Analiza wariancji i metody klasyfikacyjne

Analysis of variance and classification methods

lecture 2

14 October 2019

Ilona Anna Urbaniak (PK)

Marcin Wolter (IFJ PAN)

e-mail: marcin.wolter@ifj.edu.pl, phone: 12 662 8024

Slides: https://indico.ifj.edu.pl/event/271/

Covariance Matrix

- Let X be a p-variate random vector. The covariance matrix of X is defined as:

$$\Sigma_{XX} = Var(X) = E\{(X-\mu)^T(X-\mu)\}$$

$$= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \dots & Var(X_p) \end{pmatrix}.$$

Where:

$$Var(X_i) = E\{(X_i - \mu(X_i)) \cdot (X_i - \mu(X_i))\}$$

$$Cov(X_i, Y_j) = E\{(X_i - \mu(X_i)) \cdot (Y_j - \mu(Y_j))\}$$

Cross-covariance matrix

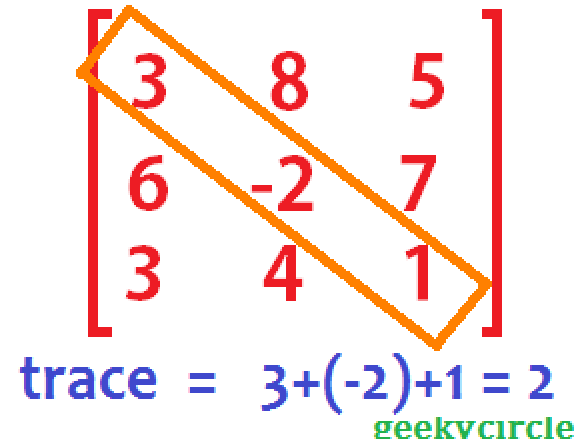
We define the covariance matrix (cross-covariance) between X and Y to be

$$\Sigma_{XY} = Cov(X,Y) = E\{(X-\mu_X)^T(Y-\mu_Y)\}$$
$$= \begin{pmatrix} Cov(X_1,Y_1) & Cov(X_1,Y_2) & \dots & Cov(X_1,Y_m) \\ Cov(X_2,Y_1) & Cov(X_2,Y_2) & \dots & Cov(X_2,X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p,Y_1) & Cov(X_p,Y_2) & \dots & Cov(X_p,Y_m) \end{pmatrix}.$$

Some relations:

$$Cov(AX,BY) = ACov(X,Y)B^T$$
$$Cov(X+a,Y+b) = Cov(X,Y)$$

Trace and variance



Let A = (a_{ij}) be a square matrices of dimension d × d.

- The trace of A is the sum of its diagonal elements:

$$tr(A) = \sum_i a_{ii}$$

- The mean is the best constant predictor of X in terms of the MSE (shown already at previous lecture):

$$E(X) = \arg\min_{c \in \mathbb{R}^p} E\|X - c\|^2$$

- The **total variance of X** is defined as the MSE of the mean:

$$\mathrm{E}\ ||\ \mathrm{X} - \mathrm{E(X)}\ ||^2 = \sum_{i=1}^{N} E(X_i - E(X_i))^2 = {}^{N}_{i=1} Var(X_i) = tr(Var(X))$$

- The **total variance of X** measures the overall variability of the components of X around the mean E(X). Commonly used measure of variability is the standard deviation.

Quadratic Forms

Let A be a symmetric matrix and x a vector.

Definition:

A quadratic form is written as:

$$x^T A x = \sum_i \sum_j a_{ij} x_i x_j$$

Note: it's a quadratic function of x.

- As a function of A, $Var(AX) = A\, Var(X) A^T$ which is a quadratic form in A.
- Quadratic forms are very common in multivariate analysis.
- Example: Chi-squared test is a quadratic form.

Quadratic forms

Forms

$$h_1(x_1, x_2, x_3) = -x_1^2 + 2x_1x_2 + x_2^2 - 4x_2x_3$$

and

$$h_2(x_1, x_2, x_3, x_4) = -x_1^2 + 2x_1x_2 + x_2^2 - 4x_2x_3$$

are quadratic forms given by matrices:

$$A_{h_1} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 1 & -2 \\ 0 & -2 & 0 \end{bmatrix} \quad \text{oraz} \quad A_{h_2} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & 1 & -2 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

These are NOT quadratic forms (think why?):

$$g_1(x_1, x_2) = x_1^2 + 2x_1x_2 + x_2 \quad \text{oraz} \quad g_2(x_1, x_2) = x_1^2 + x_2^2 + 1$$

The matrix A is a square matrix and can always be written in a symmetric form.
More general: in the complex space it is a Hermitian matrix (or self-adjoint matrix): complex square matrix that is equal to its own conjugate transpose.

Positive/negative semi-definite and positive/negative definite matrix

- A is a positive definite form, if:

$$x^T A x > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\};$$

- negative definite, if:

$$x^T A x < 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\};$$

- semi-positive definite, if:

$$x^T A x \geq 0, \quad \forall x \in \mathbb{R}^n;$$

- semi-negative definite, if:

$$x^T A x \leq 0, \quad \forall x \in \mathbb{R}^n;$$

- If none of the above, the quadratic form is not definite.

Example

Let's take the quadratic form:

$$h_1(x_1, x_2, x_3) = -x_1^2 + 2x_1x_2 + x_2^2 - 4x_2x_3.$$

Since $h_1(1,0,0) = -1$ and $h_1(0,1,0) = 1$, the quadratic form $h_1$is undefinite.

