Pattern recognition systems – Lab 5

Statistical Data Analysis

1. Objectives

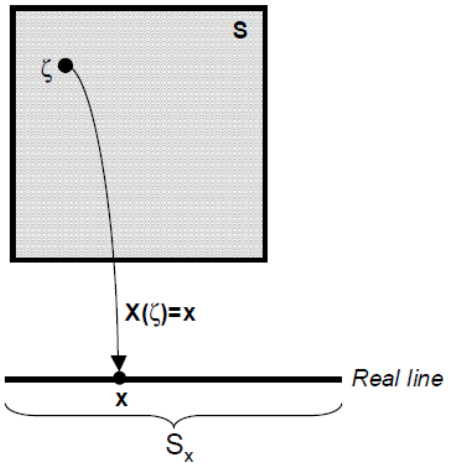
The purpose of this lab is to explore methods of analyzing statistical data used for classification and recognition. We will study the mean, standard deviation, covariance and the Gaussian probability density function. Our experiments will be done on a set of images containing faces. Using the covariance matrix we will study the correlations among different pixels.

2. Theoretical Background

2.1 Definitions

A *random variable* X is a function that assigns a real number $X(\zeta)$ to each outcome ζ in the sample space of a random experiment (see figure below). This function $X(\zeta)$ is performing a mapping from all the possible elements in the sample space onto the real line (real numbers). Random variables can be:

- Discrete: the resulting number after rolling a dice;
- Continuous: the weight of a sampled individual.



A *random variable vector* X is a function that assigns a vector of real numbers to each outcome $X(\zeta)$ in the sample space S. The notion of a random vector is an extension to that of a random variable:

$$\boldsymbol{X} = [X_1, X_2, \dots, X_N]^T$$

2.2 Statistical Characterization of Random variables

A random variable can be partially characterized by:

1. Expectation: represents the center of mass of a density.

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

2. Variance: represents the spread about the mean.

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

3. Standard deviation: The square root of the variance. It has the same units as the random variable.

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

$f_x(x)$ is the probability density function for the random variable *X*.

2.3 Statistical Characterization of Random Vectors

We can (partially) describe a random vector with the following measures:

1. Mean vector:

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

2. Covariance matrix:

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

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

The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together, i.e., to co-vary. The covariance has several important properties:

- If X_i and X_k tend to increase together, then $c_{ik} > 0$
- If X_i tends to decrease when X_k increases, then $c_{ik} < 0$
- If X_i and X_k are **uncorrelated**, then $c_{ik} = 0$
- $|c_{ij}| < \sigma_i \sigma_j$, where σ_i is the standard deviation of X_i
- $c_{ii} = VAR[X_i]$
- $c_{ij} = c_{ji}$

The covariance terms can be expressed as:

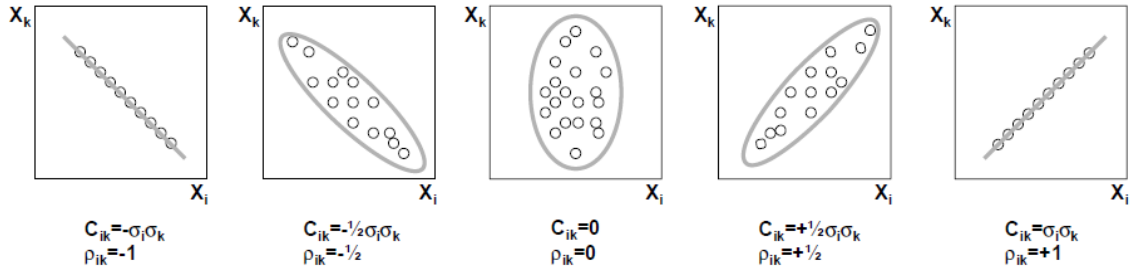
$$c_{ik} = E[(X_i - \mu_i)(X_k - \mu_k)]$$

$$c_{ii} = \sigma_i^2$$

$$c_{ik} = \rho_{ik} \sigma_i \sigma_k$$

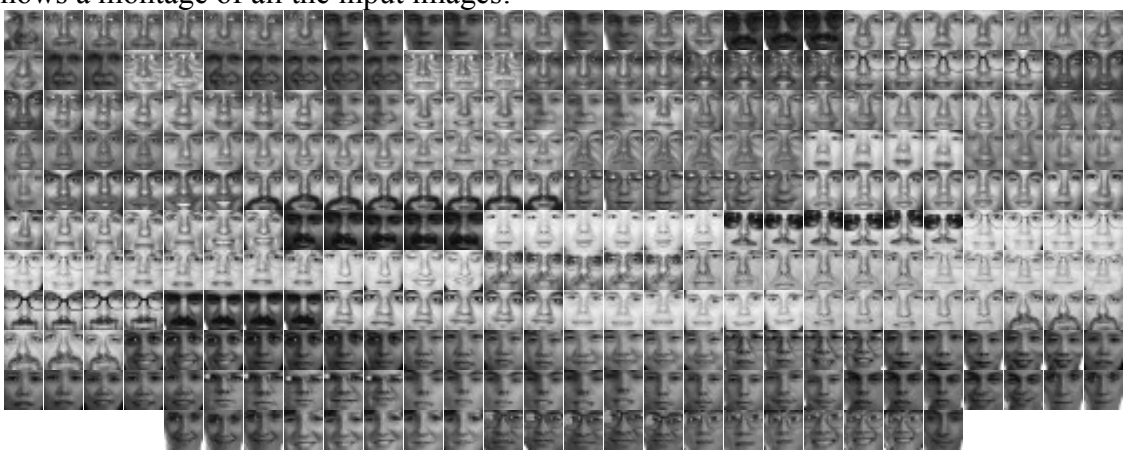
where ρ_{ik} is called the **correlation coefficient.**

