SRF L5

Statistical Data Analysis

Objective

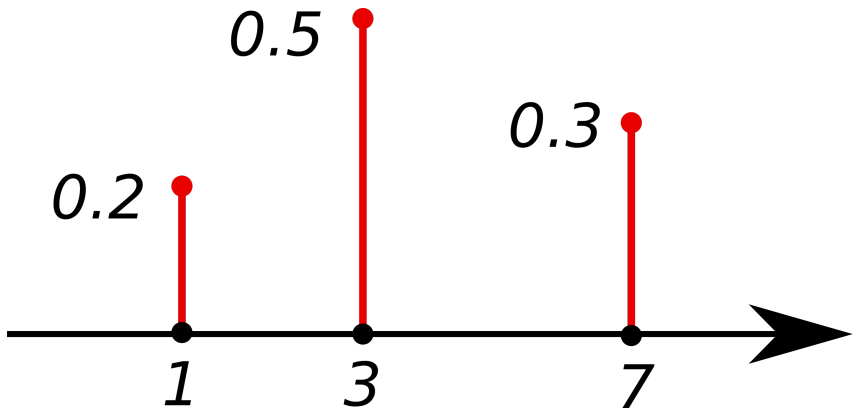
- Statistical Data Analysis: mean, standard deviation and covariance (correlation between pixels), used for classification and recognition

Definitions

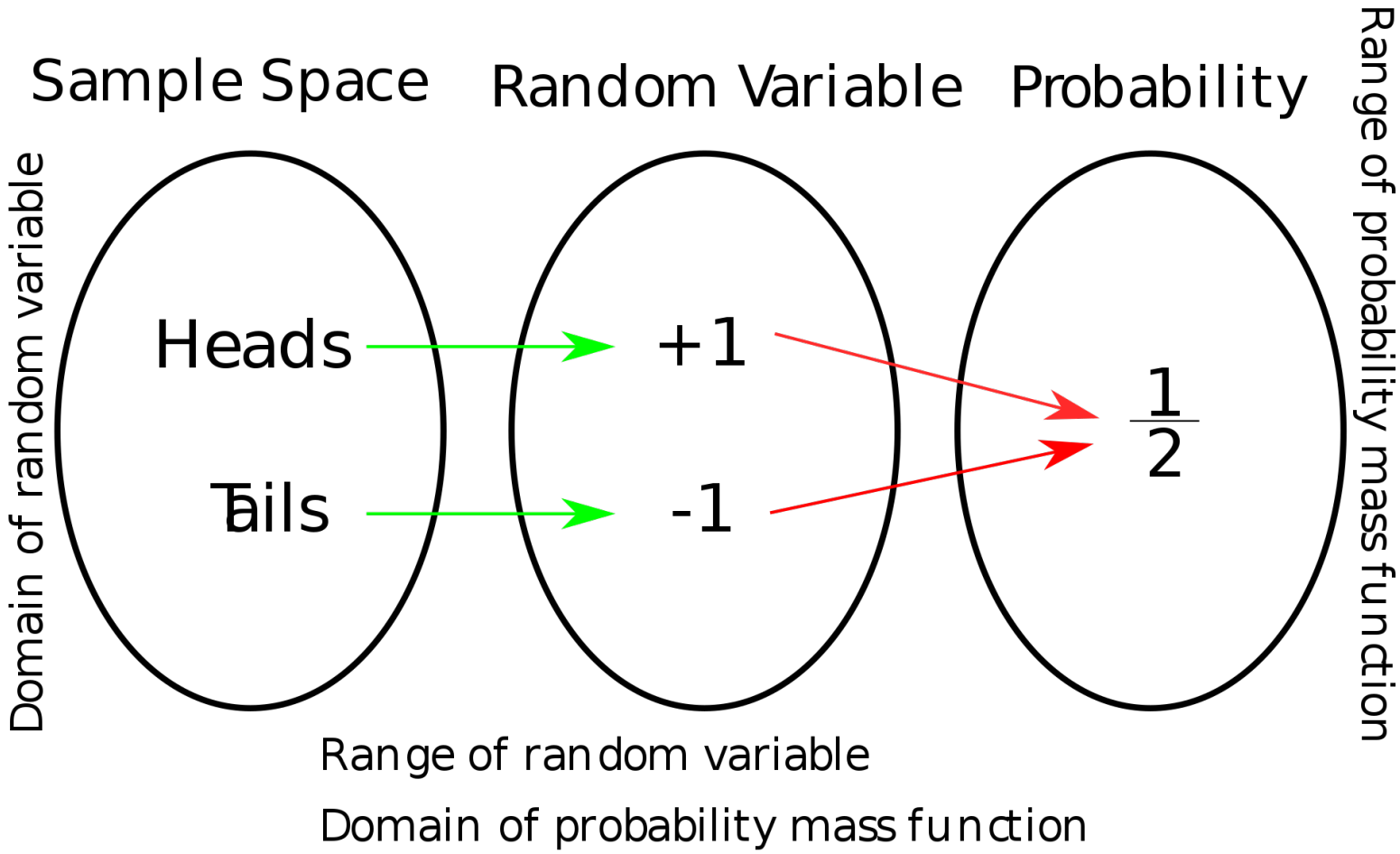
- In probability theory, the **sample space** of a random experiment is the set of possible outcomes e.g. For tossing one coin, the corresponding sample space would be {(head), (tail)}.
- A possible outcome ζ (head).
- A **random variable X** is a function $X(\zeta)$ that assigns a real number for for an outcome ζ **from the sample space**
- $X(\zeta) \in \mathbb{R}$
- X can be discrete(the resulting number after rolling a dice) or continuous (the weight of an individual)