Quadratic form

$$h(x_1, x_2) = x_1^2 + 2x_2^2$$

is positive definite, the form:

$$g(x_1, x_2, x_3) = x_1^2 + 2x_2^2$$

is semi-positive definite (think why?).

Note that covariance matrices have the following properties:

1)Every covariance matrix is a positive semi-definite matrix.

2)Every positive semi-definite matrix is a covariance matrix.

Methods of testing whether a matrix is positive/negative definite

Sylvester criterion:

Quadratic form $h(x) = x^T A x$, where $A = A^T \in \mathbb{R}^{n \times n}$, is:

1) Positive (Negative) Definite when and only when all the leading minors of matrix A are positive (negative):

$$D_j = \begin{vmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{j1} & \cdots & a_{jj} \end{vmatrix} > (<) 0, \quad (j = 1, \ldots, n);$$

Sylvester criterion – an example

For a quadratic form

$$h\left(x_1, x_2, x_3\right) = 3x_1^2 + 2x_1x_2 + x_2^2 - 2x_1x_3 + 2x_3^2$$

we have:

$$h\left(x_1, x_2, x_3\right) = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 3 & 1 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Since:

$$D_1 = 3 > 0, \quad D_2 = \begin{vmatrix} 3 & 1 \\ 1 & 1 \end{vmatrix} = 2 > 0, \quad D_3 = \begin{vmatrix} 3 & 1 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & 2 \end{vmatrix} = 3 > 0,$$

therefore a quadratic form h is positive definite.

Determinant

$$\begin{vmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{vmatrix} = a\begin{vmatrix} f & g & h \\ j & k & l \\ n & o & p \end{vmatrix} - b\begin{vmatrix} e & g & h \\ i & k & l \\ m & o & p \end{vmatrix} + c\begin{vmatrix} e & f & h \\ i & j & l \\ m & n & p \end{vmatrix} - d\begin{vmatrix} e & f & g \\ i & j & k \\ m & n & o \end{vmatrix}.$$

The simplest way to express the determinant is by considering the elements in the top row and the respective minors; starting at the left, multiply the element by the minor, then subtract the product of the next element and its minor, and alternate adding and subtracting such products until all elements in the top row have been exhausted.

Eigenvalues and eigenvectors

Let vector $\mathbf{v} > \mathbf{0}$ and let **A** be a d × d matrix.

v is an eigenvector with eigenvalue λ when

$$\mathbf{Av} = \lambda\mathbf{v}.$$

- It's typical to normalize the eigenvector to have length 1 (or have it's entries sum to 1).
- Matrix **A** has at most *d* distinct eigenvalues (think about why).
- Eigenvectors with distinct eigenvalues are orthogonal, i.e. $\mathbf{v}_1^T\mathbf{v}_2=0$

- If **A** is a positive definite matrix, then:
  - All of its eigenvalues are real-valued and positive.
  - Its inverse is also positive definite.

Eigendecomposition (spectral decomposition)

- Matrix **A** has n linearly independent eigenvectors $v_1, v_2, \ldots, v_n$ with associated eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.
- Define square matrix Q whose columns are the n linearly independent eigenvectors of A: $Q = [\,v_1 \quad v_2 \quad \cdots \quad v_n\,].$
- Since each column of Q is an eigenvector of A: $AQ = [\,\lambda_1 v_1 \quad \lambda_2 v_2 \quad \cdots \quad \lambda_n v_n\,].$
- Define matrix Λ: $\Lambda_{ii} = \lambda_i$ and $\Lambda_{ij} = 0$ for $i \neq j$, than $AQ = Q\Lambda.$
- $A = Q\,\Lambda\,Q^{-1}$ (multiplying both sides by Q^{-1})
- Or $Q^{-1} A\, Q = \Lambda$
- Matrix **A** can be decomposed into a **matrix composed of its eigenvectors, a diagonal matrix with its eigenvalues along the diagonal, and the inverse of the matrix of eigenvectors**.

  This is called the **eigendecomposition**. Matrix **A** is **diagonalizable**.

Eigendecomposition

- The eigen (spectral) decomposition allows some operations with positive definite matrices to be computed more easily:

  $A^{-1} = P\,\Lambda^{-1}\,P^{T}$.

  $A^{1/2} = P\,\Lambda^{1/2}\,P^{T}$.

- Example:

$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$ can be decomposed into $\begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}, \qquad [c, d] \in \mathbb{R}$

Eigendecomposition

Scalar $\lambda \in \mathbb{F}$ is called an **eigenvalue** of matrix $A \in \mathbb{F}^{n \times n}$ if exists a non-zero vector $v \in \mathbb{F}^n$, such that:

$$Av = \lambda v;$$

vector v is called an **eigenvector** of an eigenvalue λ.

All eigenvalues of matrix A is a spectrum of A and is denoted as $\sigma(A)$.

For matrix $A \in \mathbb{F}^{n \times n}$ the following conditions are equivalent:

(a) λ*is an eigenvalue of* A*;*

(b) system of equations $(A - \lambda I) v = 0$ has a non-zero solution;

(c) $\det(A - \lambda I) = 0$.

Eigendecomposition

For any matrix $A \in F^{n \times n}$ **det(A−λI)** is a polynomial of degree **n** (characteristic polynomial). The roots of this polynomial are the the eigenvalues.

