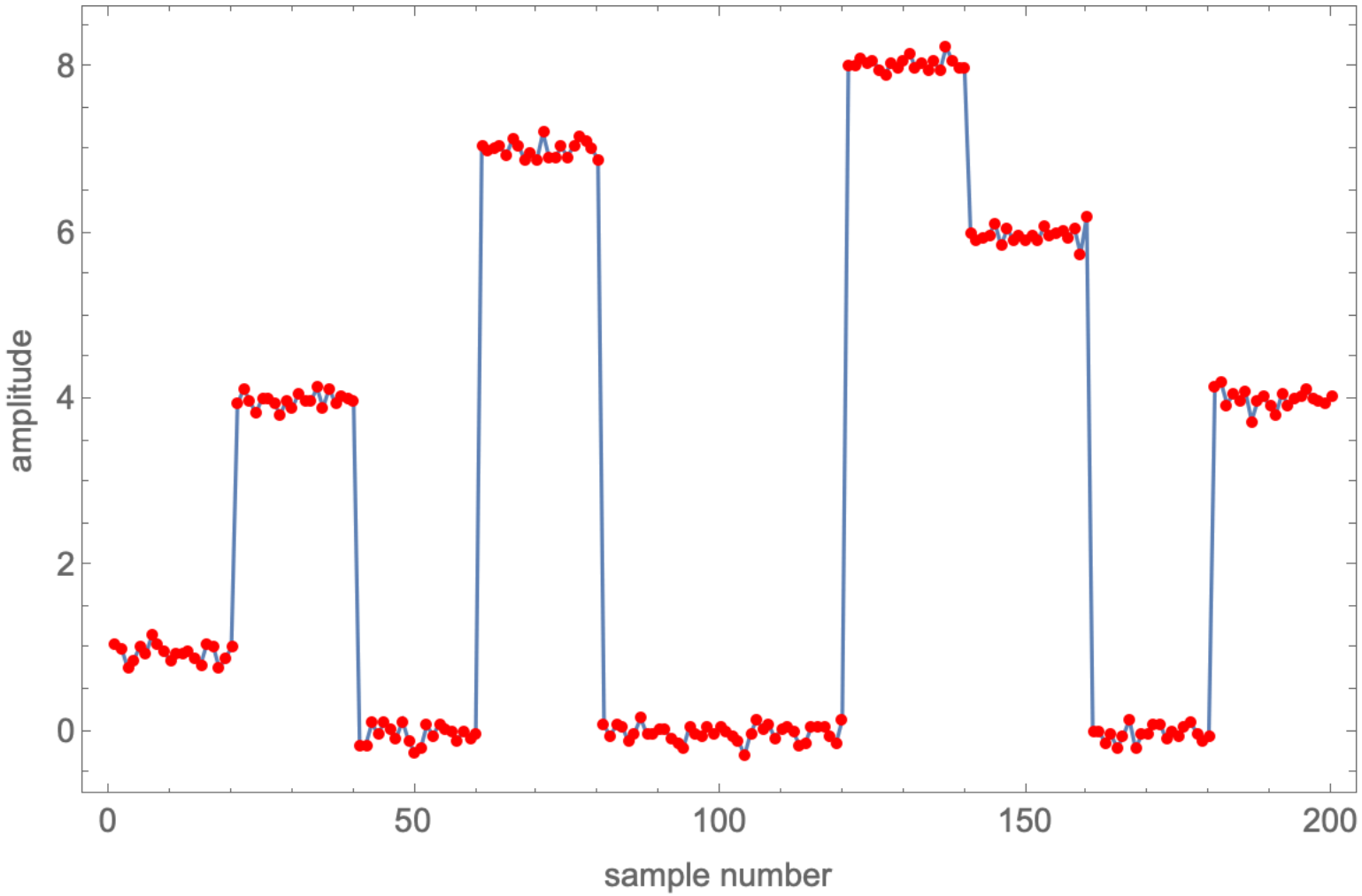
Least Squares Quantization in PCM

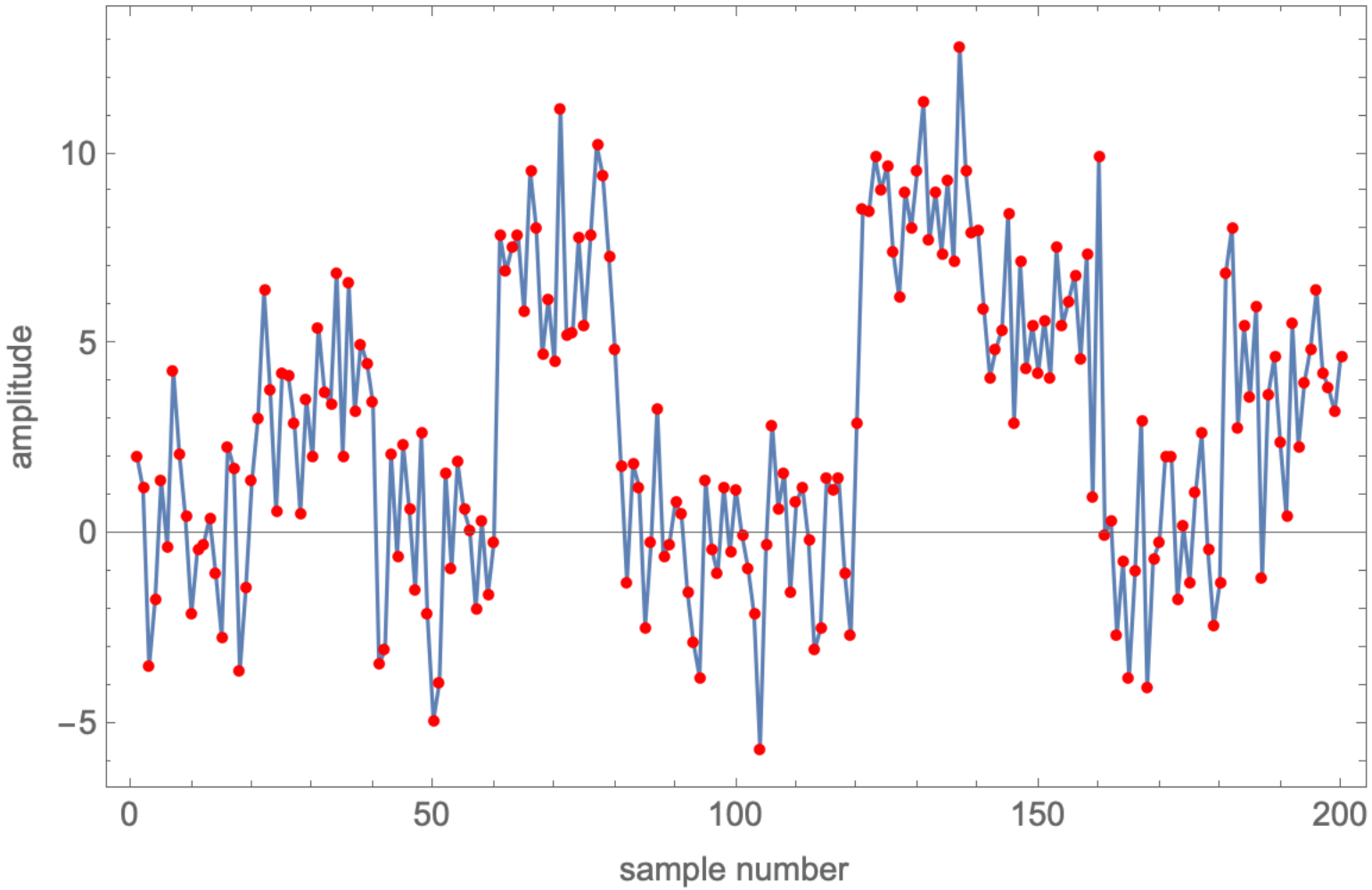
STUART P. LLOYD

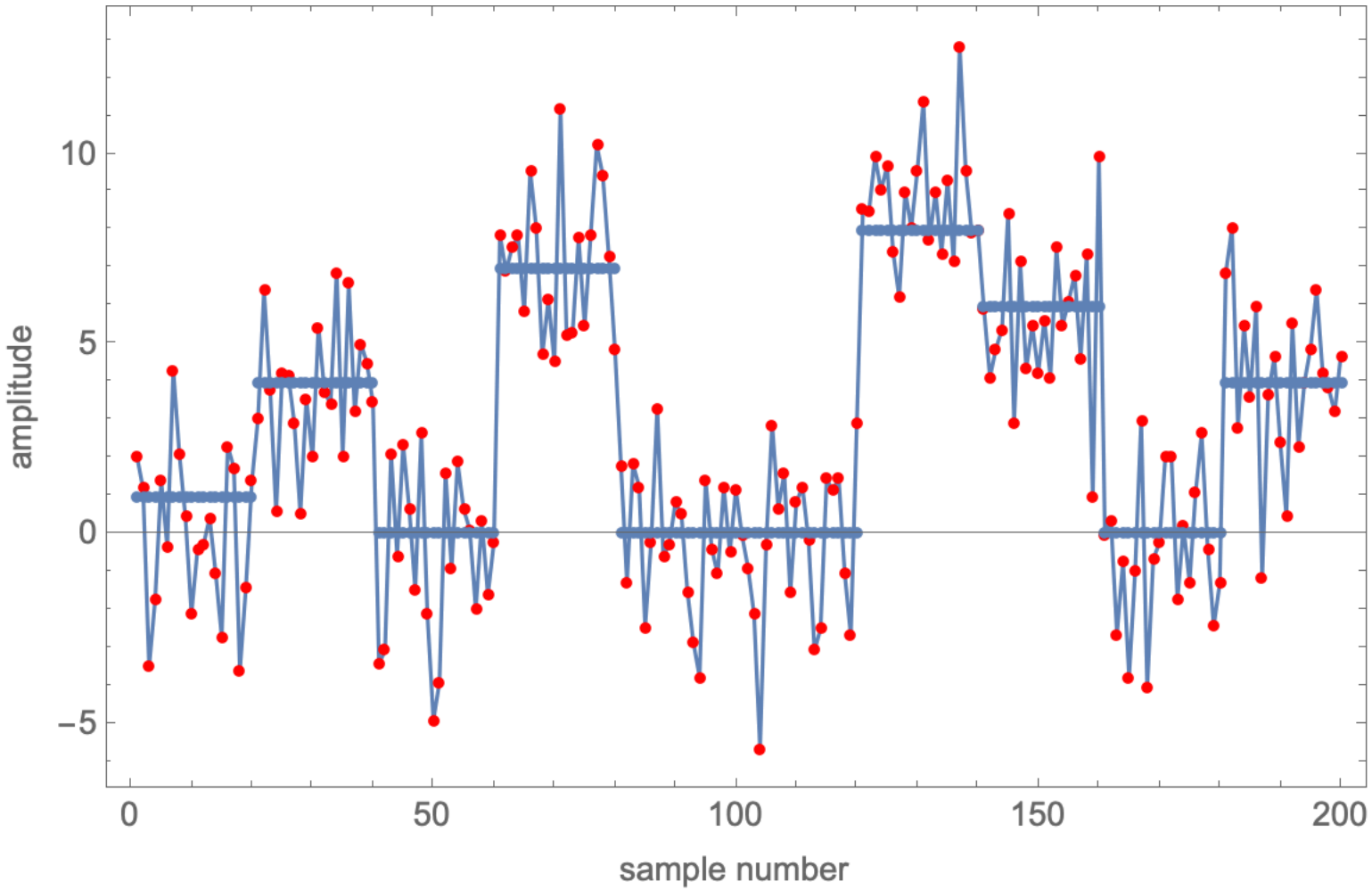


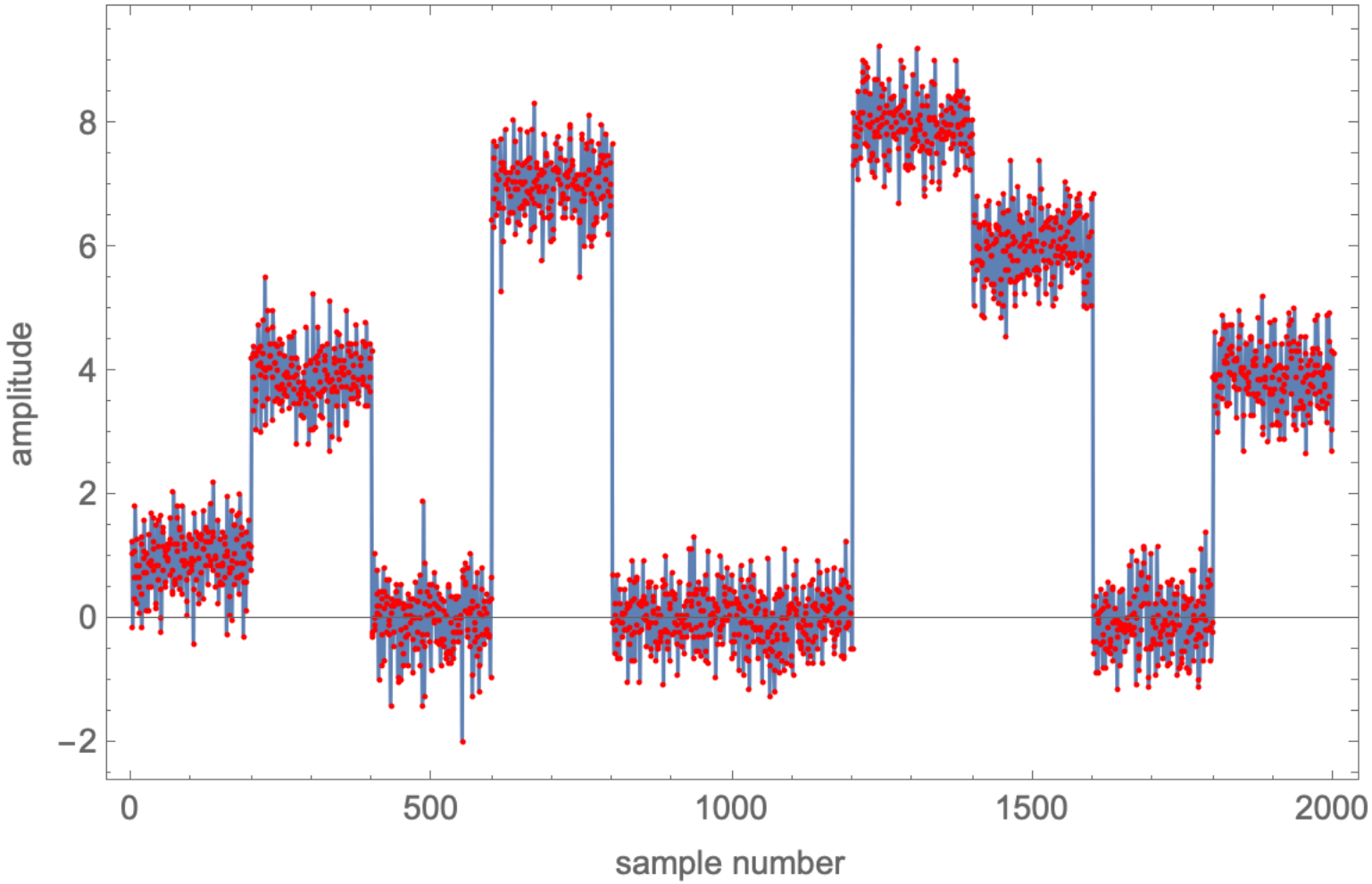
PCM is a modification of this. Instead of sending the exact sample values (3), one partitions the voltage range of the signal into a finite number of subsets and transmits to the receiver only the information as to which subset a sample happens to fall in. Built into the receiver there is a source of fixed representative voltages—"quanta"—one for each of the subsets. When the receiver is informed that a certain sample fell in a certain subset, it uses its quantum for that subset as an approximation to the true sample value and constructs a band-limited signal based on these approximate sample values.