Definitions

- A **random variable vector** X is a function that assigns a vector of real numbers to each outcome ζ in the sample space $\boldsymbol{X} = [X_1, X_2, \ldots, X_N]$
- For a discrete random variable X, we can compute the probability that X equals x: P(X=x)
- **The Probability mass function (PMF) is** a function that gives the probability that a **discrete** random variable is exactly equal to some value.
- p : $\mathbb{R}$ -> [0, 1]
- $p(x_i) = P(X=x_i)$



Definitions

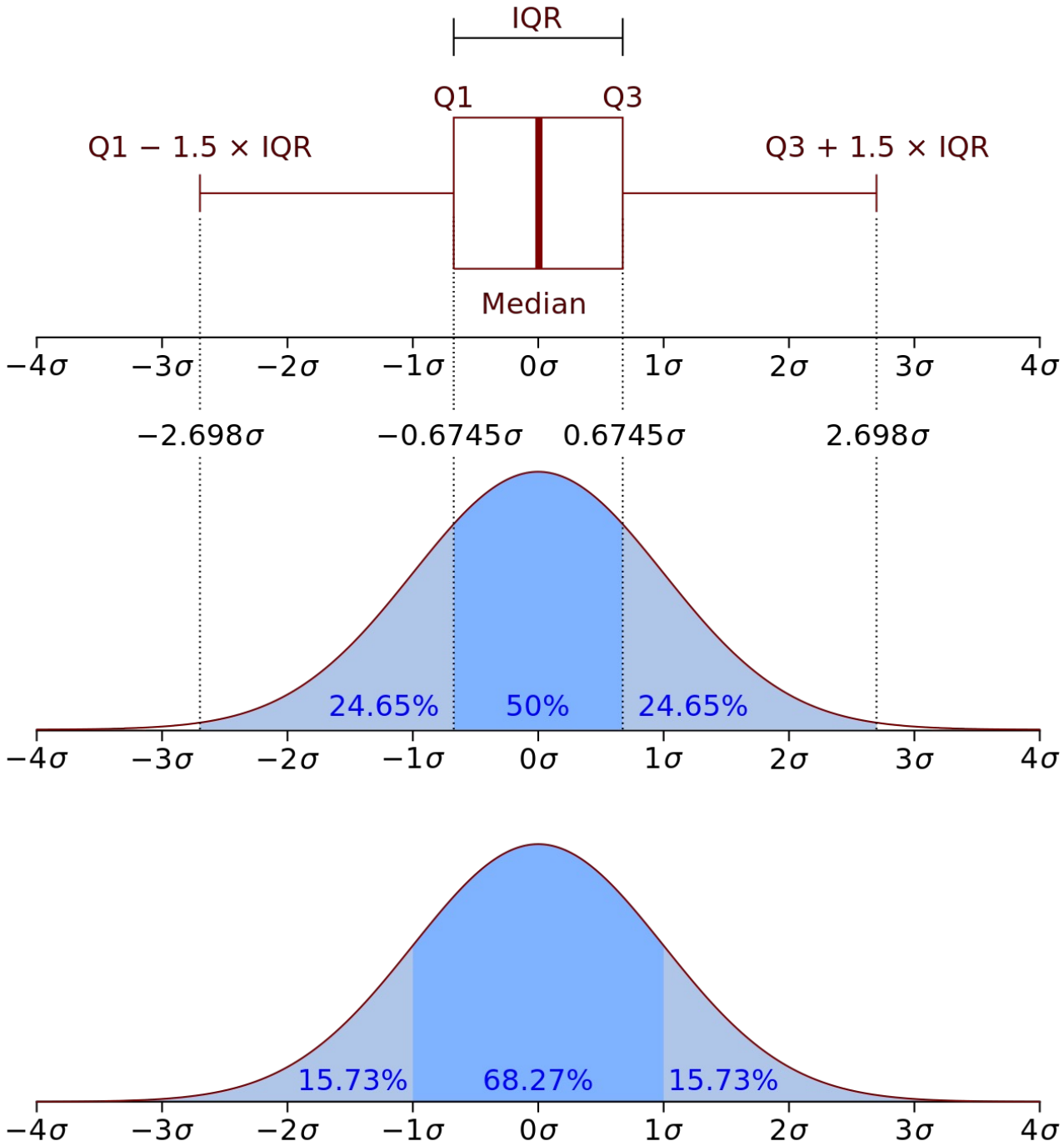


Definitions

- In the case of **continous random variables** X, the probability of X to be equal to x: P(X=x) = 0, therefore we will compute the probability that X is in [a, b]: P(a <= X <= b)
- **Probability density function (PDF) –** is used to specify the probability of the random variable falling *within a particular range of values*, as opposed to taking on any one value.
- P(a <= X <= b) = $\int_a^b f_X(x)dx$
- $f_X(x)dx$ represents the density function

Definitions

- P(a <= X <= b) = $\int_a^b f_X(x)dx$