Example:

Calculate eigenvectors and eigenvalues of matrix:

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ -2 & -2 & -1 \end{bmatrix}.$$

Since $\varphi_A(\lambda) = \det(A - \lambda I) = -(1-\lambda)(2-\lambda)(1+\lambda)$, therefore matrix A has three different eigenvalues: $\lambda_1 = -1, \lambda_2 = 1, \lambda_3 = 2$.

The equation *det(A−λI)=−(1−λ) (2−λ) (1 +λ) = 0* is a characteristic polynomial.

For each eigenvalue we find an eigenvector:

- for $\lambda_1 = -1$ we have:

$$\begin{bmatrix} 2 & 2 & 0 \\ 0 & 3 & 0 \\ -2 & -2 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2x+2y \\ 3y \\ -2x-2y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

so we get $(x, y, z) = (0, 0, t)$, $t \in \mathbb{R}$; an example eigenvector $v_{\lambda_1} = (0, 0, 1)^T$;

- for $\lambda_2 = 1$ we get:

$$\begin{bmatrix} 0 & 2 & 0 \\ 0 & 1 & 0 \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2y \\ y \\ -2x-2y-2z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

so we get $(x, y, z) = (t, 0, -t)$, $t \in \mathbb{R}$; example eigenvector is $v_{\lambda_2} = (1, 0, -1)^T$;

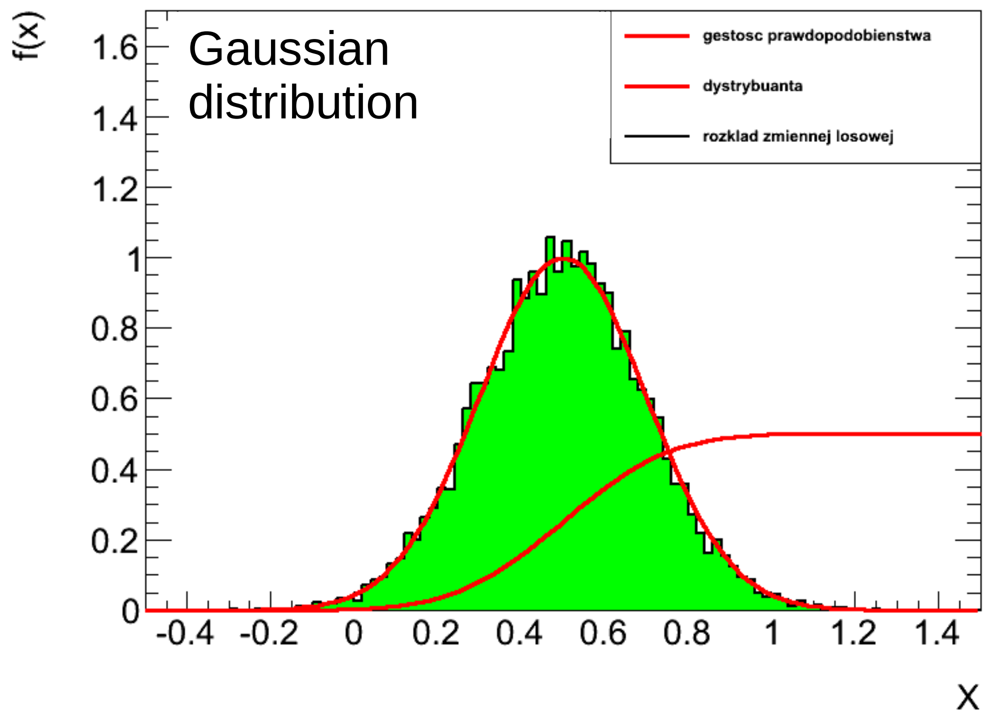
- for $\lambda_3 = 2$ we get:

$$\begin{bmatrix} -1 & 2 & 0 \\ 0 & 0 & 0 \\ -2 & -2 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -x+2y \\ 0 \\ -2x-2y-3z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

we get $(x, y, z) = (2t, t, -2t)$, $t \in \mathbb{R}$; eigenvector $v_{\lambda_3} = (2, 1, -2)^T$.

Multivariate Normal Distribution

- MVN is generalization of univariate normal distribution (gausian).



Johann Karl Friedrich Gauss (1777 – 1855)

Normal distribution (Gaussian):

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Multivariate Normal Distribution

- We assume that the population mean is μ = E(X) and
  Σ = Var(X) = E[(X − μ)(X − μ)^T], then:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

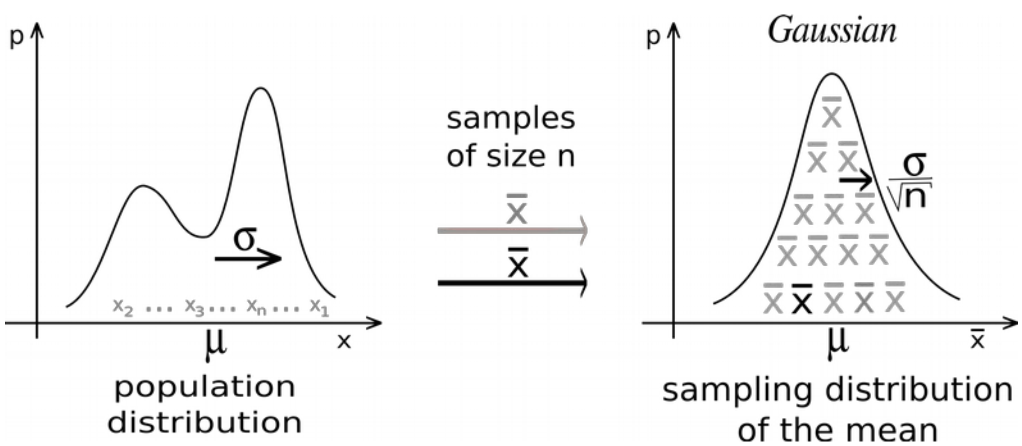
and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