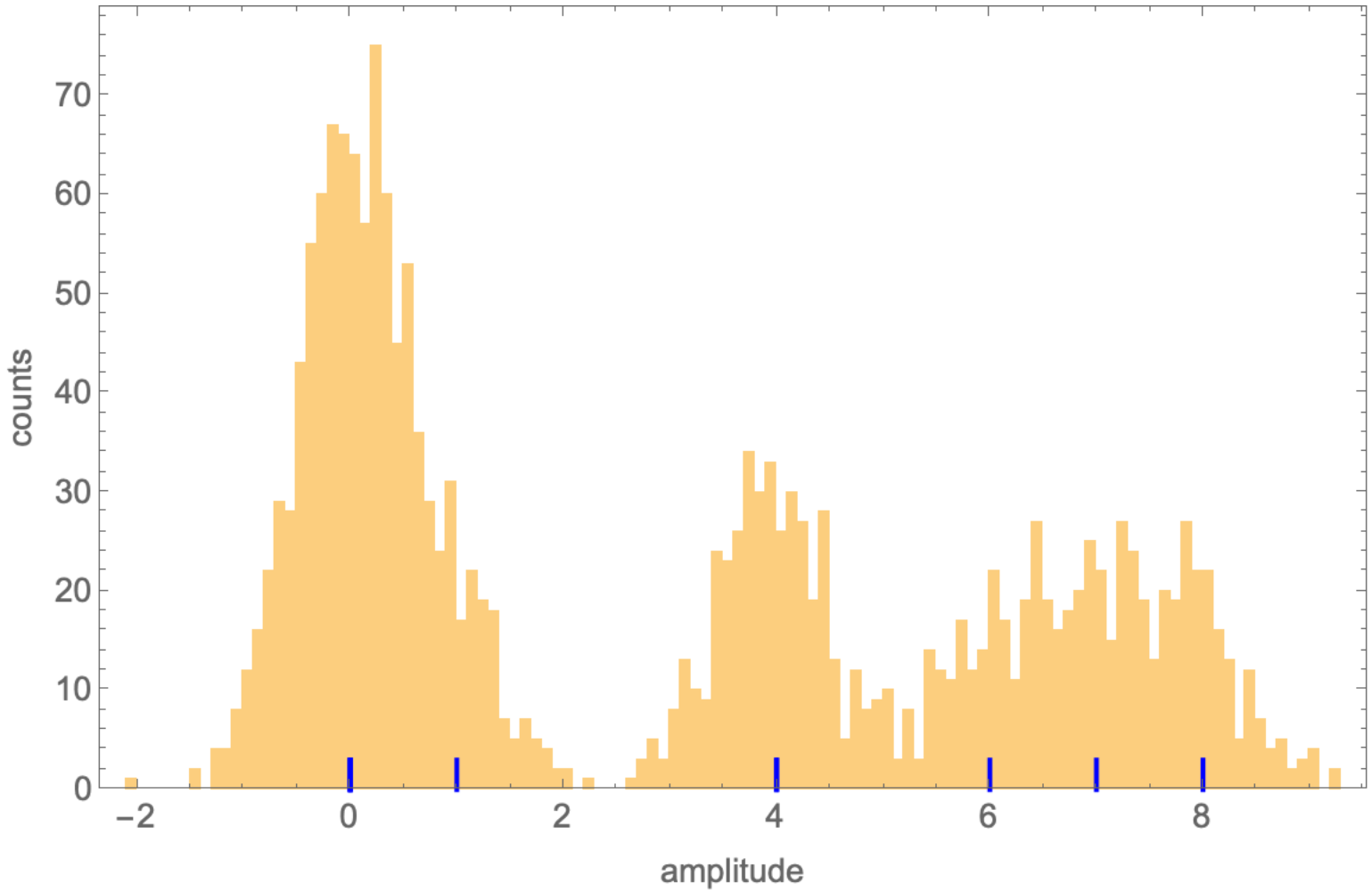
We define the *noise signal* as the difference between the receiver-output signal and the original signal and the *noise power* as the average square of the noise signal. The problem we consider is the following: given the number of quanta and certain statistical properties of the signal, determine the subsets and quanta that are best in minimizing the noise power.


amplitude
sample number
0
2
4
6
8
0
50
100
150
200


amplitude
10
5
0
−5
0
50
100
150
200
sample number


amplitude
10
5
0
−5
0
50
100
150
200
sample number


amplitude
8
6
4
2
0
−2
0
500
1000
1500
2000
sample number


70
60
50
40
30
20
10
0
counts
−2
0
2
4
6
8
amplitude

The problem can be generalized to clusters in D-dimensional space

- K means (one for each expected cluster): $\boldsymbol{\mu}_k$
- association of a data point to a given cluster is given by the matrix

$$[r_{nk}] = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & 0 & \dots & 0 \end{pmatrix}$$

- D x K matrix (K is the number of clusters)
- One row per D-dimensional data point
- Each row contains a single 1 entry in column *k* if the n-th data point is in the *k*-th cluster (*1-of-K coding*)
- This means that there is just one entry per row, but there can be more than one entry per column

- the association of a point to a cluster is obtained by minimizing the following objective function (the *distortion measure*)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \boldsymbol{\mu}_k \right\|^2$$

The minimization of the distorsion measure is achieved by iteration of the following two steps

- E step: change the assignments of data points to centers to minimize *J*, i.e.,

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

- M step: minimize with respect to the means, setting derivatives to 0,

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

to find the solution

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

- Since the denominator of this expression is just equal to the total number of points assigned to the k-th cluster and the sum is only over the data points in the cluster, this sum equals the mean of all the data points in the cluster, hence the name *k-means*.

Example, using the Old Faithful data set

```
Old Faithful Geyser Data

Description: (From R manual):

     Waiting time between eruptions and the duration of the eruption
     for the Old Faithful geyser in Yellowstone National Park, Wyoming,
     USA.

     A data frame with 272 observations on 2 variables.

eruptions  numeric  Eruption time in mins
waiting    numeric  Waiting time to next eruption

References:

     Hardle, W. (1991) Smoothing Techniques with Implementation in S.
     New York: Springer.

     Azzalini, A. and Bowman, A. W. (1990). A look at some data on the
     Old Faithful geyser. Applied Statistics 39, 357-365.
```

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85

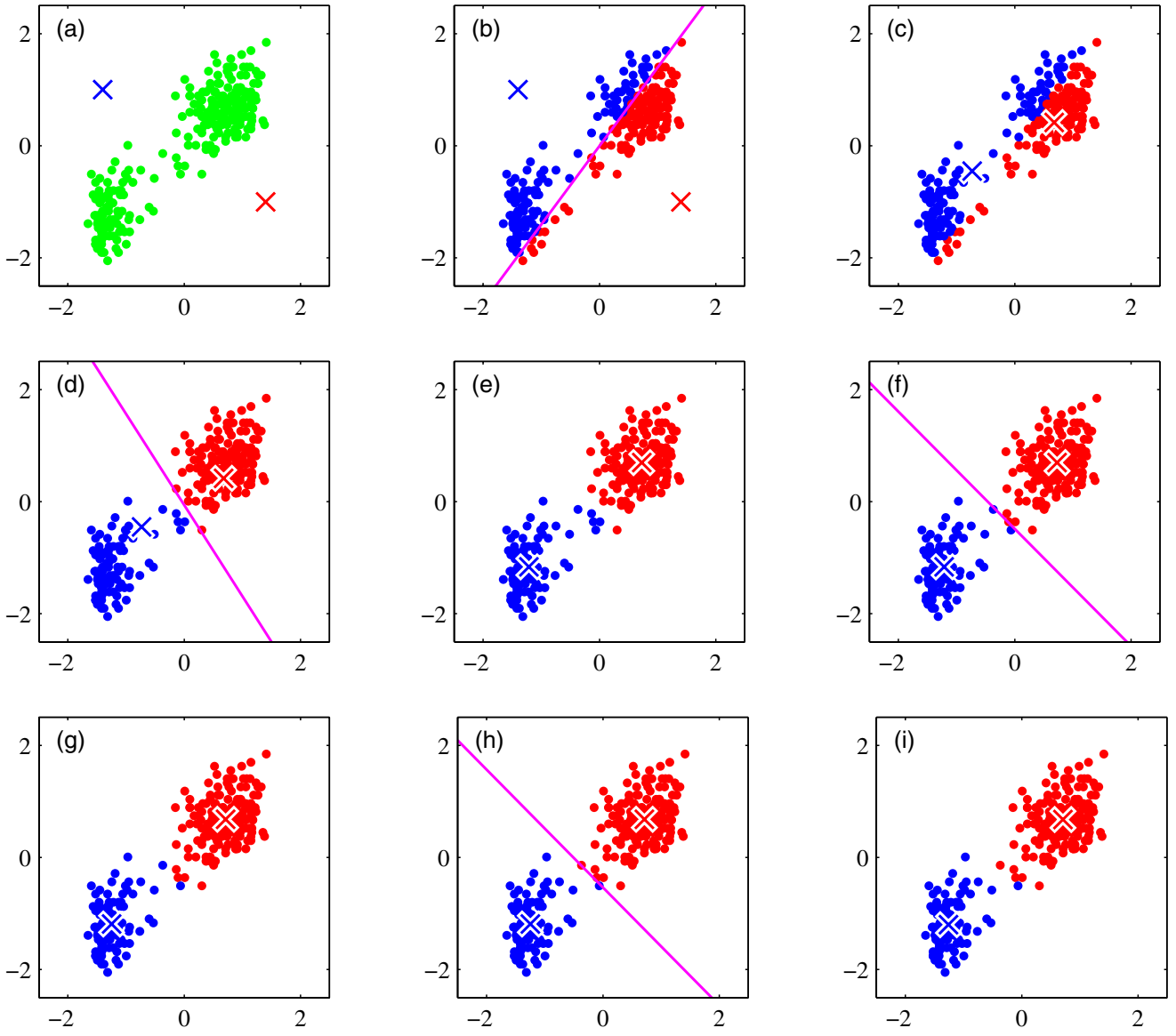
https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat

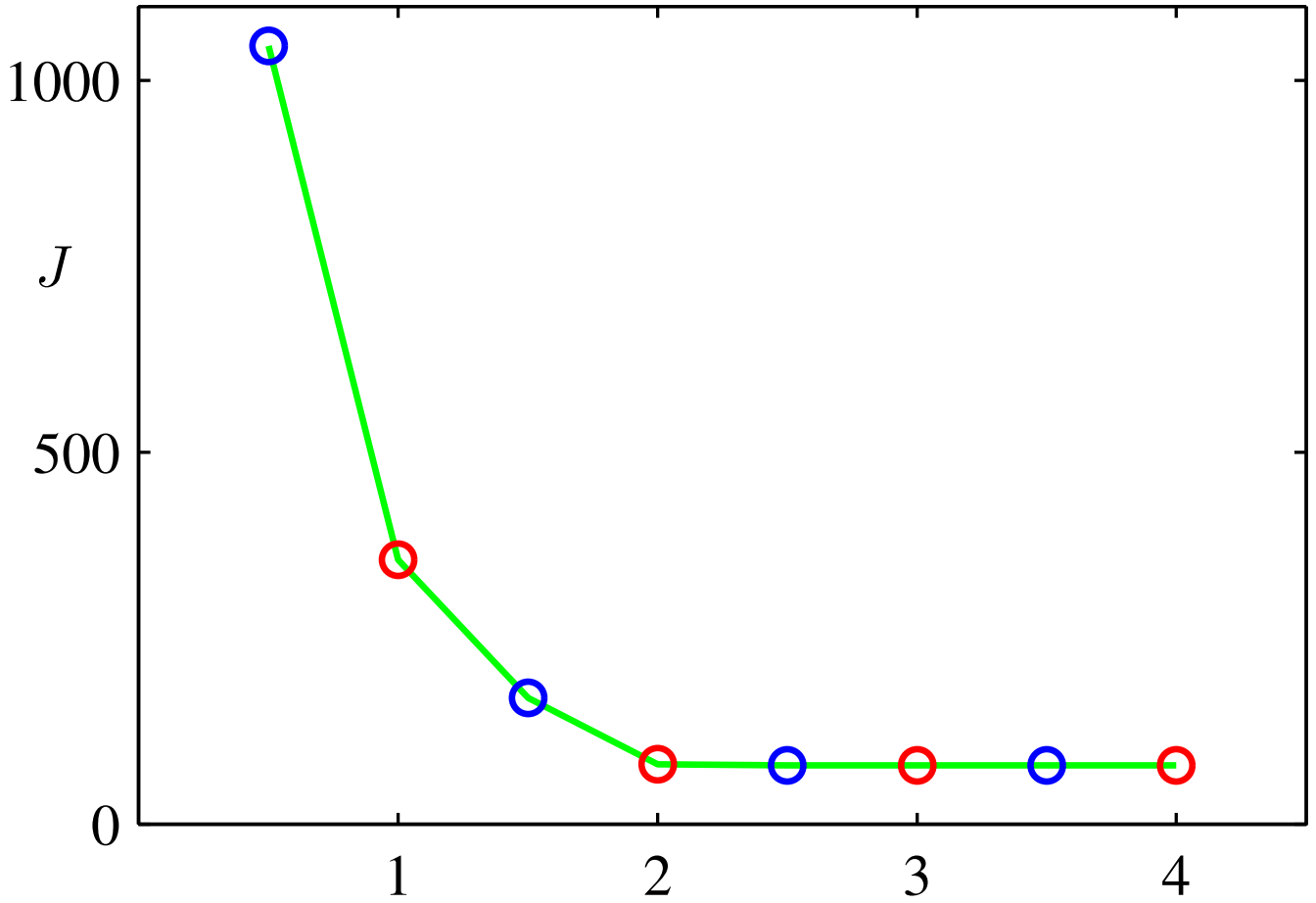
In the example run shown in the next slide, these data have been standardized, i.e., each variable has zero mean and unit standard deviation.

Illustration of the K-means algorithm using the re-scaled Old Faithful data set.

- Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ_1 and μ_2 are shown by the red and blue crosses, respectively.
- In the initial E step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer. This is equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centres, shown by the magenta line, they lie on.
- In the subsequent M step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster.
- show successive E and M steps through to final convergence of the algorithm.

(adapted from C. M. Bishop, "Pattern recognition and machine learning", Springer (2006))





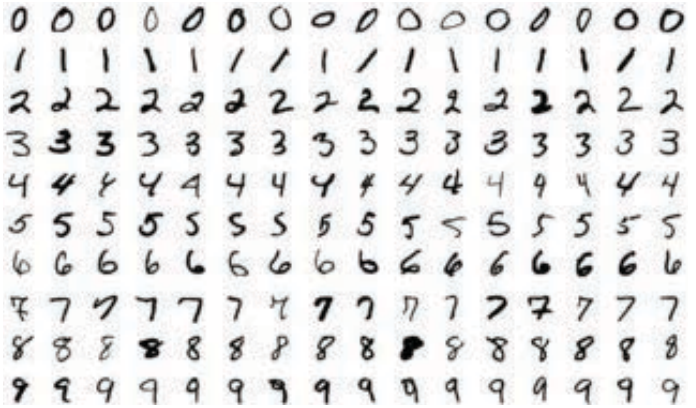
Plot of the cost function *J* after each E step (blue points) and M step (red points) of the K- means algorithm for the Old Faithful example. The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors.

(adapted from C. M. Bishop, "Pattern recognition and machine learning", Springer (2006))

Datasets for machine learning research

In addition to the Old Faithful data, there are many more datasets that can be used for ML research and training purposes.

- The MNIST database of handwritten digits (http://yann.lecun.com/exdb/mnist/), which has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST (https://www.nist.gov/srd/nist-special-database-19). The digits are size-normalized and centered in a fixed-size image.
- Visualization of the MNIST database (https://github.com/mbornet-hl/MNIST/tree/master/IMAGES/GROUPS)
- The extended EMNIST dataset (https://www.nist.gov/itl/products-and-services/emnist-dataset), which is a set of handwritten character digits derived from the NIST Special Database 19 and converted to a 28x28 pixel image format and dataset structure that directly matches the MNIST dataset. For more information, see https://arxiv.org/pdf/1702.05373.pdf.
- Wikipedia on MNIST, https://en.wikipedia.org/wiki/MNIST_database
- Wikipedia on datasets for ML research https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research



Sample images from MNIST test dataset (https://en.wikipedia.org/wiki/MNIST_database#/media/File:MnistExamplesModified.png)

B. Principal Component Analysis

Reprinted from Journal of the Optical Society of America A, Vol. 4, page 519, March 1987
Copyright © 1987 by the Optical Society of America and reprinted by permission of the copyright owner.

Low-dimensional procedure for the characterization of human faces

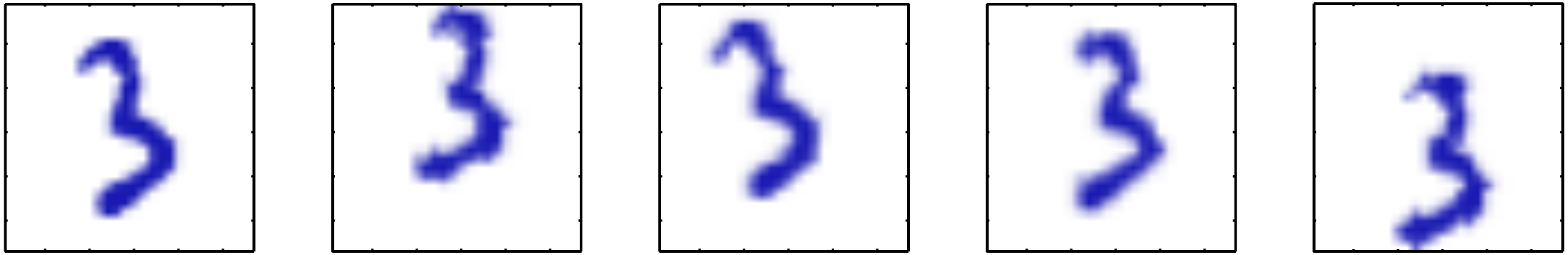
L. Sirovich and M. Kirby

Division of Applied Mathemetics, Brown University, Providence, Rhode Island 02912

Received August 25, 1986; accepted November 10, 1986

A method is presented for the representation of (pictures of) faces. Within a specified framework the representation is ideal. This results in the characterization of a face, to within an error bound, by a relatively low-dimensional vector. The method is illustrated in detail by the use of an ensemble of pictures taken for this purpose.