- P(a <= X <= b) is equal to the area under the curve $f_X(x)$
- **Example:** the PDF of a normal (Gaussian) density function



Statistical Characterization of Random variables

- **Mean (expectation):** represents the center of mass of a density

$$E[X] = \mu = \int_{-\infty}^{\infty} x f_x(x) dx$$

- **Variance**: represents the spread about the mean

$$VAR[X] = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_x(x) dx$$

- **Standard deviation**: The square root of the variance.

$$STD[X] = VAR[X]^{1/2}$$

Statistical Characterization of Random Vectors

- **Mean vector**

$$E[\boldsymbol{X}] = \left[E[X_1], E[X_2], \dots, E[X_N]\right] = [\mu_1, \mu_2, \dots, \mu_N] = \boldsymbol{\mu}$$

- **Covariance matrix**

$$COV[\boldsymbol{X}] = \boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)^T] & \dots & E[(X_1 - \mu_1)(X_N - \mu_N)^T] \\ \dots & \dots & \dots \\ E[(X_N - \mu_N)(X_1 - \mu_1)^T] & \dots & E[(X_N - \mu_N)(X_N - \mu_N)^T] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \dots & c_{1N} \\ \dots & \dots & \dots \\ c_{N1} & \dots & \sigma_N^2 \end{bmatrix}$$

