Orthogonality and symmetric matrices *(4.5 points)*

This exercise is a practical example of how to perform a *Principal Component Analysis* (PCA) of a collection of data.

The answers should be given in the answer sheet provided in the course webpage. The Octave history file must be uploaded. Deadline: 16th October, 16pm.

To be done with Octave. Consider the data matrix

$$X = \begin{pmatrix} 2.5 & 3.4 & 5.6 \\ 1.3 & 4.3 & 5.2 \\ 6.1 & 2 & 3.1 \\ 3.2 & 1.5 & 3.4 \\ 6.2 & 4.4 & 5.1 \\ 2.4 & 7.2 & 4.6 \end{pmatrix}.$$

Imagine this is the data of some study, where the rows correspond to different patients, or genes $g_1, \ldots, g_5$, and the columns to different conditions c_1, c_2, c_3. We think of the rows as vectors in $\mathbb{R}^3$, written as a linear combination of c_1, c_2, c_3. In this way, c_1, c_2, c_3 are identified with the canonical basis of $\mathbb{R}^3$.

(i) Center the matrix X by subtracting from each column the mean of that column $\bar{c}_1, \bar{c}_2, \bar{c}_3$. Store the result in a variable called Y.

(ii) Store in a variable C the *covariance matrix* (with respect to the conditions) of the data X:

$$C = \frac{1}{5} Y^t Y.$$

This matrix satisfies the following property: given two linear combinations of c_1, c_2, c_3, $x_1 = a_1c_1 + a_2c_2 + a_3c_3$ and $x_2 = b_1c_1 + b_2c_2 + b_3c_3$, the covariance between x_1 and x_2 is given by the matrix product $(b_1, b_2, b_3) \cdot C \cdot (a_1, a_2, a_3)^t$.

(iii) Find the eigenvalues and an **orthonormal** basis of eigenvectors of the symmetric matrix C. Store the diagonal matrix in the variable D, and the matrix with columns the chosen eigenvectors in the variable V. Order the columns of eigenvectors from higher to lower eigenvalue.

The reference system defined by the chosen eigenvectors are called the Principal Components associated to X. We denote them by pc_1, pc_2, pc_3. Observe that they give an orthonormal basis of $\mathbb{R}^3$, and that the reference origin is in $(\bar{c}_1, \bar{c}_2, \bar{c}_3)$.

(iv) Find the orthogonal projections of each of the samples $g_1, \ldots, g_5$ into each principal component pc_1, pc_2, pc_3.

Questions. With the results found above, answer the following questions:

(i) With the theory given in class, justify the following claim: "The covariance between the different principal components is zero; The first principal component is the one with higher variance".

(ii) Give the coefficients of each of the components pc_i as a linear combination of c_1, c_2, c_3. These coefficients are called the *loadings*.

(iii) Give the coefficients of each g_i in the basis pc_1, pc_2, pc_3:

$$g_i = s_{i1}pc_1 + s_{i2}pc_2 + s_{i3}pc_3.$$

These coefficients are called *scores*.