Image from Turk and Pentland, J. of Cognitive Neuroscience, 3 (1991) 71



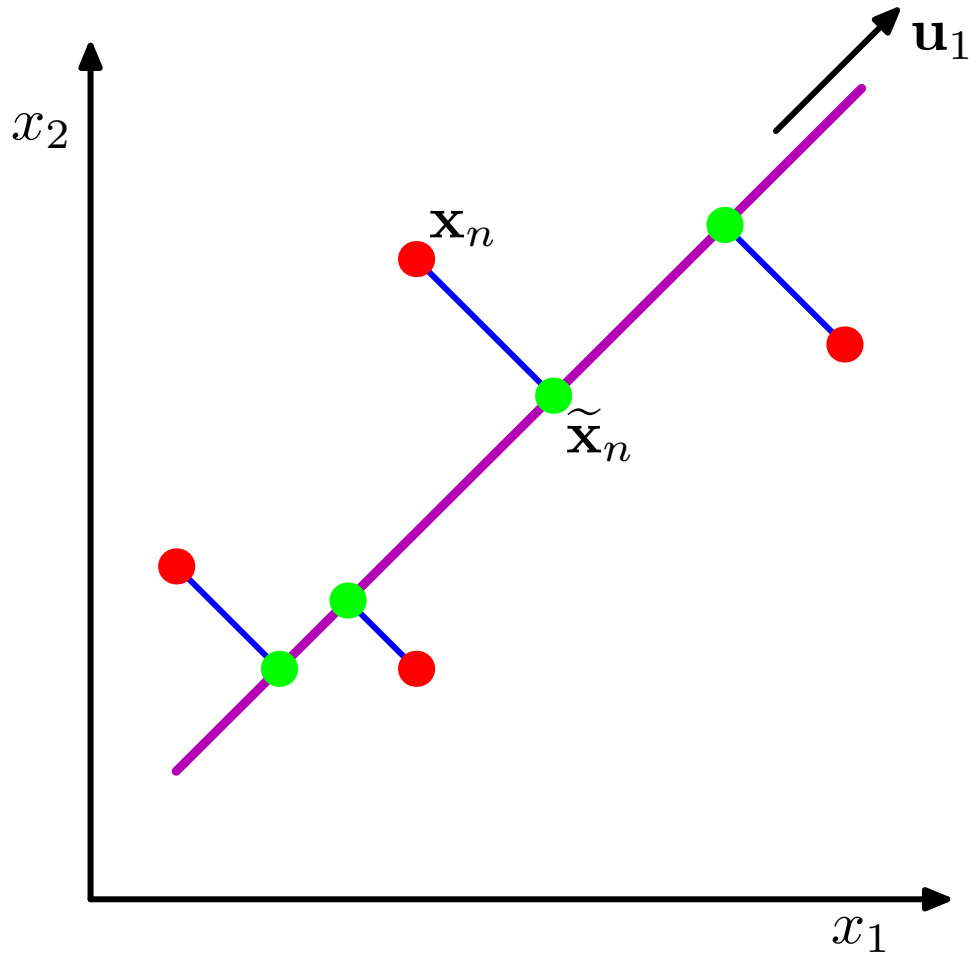
Consider an artificial data set constructed by taking one of the off-line digits, represented by a 64 × 64 pixel grey-level image, and embedding it in a larger image of size 100×100 by padding with pixels having the value zero (corresponding to white pixels) in which the location and orientation of the digit is varied at random.

Each of the resulting images is represented by a point in the 100 × 100 = 10,000-dimensional data space. However, across a data set of such images, there are only three degrees of freedom of variability, corresponding to the vertical and horizontal translations and the rotations. The data points will therefore live on a subspace of the data space whose intrinsic dimensionality is three.

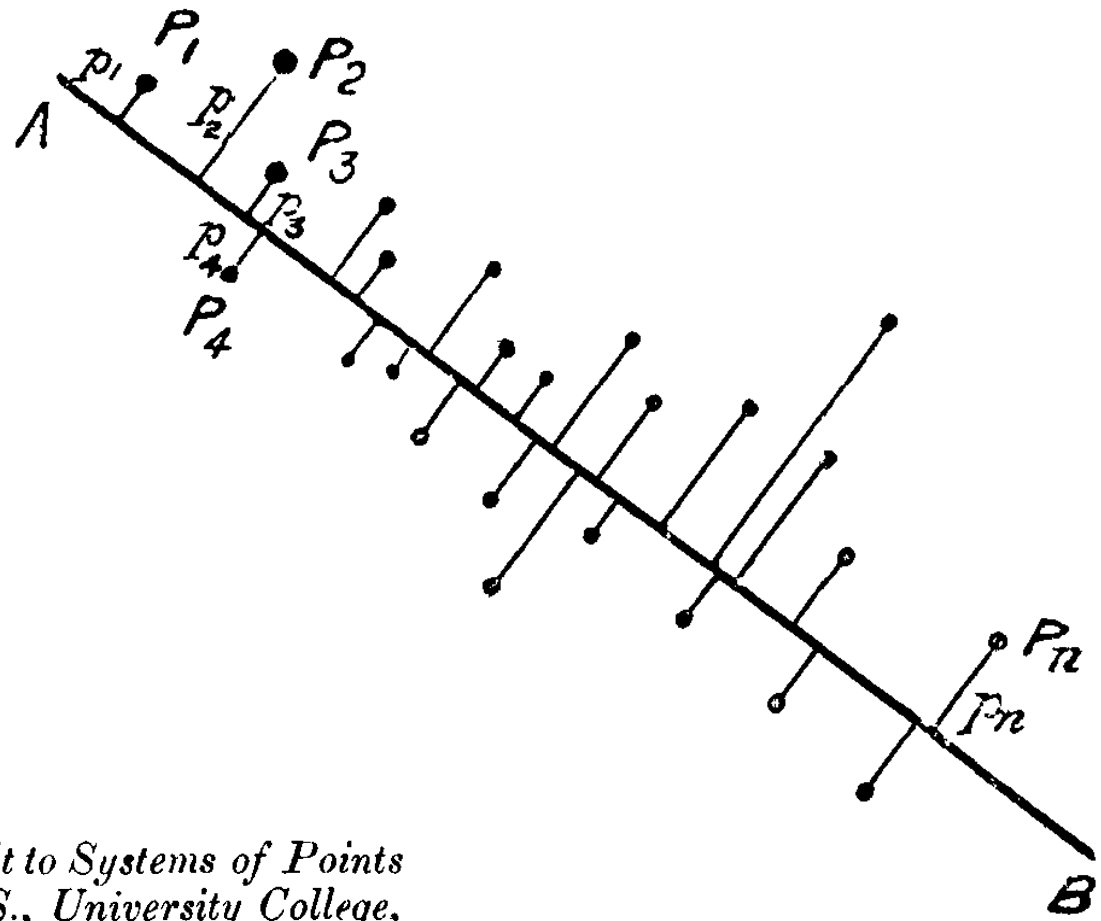
(this and other material is adapted from C. M. Bishop, "Pattern recognition and machine learning", Springer (2006)).

Principal component analysis seeks a space of lower dimensionality, known as the principal subspace and denoted by the magenta line, **such that the orthogonal projection of the data points** (red dots) **onto this subspace maximizes the variance of the projected points** (green dots).

An **alternative definition of PCA** is based on **minimizing the sum-of-squares of the projection errors**, indicated by the blue lines (a least-squares method).



As a curiosity, consider the very similar figure in a paper published in 1901 by Karl Pearson



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

Maximum variance of the projected data samples

- N D-dimensional data samples $\{\mathbf{x}_n\}|_{n=1,\ldots,N}$

- Projection of the D-dimensional data samples onto an 1-dimensional subspace identified by $\hat{\mathbf{u}}_1$
  - mean of data samples $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$
  - projected mean $\hat{\mathbf{u}}_1^T\bar{\mathbf{x}}$
  - variance of projected data

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{\mathbf{u}}_1^T\mathbf{x}_n - \hat{\mathbf{u}}_1^T\bar{\mathbf{x}}\right)^2$$

$$= \frac{1}{N-1}\sum_{n=1}^{N}\hat{\mathbf{u}}_1^T\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^T\hat{\mathbf{u}}_1$$

$$= \hat{\mathbf{u}}_1^T S\hat{\mathbf{u}}_1$$

with $S = \frac{1}{N-1}\sum_{n=1}^{N}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^T$

(covariance matrix)

Maximum variance of the projected data samples – 2

- variance of projected data $\sigma^2 = \hat{\mathbf{u}}_1^T S \hat{\mathbf{u}}_1$
  - variance maximization with constraint

$$\hat{\mathbf{u}}_1^T S \hat{\mathbf{u}}_1 + \lambda_1 \left(1 - \hat{\mathbf{u}}_1^T \hat{\mathbf{u}}_1\right) \quad \Rightarrow \quad S\hat{\mathbf{u}}_1 - \lambda_1 \hat{\mathbf{u}}_1 = 0 \quad \Rightarrow \quad S\hat{\mathbf{u}}_1 = \lambda_1 \hat{\mathbf{u}}_1$$

i.e., **the variance is maximized taking a vector that is the normalized eigenvector with the largest eigenvalue (this eigenvector is the *first principal component*)**

In particular, the variance is

$$\hat{\mathbf{u}}_1^T S \hat{\mathbf{u}}_1 = \lambda_1 \hat{\mathbf{u}}_1^T \hat{\mathbf{u}}_1 = \lambda_1$$

- we can find more components by iterating this procedure (they are the eigenvectors of the covariance matrix, in order of eigenvalue magnitude)

Singular Value Decomposition (SVD)

This is a general method to find all the components.