- σi is the standard deviation of X_{I}
- **Correlation coefficient**

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j}$$

Statistical Characterization of Random Vectors

Covariance matrix

$\boldsymbol{X} = [X_1, X_2, \ldots, X_N]$ random variable vector

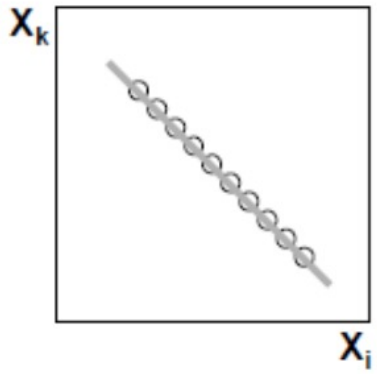
$$COV[\boldsymbol{X}] = \boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)^T] & \ldots & E[(X_1 - \mu_1)(X_N - \mu_N)^T] \\ \ldots & \ldots & \ldots \\ E[(X_N - \mu_N)(X_1 - \mu_1)^T] & \ldots & E[(X_N - \mu_N)(X_N - \mu_N)^T] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \ldots & c_{1N} \\ \ldots & \ldots & \ldots \\ c_{N1} & \ldots & \sigma_N^2 \end{bmatrix}$$

- The covariance matrix indicates the **tendency of each pair of features** (elements in a random vector) **to vary together**, i.e., to co-vary.
- The covariance matrix can be used for feature selection (e.g. we have a dataset, compute a very large number of features (color, gradient, edges etc.) on the images, all features might not be useful for building machine learning models, therefore we perform feature selection and dimensionality reduction. Features that have high correlation have the same effect on the output, so we can drop one of the two features.)

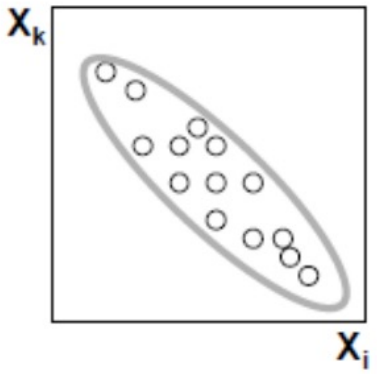
Statistical Characterization of Random Vectors

The next figures represent the correlation charts between two features, X_i and X_k.



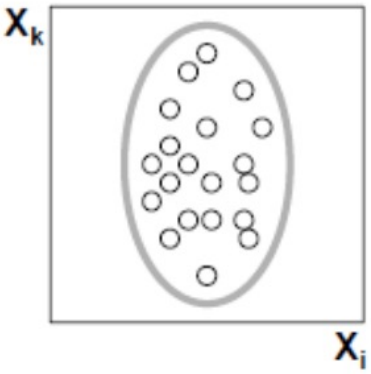
$C_{ik}=-\sigma_i\sigma_k$
$\rho_{ik}=-1$

Strong negative correlation: if Xi increases, Xk decreases



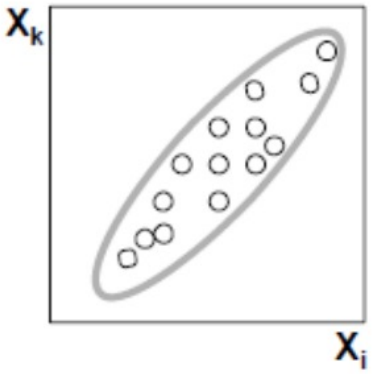
$C_{ik}=-\frac{1}{2}\sigma_i\sigma_k$
$\rho_{ik}=-\frac{1}{2}$