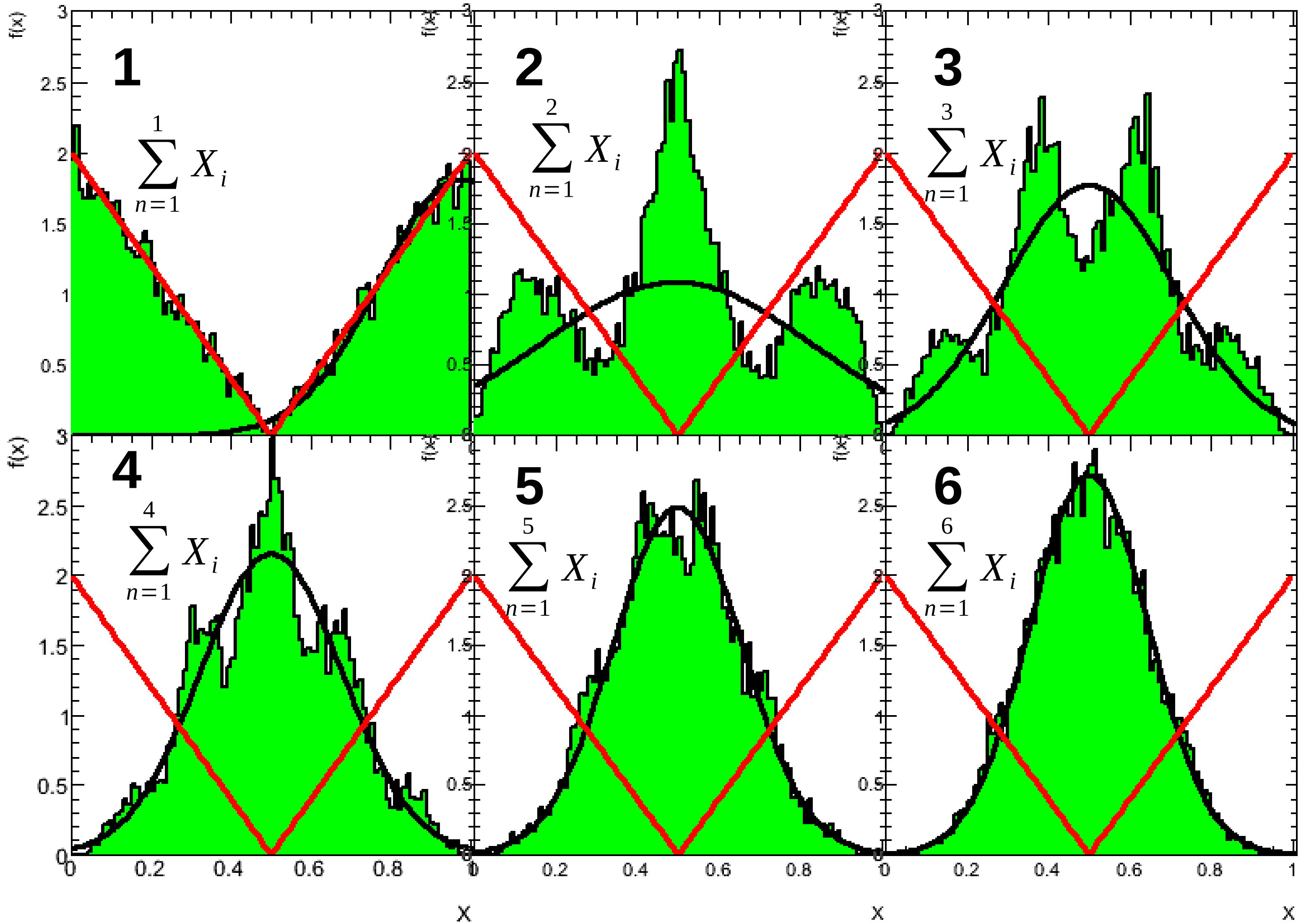
Central limit theorem (CLT)

- One of the **most important theorems** of the statistics – this is why we observe in nature mostly Gaussian distributions.

Let $\{X_1, \ldots, X_n\}$ be a random sample of size n of independent and identically distributed random variables drawn from a distribution of expected value given by μ and finite variance given by σ^2. CLT states that as n gets larger, the distribution of the difference between the sample average and its limit μ, when multiplied by the factor √n, approximates the normal distribution with mean 0 and variance σ^2. For large enough n, the distribution of average is close to the normal distribution with mean μ and variance σ^2/n.