We start with the following matrix where the rows (N) list the D-dimensional data vectors, where we assume that the mean value has already been subtracted from all data vectors

$$A = [x_{i,j}] = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1,D} \\ x_{21} & x_{22} & \cdots & x_{2,D} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,D} \end{pmatrix}$$

(i.e., A is a N x D matrix). The covariance matrix is a D x D matrix given by the matrix product

$$(A^T A)_{i,j} = \sum_{k=1}^{N} x_{k,i} x_{k,j} = (n-1)\mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j) = V_{i,j}$$

Since

$$(A^T A)_{i,j} = \sum_{k=1}^{N} x_{k,i} x_{k,j} = (n-1)\mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j) = V_{i,j}$$

then the same D x D orthogonal matrix U of normalized eigenvectors of V diagonalizes both V and A^TA, i.e.,

$$S^T S = U^T (A^T A) U = (AU)^T (AU) = (W^T AU)^T (W^T AU)$$

where S^TS is a diagonal D x D matrix, $\mathrm{diag}(\lambda_1, \ldots, \lambda_D)$, the λ_k are the eigenvalues of A^TA usually listed in descending order, and W is an N x N matrix such that

$$S = W^T AU \quad \Rightarrow \quad A = W S U^T$$

so that S is an N x D matrix. The transpose operations have been chosen so that the final transformation of the A matrix looks like a transformation of a square matrix.

We have defined matrices so that A is an $N \times D$ matrix, U is a $D \times D$ matrix, S is $N \times D$ matrix and W is an N x N matrix.

Notice also that we can easily diagonalize the $N \times N$ matrix AA^T as follows

$$AA^T = (WSU^T)\,(WSU^T)^T = WSU^TUS^TW^T = WSS^TW^T$$

and we conclude that W is the matrix of the normalized eigenvectors of AA^T .

The matrix $A = WSU^T$ with

$$S = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sqrt{\lambda_D} \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

is called the **singular value decomposition** of A, and it works even when the S matrix is singular.

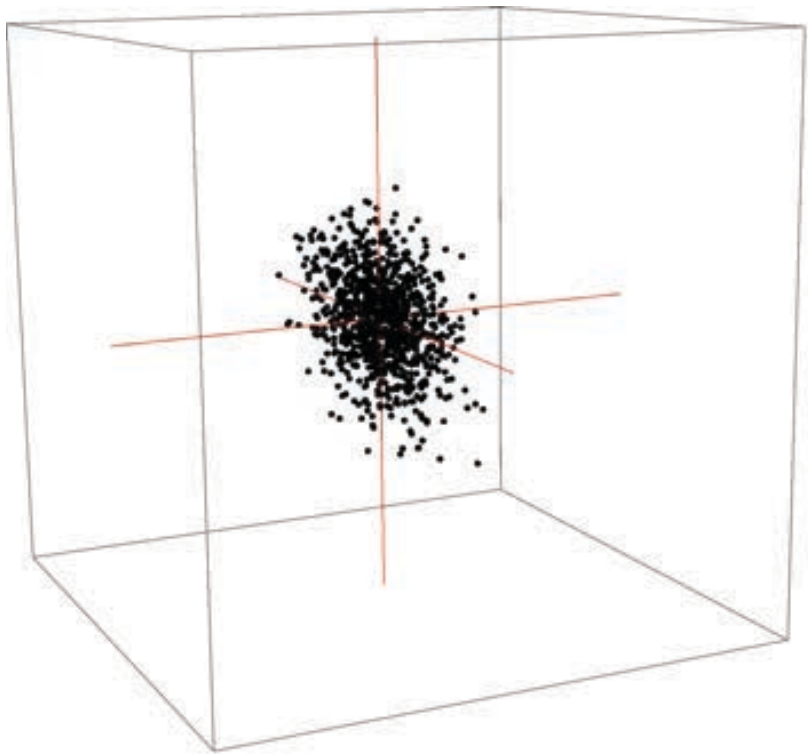
We can use the matrix equations to obtain a straightforward geometric interpretation.

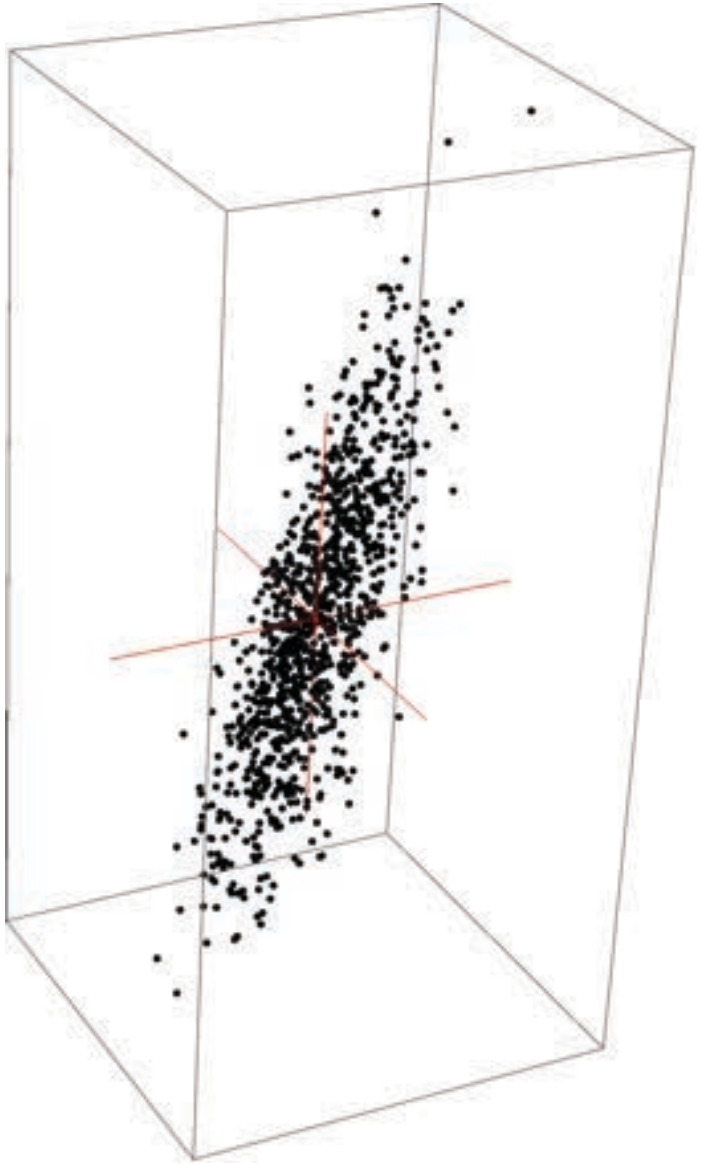
Since

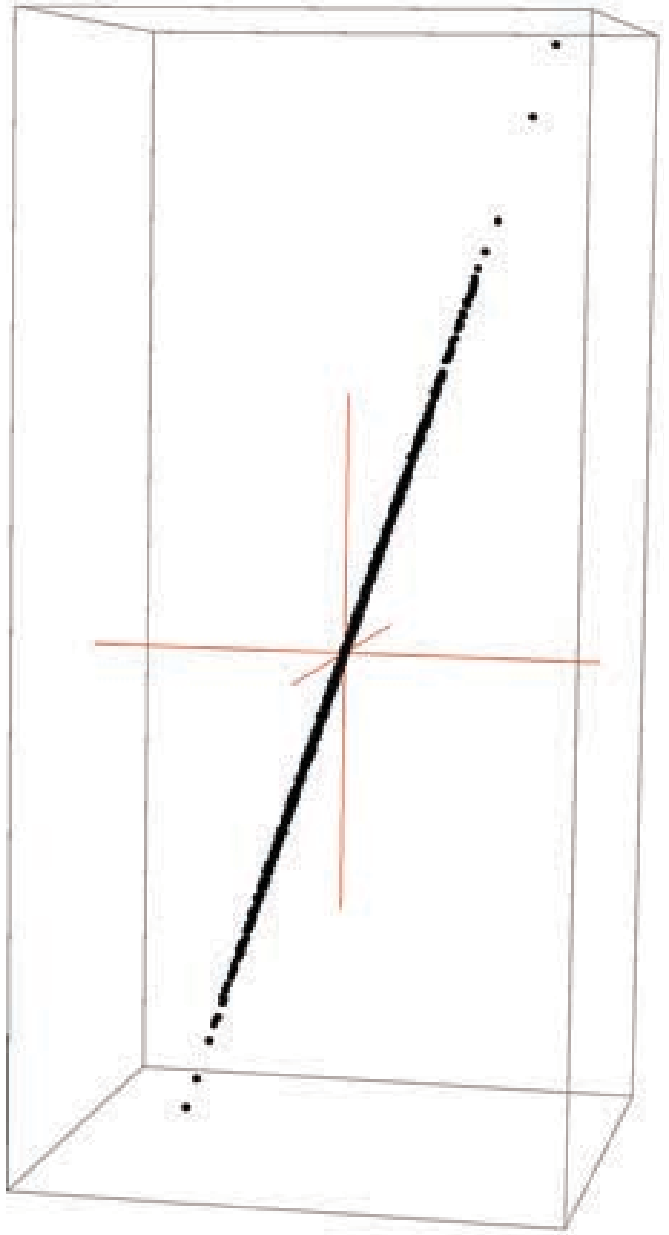
$$A = WSU^T \quad \Rightarrow \quad AU = WS \quad \Rightarrow \quad \sum_{k=1}^{D} x_{ik} U_{kl} = \sum_{k=1}^{D} (U^T)_{lk} x_k^{(i)}$$