Medium negative correlation: if Xi increases, Xk decreases



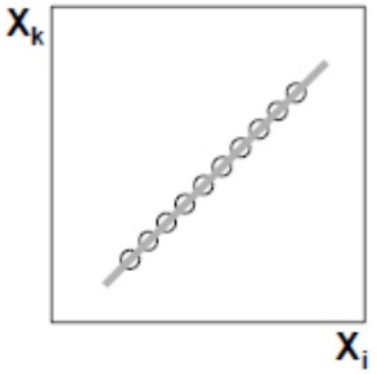
$C_{ik}=0$
$\rho_{ik}=0$

No correlation



$C_{ik}=+\frac{1}{2}\sigma_i\sigma_k$
$\rho_{ik}=+\frac{1}{2}$

Medium positive correlation: if Xi increases, Xk increases



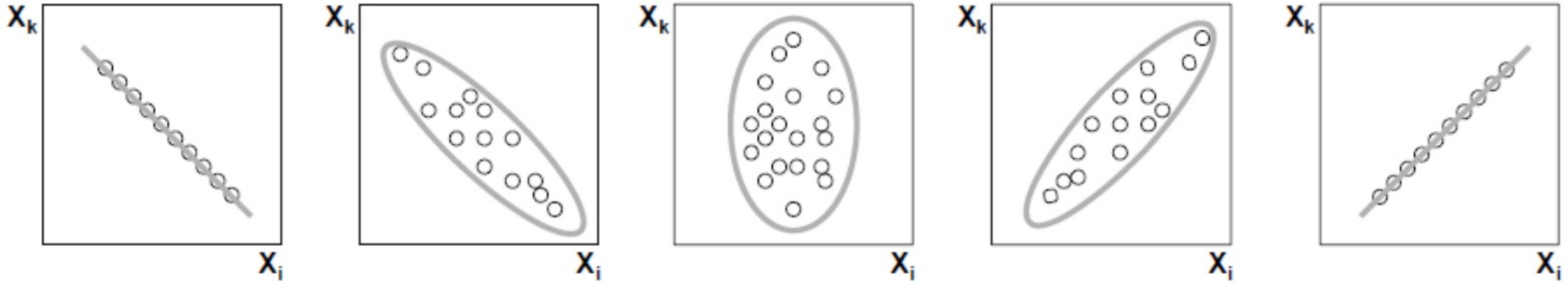
$C_{ik}=\sigma_i\sigma_k$
$\rho_{ik}=+1$

Strong Positive correlation: if Xi increases, Xk increases

- The covariance has several important properties:
  - - If Xi and Xk tend to increase together, then $c_{ik} > 0$
  - - If Xi tends to decrease when Xk increases, then $c_{ik} < 0$
  - - If Xi and Xk are uncorrelated, then $c_{ik} = 0$

Statistical Characterization of Random Vectors

The next figures represent the correlation charts between two features, X_i and X_k.



- The **correlation coefficient p** between two features X_k and X_i represents the **normalized covariance** between the two features and measures the direction of the linear relationship between the two features
- The correlation coefficient takes values in [-1, 1]. The closer is to +1 or -1, the more closely the two features are related.
- We can use the covariance when the two features have the same scale, and use the correlation when they have different scales.

Practical Issues

In this lab session you are required to study the correlation between pixels belonging to human faces. You are given $p=400$ images that contain human faces. The figure below shows a montage of all the input images:

Practical Issues

Let $\boldsymbol{I}$ be the feature matrix which will hold all the intensity values from the image set. $\boldsymbol{I}$ is of dimension $p \times N$, where p is the number of images and N is the number of pixels in each image. The k^{th} row contains all the pixel intensities from the k^{th} image in row-major order. Example for 3x3 matrix:

$$\begin{bmatrix} A_{00} & A_{01} & A_{02} \\ A_{10} & A_{11} & A_{12} \\ A_{20} & A_{21} & A_{22} \end{bmatrix} \rightarrow [A_{00}, A_{01}, A_{02}, A_{10}, A_{11}, A_{12}, A_{20}, A_{21}, A_{22}]$$

Each image in the set has the dimension of *19x19* pixels. The interpretation of the feature matrix $\boldsymbol{I}$ is that each row holds a sample for the N dimensional random variable $\boldsymbol{X}$ which is drawn from the distribution underlying the dataset.

Your task will be to compute the covariance matrix of the given set of images and to study how different features vary with respect to each other.

Practical Issues

The mean value of a feature located at position i in the image is:

$$\mu_i = \frac{1}{p}\sum_{k=1}^{p} I_{ki}$$

Where I_{ki} represents the value of feature i in image k.

The mean is a vector of 19x19=361 values

The standard deviation of a feature i is:

$$\sigma_i = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(I_{ki} - \mu_i)^2}$$

The standard deviation is a vector of 19x19=361 values

The elements of the covariance matrix, c_{ij} can be computed by:

$$c_{ij} = \frac{1}{p}\sum_{k=1}^{p}(I_{ki} - \mu_i)(I_{kj} - \mu_j)$$

The covariance matrix is a matrix of 361x361 values

The correlation coefficient is:

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j}$$

Note that $c_{ii} = \sigma_i^2$ and $\rho_{ii} = 1$.

Activitate practică

- Ex. 1 . Încărcați cele 400 de imagini și stocați valorile de intensitate din fiecare imagine ca și rânduri in matricea de trăsături ***I***.