Central limit theorem (CLT)

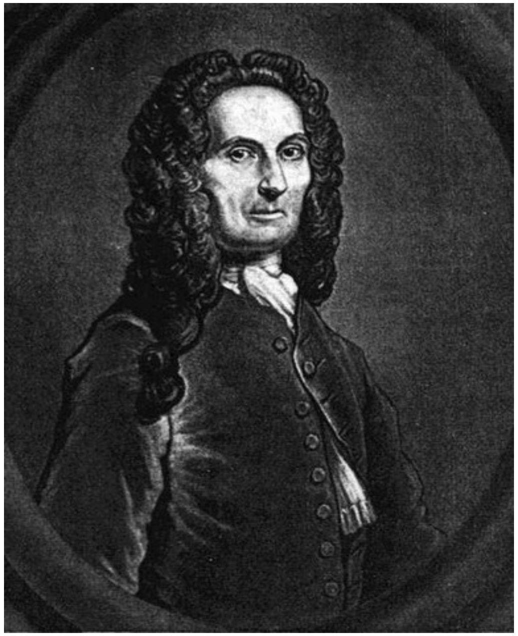


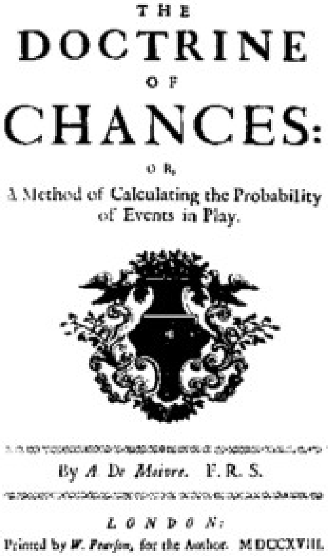
Central limit theorem (CLT)

For the first time CLT for binomial distributions was postulated in the second edition of the „*The Doctrine of Chances*” by Abraham de Moivre’a, published in 1738. It was forgotten for over 80 years, and in 1812 Pierre-Simon Laplace proved CLT for the binomial distributions.

CLT in the version of Lindeberg & Levy was published in 1920’ties, however independently it was proven earlier by Aleksandr Lyapunov in 1901.

Abraham de Moivre (1667 – 1754)




THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability of Events in Play.

By A. De Moivre. F. R. S.

LONDON:
Printed by W. Pearson, for the Author. MDCCXVIII.

Proof of classical CLT

- **Characteristic function** of a random variable defines its probability distribution (if a random variable admits a probability density function, then the characteristic function is the Fourier transform of the probability density function).
- Assume $\{X_1, \ldots, X_n\}$ are independent and identically distributed random variables, each with mean μ and finite variance σ^2. The sum $X_1 + \ldots + X_n$ has mean nμ and variance $n\sigma^2$. The random variable:

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i, \qquad Y_i = \frac{X_i - \mu}{\sigma}$$

- The characteristic function of Zn is given by:

$$\varphi_{Z_n}(t) = \varphi_{\sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i}(t) = \varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) \varphi_{Y_2}\left(\frac{t}{\sqrt{n}}\right) \cdots \varphi_{Y_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n,$$

- since all of the Y_i are identically distributed (zero mean, σ=1).

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right), \quad \left(\frac{t}{\sqrt{n}}\right) \to 0$$

Taylor's theorem

“*An Introduction to Stochastic Processes in Physics*”, *Don S. Lemons*

Proof of classical CLT

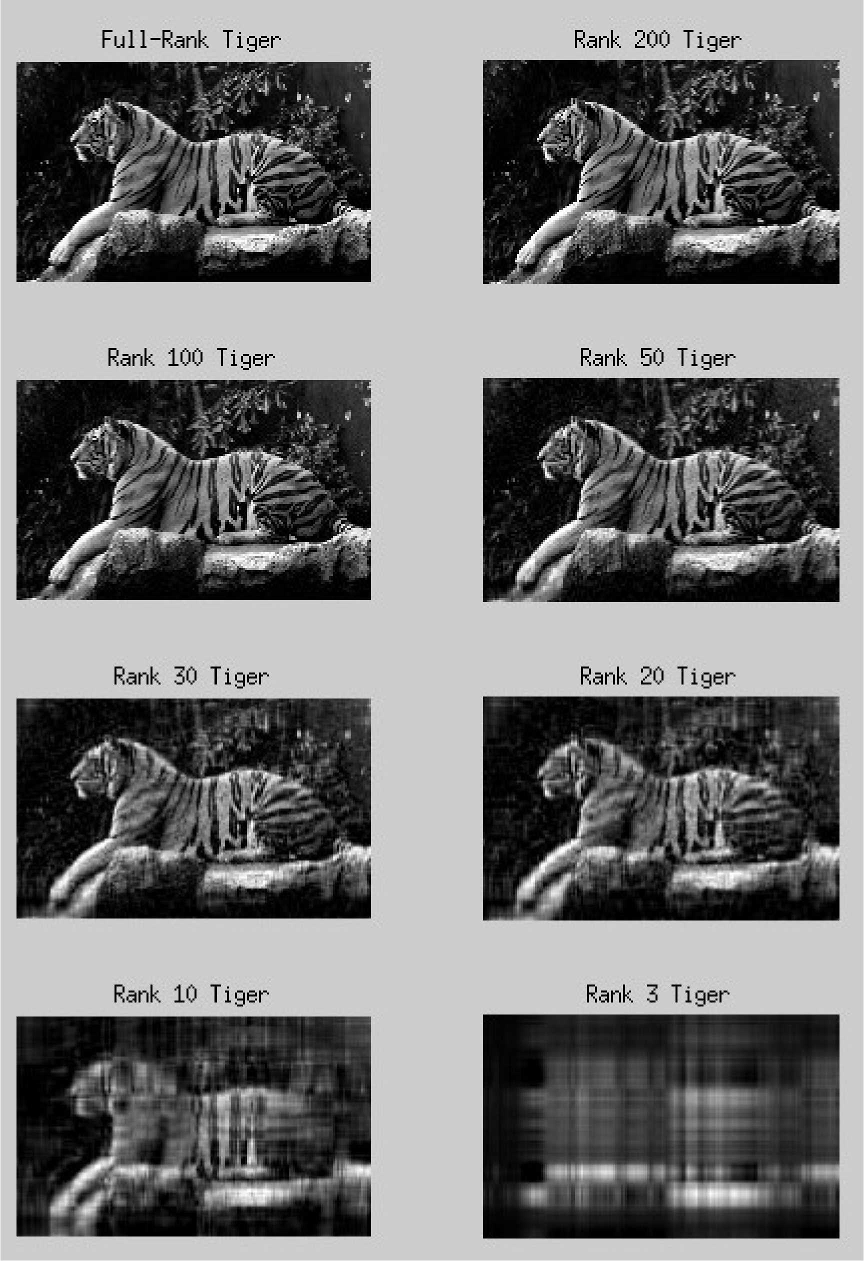
- Since $e^x = \lim(1 + x/n)^n$, the characteristic function of Z_n equals:

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \rightarrow e^{-\frac{1}{2}t^2}, \quad n \rightarrow \infty.$$

- All of the higher order terms vanish in the limit $n \rightarrow \infty$.

- The right hand side equals the characteristic function of a standard normal distribution N(0,1) → in the limit of $n \rightarrow \infty$ the $Z_n \rightarrow$ N(0,1)

Singular Value Decomposition SVD



SVD

Singular value decomposition is a method of decomposing a matrix into three other matrices:

$$A = USV^T$$

Where:

$$AA^TU \quad = \quad US^2$$

A is an m × n ma

U is an m × n orthogonal matrix

S is an n × n diagonal matrix

V is an n × n orthogonal matrix

Orthogonal matrices: $$U^TU = VV^T = I$$

https://blog.statsbot.co/singular-value-decomposition-tutorial-52c695315254

SVD

$$A = USV^T$$

$$a_{ij} = \sum_{k=1}^{n} u_{ik} s_k v_{jk}$$

The variables, $\{s_i\}$, are called singular values and are normally arranged from largest to smallest:

$$s_{i+1} \leq s_i$$

The columns of U are called left singular vectors, while those of V are called right singular vectors.

SVD

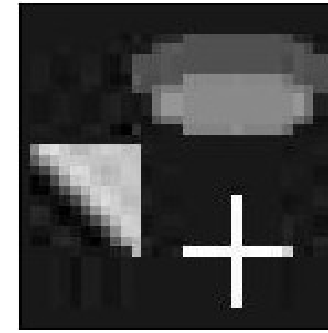
Using orthogonality property we get:

$$A = USV^T \quad \Rightarrow \quad \begin{aligned} AA^TU &= US^2 \\ A^TAV &= VS^2 \end{aligned}$$

The standard procedure (or eigenvalue calculator) can be used to solve these equations and find the U, V and S^2.