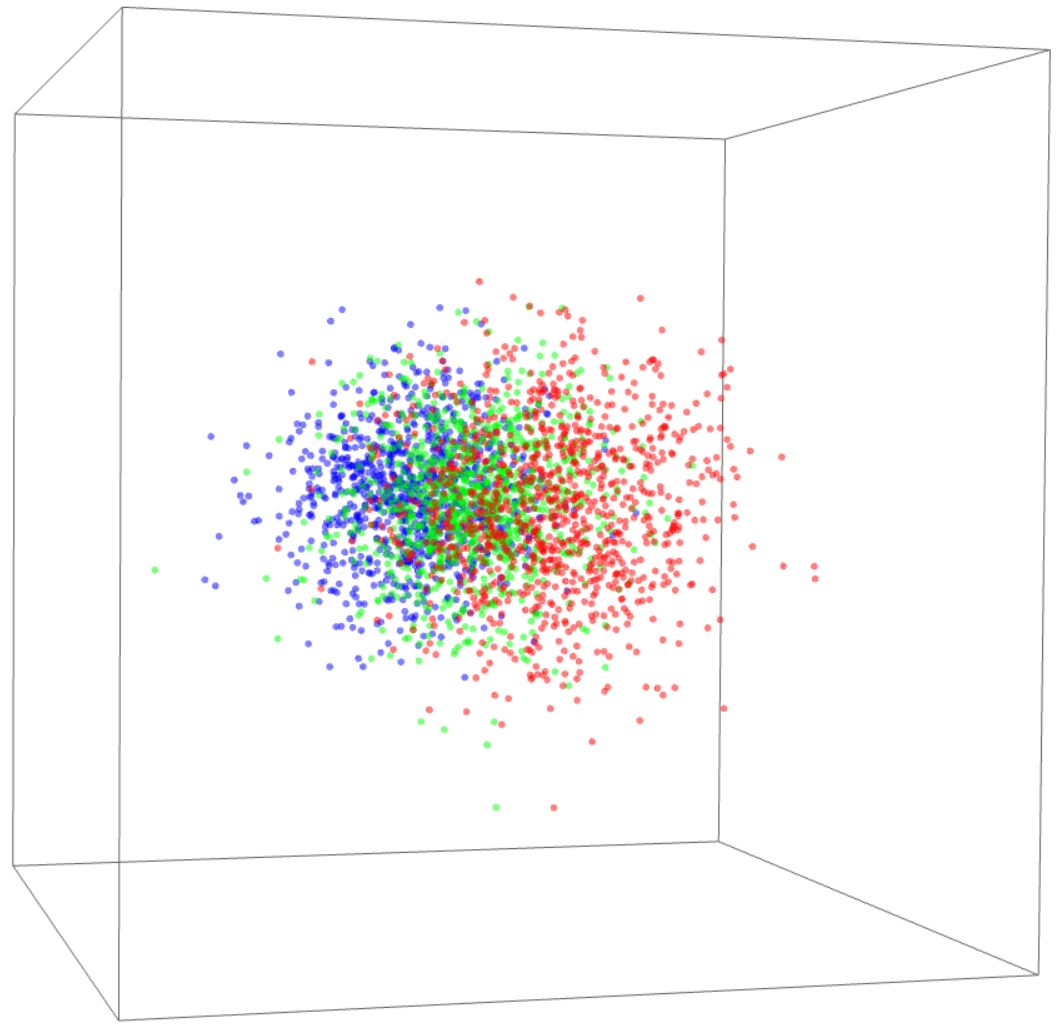
where $x^{(i)}$ is the vector result of the *i*-th measurement, and **we see that the term** AU **corresponds to a rotation of all the initial measurement vectors to a new frame of reference**.
The r.h.s. of the equation in the middle shows what happens: *W* is a large N x N matrix that encodes the measurement data in a different way, and each of its rows is acted upon by the matrix *S* which may contain null eigenvalues and has plenty of zeros.

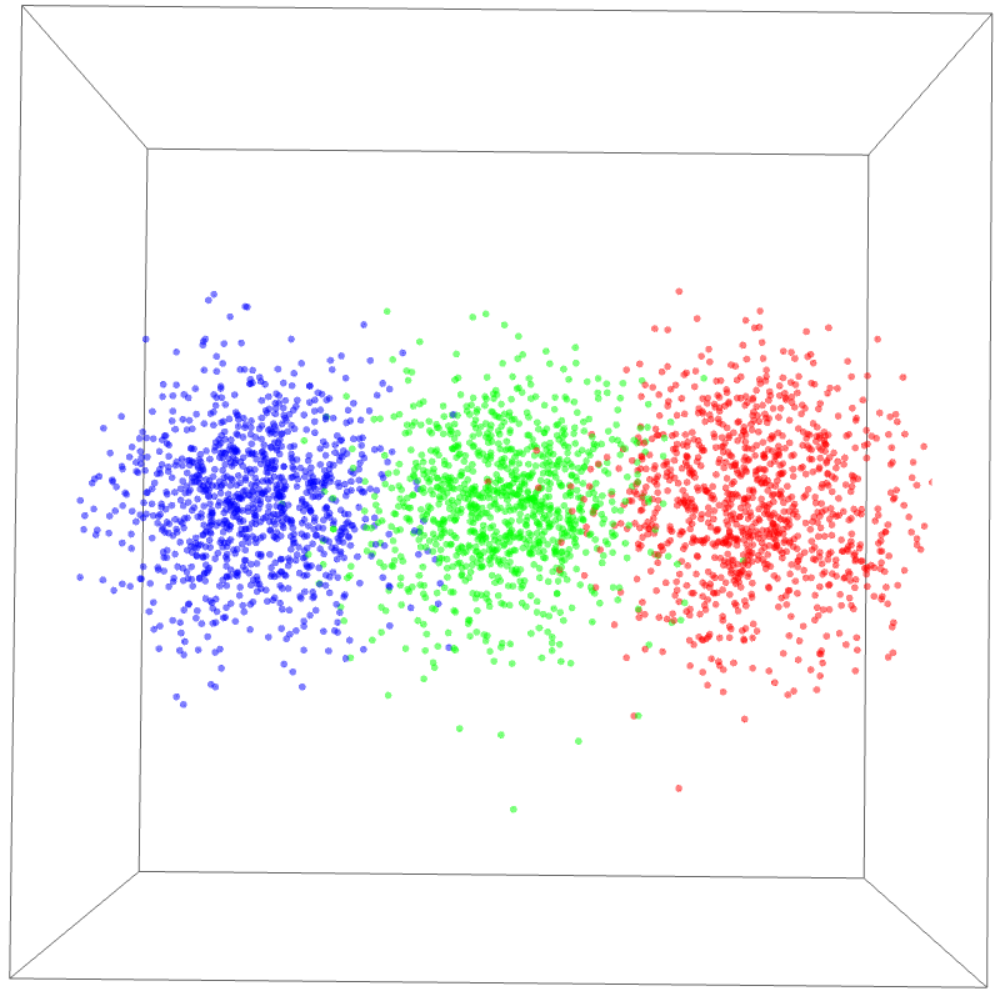
This means that most of the information of *W* is nulled and that additional components can be neglected because of null or near-zero eigenvalues, i.e., the transformation AU acts both as a rotation that makes the transformed random variable independent and clearly points to the dependent variables (those with null or near-zero eigenvalues).











Face recognition

When viewed as vectors, images are very-high-dimensional vectors. However, amid all these vectors, very few correspond to actual faces.

With PCA, we can find a compact representation of human (or animal!) faces.

Image from Turk and Pentland, J. of Cognitive Neuroscience, 3 (1991) 71

Turk and Pentland, J. of Cognitive Neuroscience **3** (1991) 71

Eigenfaces for Recognition

Matthew Turk and Alex Pentland
Vision and Modeling Group
The Media Laboratory
Massachusetts Institute of Technology

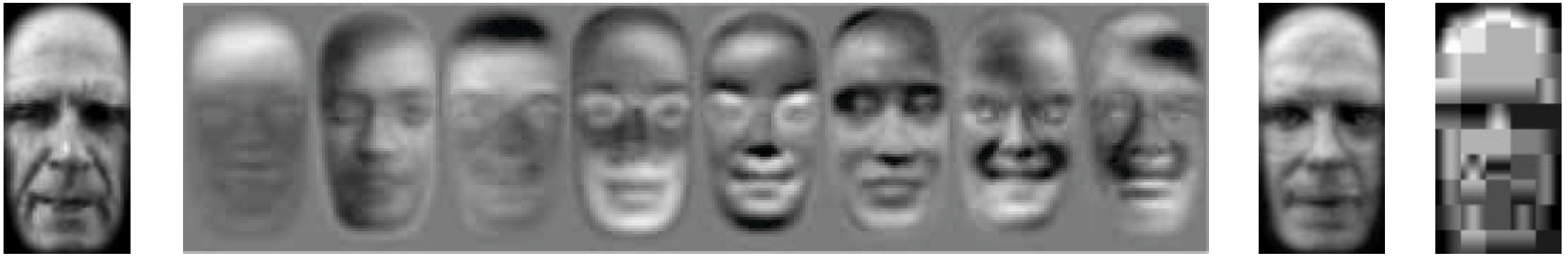
Abstract

■ We have developed a near-real-time computer system that can locate and track a subject's head, and then recognize the person by comparing characteristics of the face to those of known individuals. The computational approach taken in this system is motivated by both physiology and information theory, as well as by the practical requirements of near-real-time performance and accuracy. Our approach treats the face recognition problem as an intrinsically two-dimensional (2-D) recognition problem rather than requiring recovery of three-dimensional geometry, taking advantage of the fact that faces are normally upright and thus may be described by a small set of 2-D characteristic views. The system functions by projecting face images onto a feature space that spans the significant variations among known face images. The significant features are known as "eigenfaces," because they are the eigenvectors (principal components) of the set of faces; they do not necessarily correspond to features such as eyes, ears, and noses. The projection operation characterizes an individual face by a weighted sum of the eigenface features, and so to recognize a particular face it is necessary only to compare these weights to those of known individuals. Some particular advantages of our approach are that it provides for the ability to learn and later recognize new faces in an unsupervised manner, and that it is easy to implement using a neural network architecture. ■