  Feature matrix **I** has the size 400 x 361 (HxW). 361 = 19x19

  Each column of the matrix represents a feature vector (in our case the feature vector contains only the intensities). For ex. col j wil have all intensities of pixels at position j in all the images

Activitate practică

1. Ex 2. Calculați vectorul cu valorile medii si salvați-l într-un fișier text de tip CSV (comma separated values). Se scriu componentele cu virgule între ele și se salvează într-un fișier text cu extensia CSV.

The mean value of a feature located at position i in the image is:

$$\mu_i = \frac{1}{p}\sum_{k=1}^{p} I_{ki}$$

Where I_{ki} represents the value of feature i in image k.

The mean values will be computed as the mean of each column.

- The vector of means will have 361 (19x19) values.
- p – is the number of samples (400)

Activitate practică

1. Ex 3. Calculați matricea de covarianță si salvați-o într-un fișier CSV.

The elements of the covariance matrix, c_{ij} can be computed by:

$$c_{ij} = \frac{1}{p}\sum_{k=1}^{p}(I_{ki} - \mu_i)(I_{kj} - \mu_j)$$

- The covariance matrix has size 361x361.
- Dot product between every pair of columns i and j

Activitate practică

Ex 4. Calculați matricea coeficienților de corelație și salvați-o într-un fișier CSV.

The standard deviation of a feature i is:

$$\sigma_i = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(I_{ki} - \mu_i)^2}$$

The correlation coefficient is:

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j}$$

Note that $c_{ii} = \sigma_i^2$ and $\rho_{ii} = 1$.

The standard deviation is computed for each column

The std deviation vector has 361 values

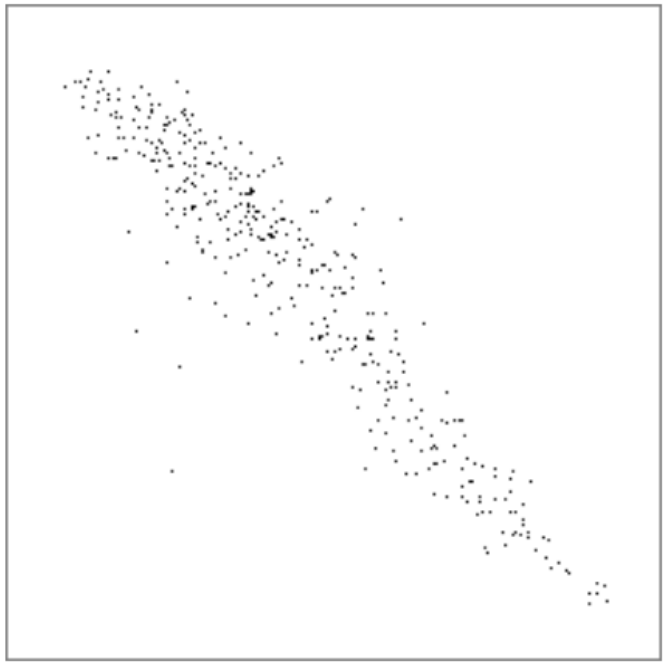
The correlation coefficients matrix has size 361x361

Activitate practică

Ex 5. Afișați coeficientul de corelație si graficul de corelație pentru următoarele poziții (linie, coloana). Graficul de corelație este o imagine albă de dimensiune 256x256 care conține puncte negre la fiecare poziție (I_{kj}, I_{ki}), unde *k=1:p* iar *i* și *j* au fost fixate și reprezintă poziții liniarizate din imagine.

Activitate practică

(5,4) and (5,14). These points correspond to pixels belonging to left eye and right eye. Your result should resemble the one in figure below having the correlation coefficient ~ 0.94.