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



3. 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:



Let ***I*** be the feature matrix which will hold all the intensity values from the image set. ***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 ***I*** is that each row holds a sample for the N dimensional random variable ***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.

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 standard deviation of a feature *i* is:

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

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 correlation coefficient is:

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

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

4. Practical Work

1. Load the 400 images and store the intensity values as rows in the feature matrix ***I***. The code that loads several images from a folder is:

```
char folder[256] = "faces";
char fname[256];
for(int i=1; i<=400; i++){
   sprintf(fname,"%s/face%05d.bmp", folder, i);
   Mat img = imread(fname,0);
}
```

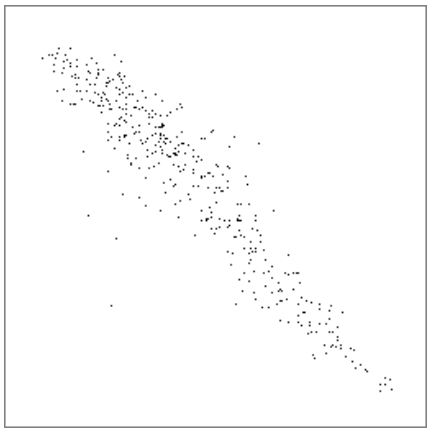
2. Compute mean values for each feature and save them to a csv text file (comma separated values). Write the components in a text file separated by commas and save it with a csv extension. Csv files are viewable in Microsoft Excel as tables.

3. Compute the covariance matrix and save it to a csv text file.

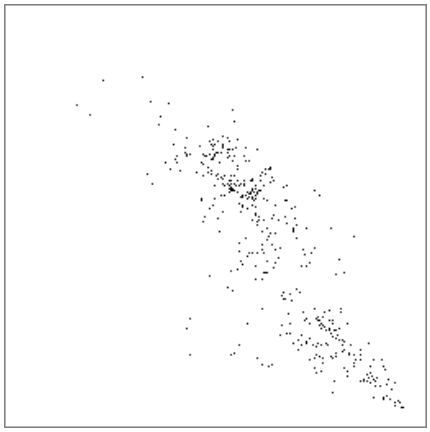
4. Compute the correlation coefficients matrix and save it to a csv text file.

5. Compute the correlation coefficient and display the correlation chart between selected intensity feature pairs. The correlation chart between the ith and jth features is a 256x256 white image with black points at locations (I_{kj}, I_{ki}), for each possible *k*. Use the following coordinate pairs (row, column) which must be linearized (transformed to a single value using the row-major order presented above) to find the correct column index from ***I***:

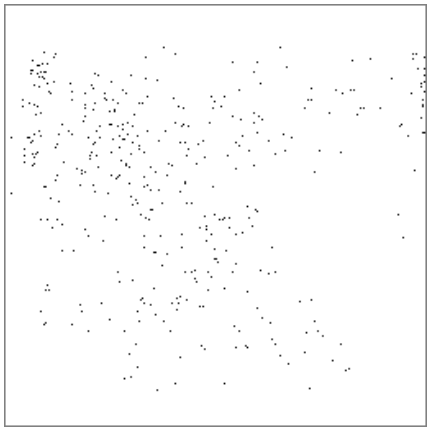
a. (5,4) and (5,14). These points correspond to pixels belonging to left eye and right eye. Your result should resemble the one in figure below having the correlation coefficient ~ 0.94.



b. (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.



c. (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.



6. Plot the probability density function for a selected feature having the form of a one dimensional Gaussian probability density function:

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

where μ is the mean and σ is the standard deviation for the selected feature. Normalize the density values so that the peak reaches the height of the image.

7. Optionally, plot the 2D probability density function as a grayscale image for two selected features using the 2D Gaussian probability density function:

$$p(x_i, x_j) = \frac{1}{2\pi\sqrt{\det(C_{ij})}} exp\left(-0.5\left([x_i - \mu_i, x_j - \mu_j] C_{ij}^{-1} \begin{bmatrix} x_i - \mu_i \\ x_j - \mu_j \end{bmatrix}\right)\right)$$

where μ_i is the mean for feature i and C_{ij} is the covariance matrix between features i and j. Normalize the density values to fit inside the range 0:255.

5. References

MIT CBCL FACE dataset http://www.ai.mit.edu/courses/6.899/lectures/faces.tar.gz