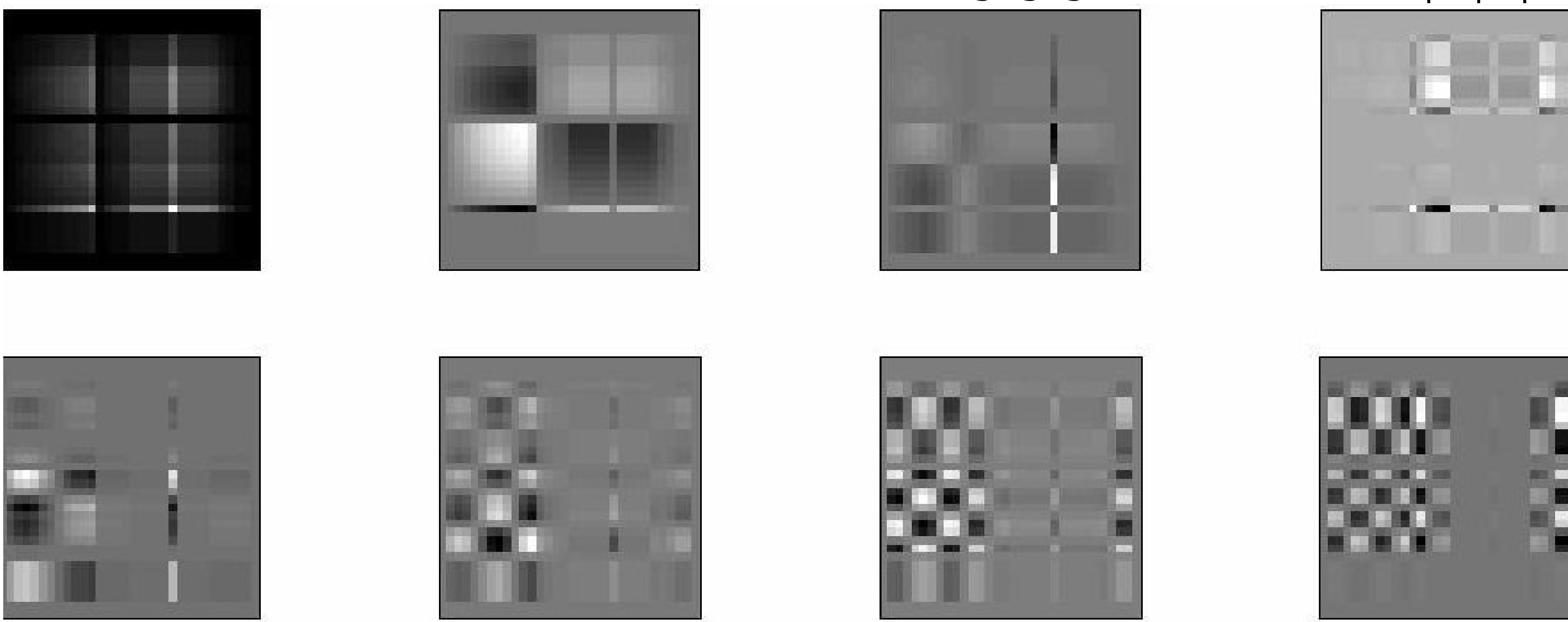
Applications of the SVD

the SVD of a 32-times-32 digital image *A is computed*



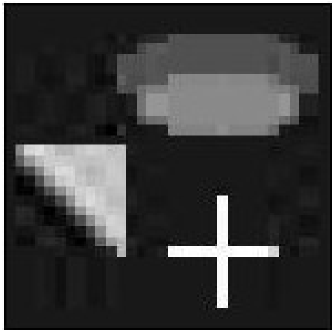
the activities are lead by Prof. Per Christian Hansen.

$\sigma_1 u_1 v_1^T$ $\sigma_2 u_2 v_2^T$ $\sigma_3 u_3 v_3^T$ $\sigma_4 u_4 v_4^T$



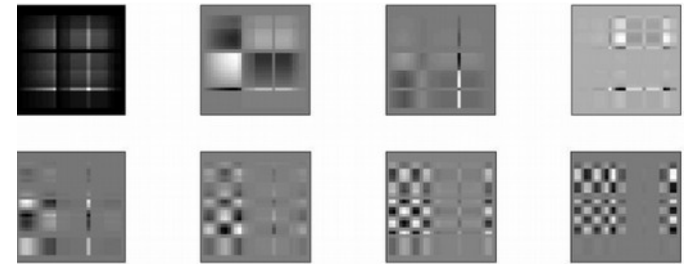
$\sigma_5 u_5 v_5^T$ $\sigma_6 u_6 v_6^T$ $\sigma_7 u_7 v_7^T$ $\sigma_8 u_8 v_8^T$

Applications of the SVD

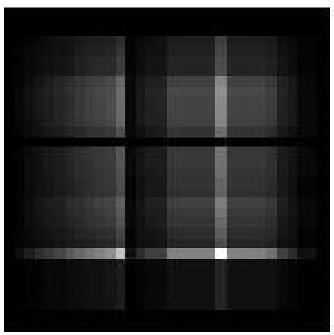
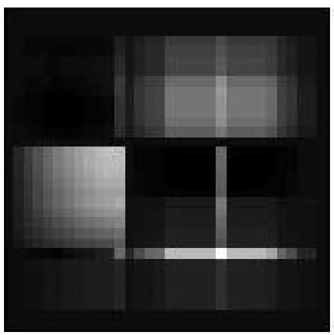
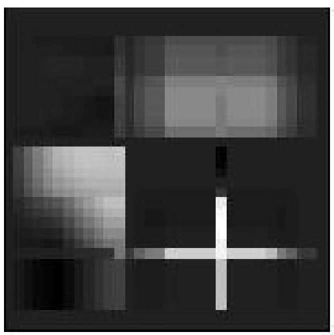
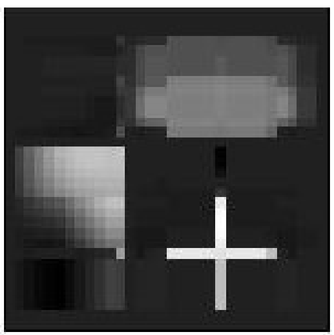


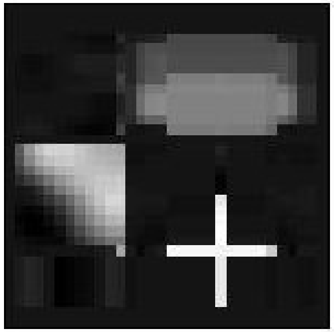
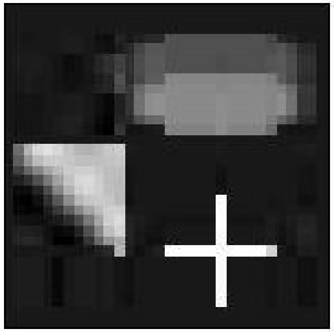
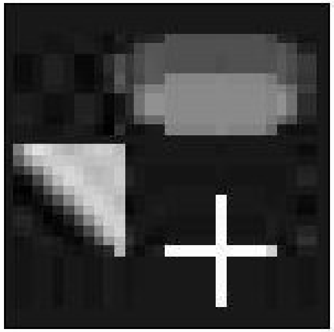
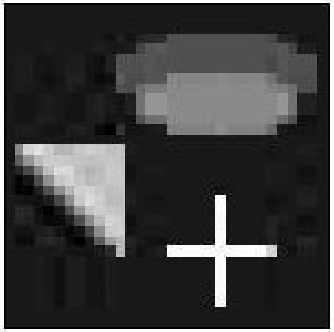
$$A_s = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_s u_s v_s^T$$