1. Initialize a white image of size 256x256

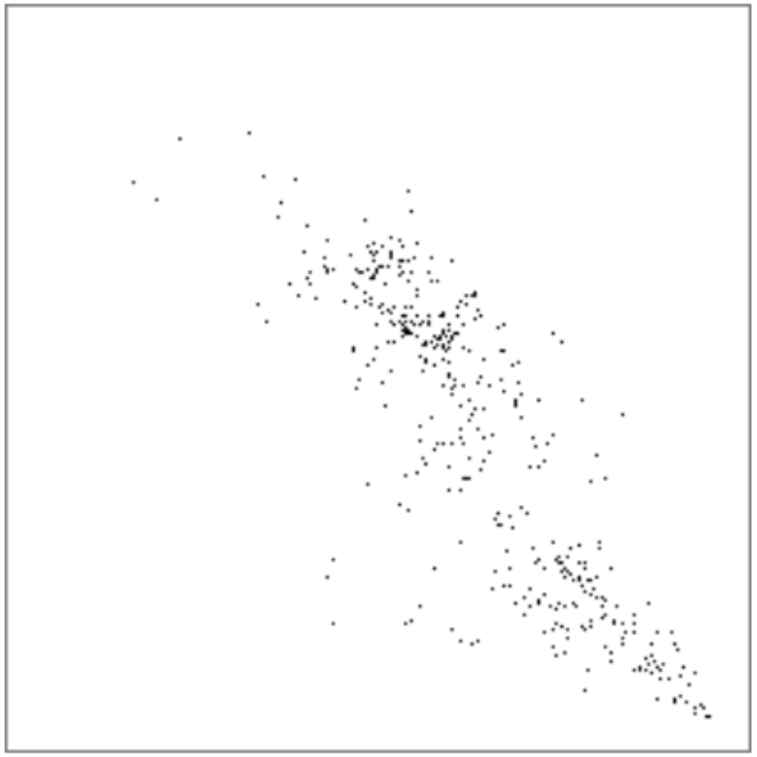
2. Compute i, j

Ex. i = 5 * 19 + 4
j = 5 * 19 + 14

At location (I_{kj},I_{ki})
where k=0; k < 400, set the pixel in the image to 0

Activitate practică

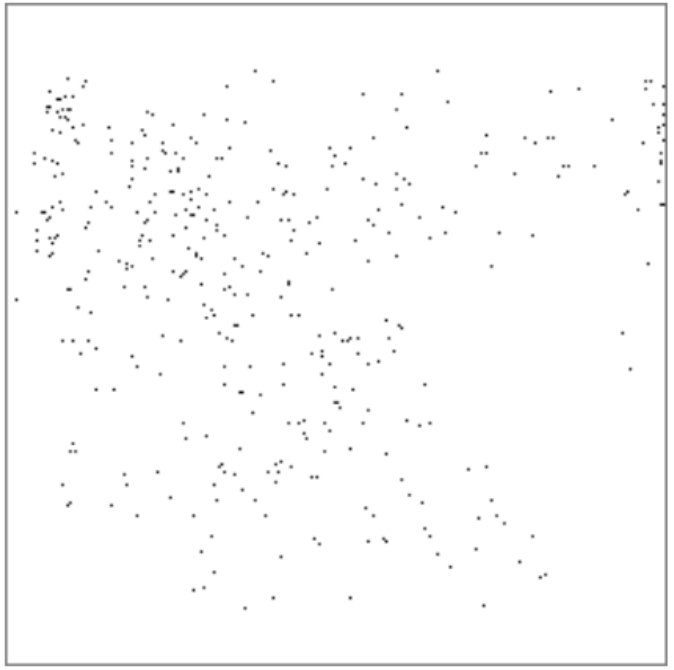
- Ex 5. (10,3) and (9, 15). These points correspond to pixels belonging to left cheek and right cheek. Your result should resemble the one in figure below having the correlation coefficient ~ 0.84.



Activitate practică

- Ex 5.

(5,4) and (18,0). These points correspond to pixels belonging to left eye and the left bottom corner of the face images (notice these points are not highly correlated). Your result should resemble the one in figure below having the correlation coefficient ~ 0.07.



Activitate practică

1. Ex 6. Afișați graficul funcției de densitate de probabilitate unidimensională pentru o trăsătură aleasă. Formula funcției de densitate gaussiană este:

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

- unde μ este valoarea medie și σ este deviația standard pentru trăsătura selectată. Se normalizează valorile astfel încât maximul sa fie egal cu înălțimea imaginii.

- The chart image could be of size 256x256, x =0; x < 256