PCA is used in many applications such as *face recognition*. Start with N pictures of faces, then the basic idea goes as follows:

- take the image vector $\phi_{i,j}$ with $i, j = 1, \ldots, D$ (each element represents the gray level encoded in a pixel at position i,j) and form the new vector $\phi_{j+(i-1)D} = \phi_{i,j}$
- average the individual pixel values to form an *average face* ϕ_{AV}
- for each image calculate the difference from the mean $\varphi = \phi - \phi_{AV}$; this is called *the caricature*
- form an $N \times D^2$ matrix where the N rows are the individual caricatures φ
- apply the SVD to this matrix
- it is empirically found that there are only few large eigenvectors, and all images fall in a low dimensional subspace of the global space (potentially a dimension D^2 space, or at least that there are only few dimensions that really matter in the rotated reference frame)
- different faces correspond to different regions in the subspace, and to recognize a face it is sufficient to find the distance (limited to this subspace) of the new rotated caricature from the centers of the regions that represent the individual faces (the *eigenfaces*)
- This method was first developed by Sirovich and Kirby in 1987, and their original paper contains a nice selection of illustrative images.

Any face in the training set can be expressed as a linear combination of a set of eigenfaces



(a) (b) (c) (d)

Figure 5.18 *Face modeling and compression using eigenfaces (Moghaddam and Pentland 1997) © 1997 IEEE: (a) input image; (b) the first eight eigenfaces; (c) image reconstructed by projecting onto this basis and compressing the image to 85 bytes; (d) image reconstructed using JPEG (530 bytes).*

From R. Szeliski, "Computer Vision: Algorithms and Applications, 2nd ed.", Springer (2022)

Figure 4. Three images and their projections onto the face space defined by the eigenfaces of Figure 2. The relative measures of distance from face space are **(a)** 29.8, **(b)** 58.5, **(c)** 5217.4. Images **(a)** and **(b)** are in the original training set.

Image from Turk and Pentland, J. of Cognitive Neuroscience, 3 (1991) 71

Figure 13. **(a)** Partially occluded face image and **(b)** its reconstruction using the eigenfaces.

Image from Turk and Pentland, J. of Cognitive Neuroscience, 3 (1991) 71

Short list of applications of PCA in physics/astrophysics

- Astrophysical object classification (https://www.aanda.org/articles/aa/pdf/2011/11/aa17529-11.pdf)
- Eigenspectra of the VIPERS spectral database (www.vipers.inaf.it)
- Correlation analysis in LHC data (https://arxiv.org/pdf/1708.07113.pdf)
- Analysis of astro-geodetic data (https://ui.adsabs.harvard.edu/abs/2018RoAJ...28..113G/abstract)

A&A 535, A36 (2011)
DOI: 10.1051/0004-6361/201117529
© ESO 2011

Astronomy & Astrophysics

SDSS DR7 superclusters

Principal component analysis

M. Einasto[1], L. J. Liivamägi[1,2], E. Saar[1,3], J. Einasto[1,3,4], E. Tempel[1], E. Tago[1], and V. J. Martínez[5]

[1] Tartu Observatory, 61602 Tõravere, Estonia
e-mail: maret@aai.ee
[2] Institute of Physics, Tartu University, Tähe 4, 51010 Tartu, Estonia
[3] Estonian Academy of Sciences, 10130 Tallinn, Estonia
[4] ICRANet, Piazza della Repubblica 10, 65122 Pescara, Italy
[5] Observatori Astronòmic, Universitat de València, Apartat de Correus 22085, 46071 València, Spain

Received 20 June 2011 / Accepted 26 July 2011

ABSTRACT

Context. The study of superclusters of galaxies helps us to understand the formation, evolution, and present-day properties of the large-scale structure of the Universe.
Aims. We use data about superclusters drawn from the SDSS DR7 to analyse possible selection effects in the supercluster catalogue, to study the physical and morphological properties of superclusters, to find their possible subsets, and to determine scaling relations for our superclusters.
Methods. We apply principal component analysis and Spearman's correlation test to study the properties of superclusters.
Results. We have found that the parameters of superclusters do not correlate with their distance. The correlations between the physical and morphological properties of superclusters are strong. Superclusters can be divided into two populations according to their total luminosity: high-luminosity ones with $L_g > 400 \times 10^{10}\, h^{-2}\, L_\odot$ and low-luminosity systems. High-luminosity superclusters form two sets, which are more elongated systems with the shape parameter $K_1/K_2 < 0.5$ and less elongated ones with $K_1/K_2 > 0.5$. The first two principal components account for more than 90% of the variance in the supercluster parameters. We use principal component analysis to derive scaling relations for superclusters, in which we combine the physical and morphological parameters of superclusters.
Conclusions. The first two principal components define the fundamental plane, which characterises the physical and morphological properties of superclusters. Structure formation simulations for different cosmologies, and more data about the local and high redshift superclusters are needed to understand the evolution and the properties of superclusters better.

PHYSICAL REVIEW C **96**, 064902 (2017)

Principal-component analysis of two-particle azimuthal correlations in PbPb and *p*Pb collisions at CMS

A. M. Sirunyan *et al.**
(CMS Collaboration)

(Received 23 August 2017; published 5 December 2017)

For the first time a principle-component analysis is used to separate out different orthogonal modes of the two-particle correlation matrix from heavy ion collisions. The analysis uses data from $\sqrt{s_{NN}} = 2.76$ TeV PbPb and $\sqrt{s_{NN}} = 5.02$ TeV pPb collisions collected by the CMS experiment at the CERN Large Hadron Collider. Two-particle azimuthal correlations have been extensively used to study hydrodynamic flow in heavy ion collisions. Recently it was shown that the expected factorization of two-particle results into a product of the constituent single-particle anisotropies is broken. The new information provided by these modes may shed light on the breakdown of flow factorization in heavy ion collisions. The first two modes ("leading" and "subleading") of two-particle correlations are presented for elliptical and triangular anisotropies in PbPb and pPb collisions as a function of p_{T} over a wide range of event activity. The leading mode is found to be essentially equivalent to the anisotropy harmonic previously extracted from two-particle correlation methods. The subleading mode represents a new experimental observable and is shown to account for a large fraction of the factorization breaking recently observed at high transverse momentum. The principle-component analysis technique was also applied to multiplicity fluctuations. These also show a subleading mode. The connection of these new results to previous studies of factorization is discussed.

DOI: 10.1103/PhysRevC.96.064902

STATISTICAL ANALYSIS OF ASTRO-GEODETIC DATA THROUGH PRINCIPAL COMPONENT ANALYSIS, LINEAR MODELLING AND BOOTSTRAP BASED INFERENCE

ANDREEA IOANA GORNEA[1,2], ALEXANDRU CALIN[1], PAUL DANIEL DUMITRU[1], DAN ALIN NEDELCU[2], RADU STEFAN STOICA[3]

[1]*Technical University of Civil Engineering Bucharest*
Lacul Tei Bvd. 122 - 124, 020396 Bucharest, Romania, gornea.andreea@gmail.com
[2]*Astronomical Institute of Romanian Academy*
Str. Cutitul de Argint 5, 040557 Bucharest, Romania
[3]*Université de Lorraine, Institut Elie Cartan de Lorraine*
54506 Vandoeuvre-lés-Nancy Cedex, France

Abstract. The paper demonstrates the application of statistical based methodology for the analysis of the vertical deviation angle. The studied data set contains astro-geodetic observations. The Principal Component Analysis and the Multiple Linear Regression models are embedded within a bootstrap procedure, in order to overcome the difficulties related to data correlation, while taking advantage of all the information provided. The methodology is applied on real data. The obtained results indicate that the pressure, the temperature and the humidity are variables that may influence the measure of the vertical deviation.

Key words: vertical deviation – astro-geodetic data – principal component analysis – multi-linear regression – bootstrap – statistics.

The END!