$$s \le r$$



A_1 A_2 A_3 A_4

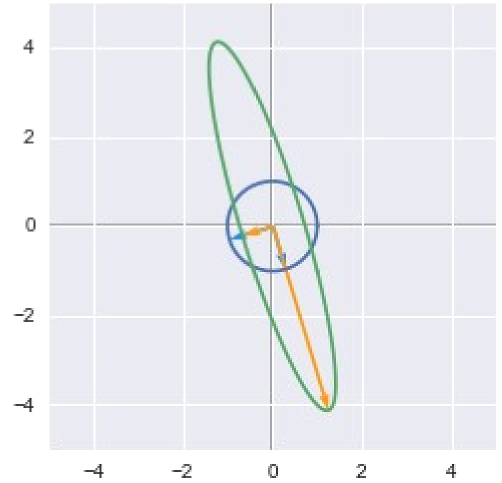
A_5 A_6 A_7 A_8

G. Strang " at first you see nothing, and suddenly you recognize everything."

Exercises

1) Write a code (python, R) which performs eigendecomposition of a given symmetric matrix.

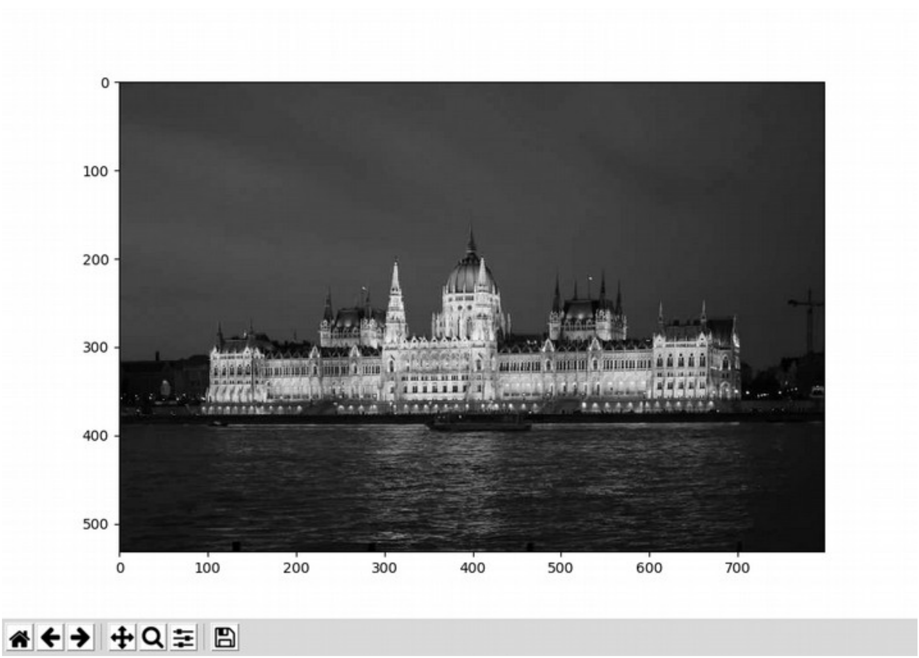
- Code it yourself
- Try maybe numpy.roots(p) to find the roots
- Test against numpy.linalg.eig()
- If you want to play more try to visualize the linear transformation (see “BONUS: visualizing linear transformations” in https://hadrienj.github.io/posts/Deep-Learning-Book-Series-2.7-Eigendecomposition/)

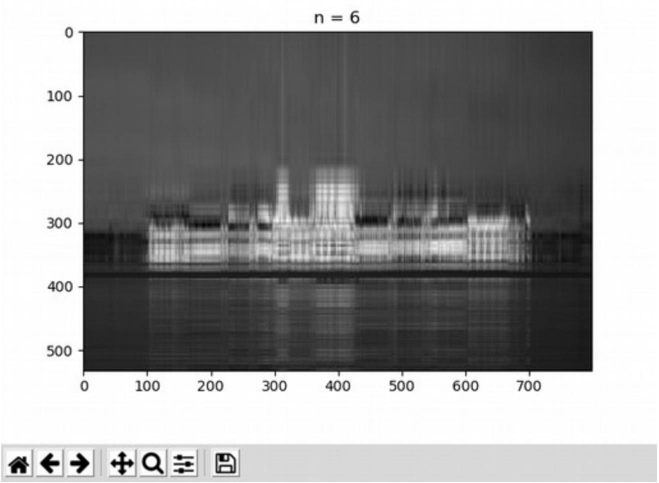


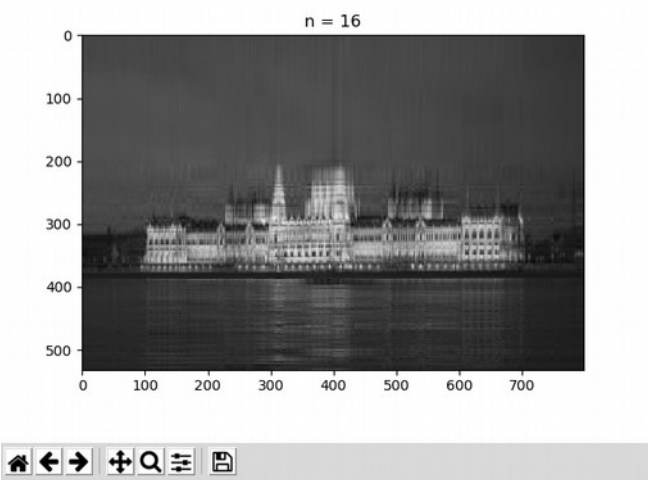
Exercises

- Apply Singular Value Decomposition to a photograph:

  numpy.linalg.svd







Exercises

- Write a script showing, that the CLT works :)