Tutorial 8: Principal Component Analysis (PCA)

As we have seen in the lecture, PCA is a multivariate statistical technique that is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.

1. Describe a step-by-step procedure that calculates a PCA transformation matrix P of the *n*-dimensional sample *X*. This transformation P should retain as many principal components *k* as necessary in order to explain a certain amount *v* (for instance, 90%) of the total sample variance. Your first step could be "calculate the covariance matrix *S* of *X*".

2. Consider the following covariance matrix

$$S = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}.$$

We will analyse its corresponding correlation matrix and check that the principal components obtained from covariance and correlation matrices are different.

2a. Calculate the derived correlation matrix *R*. The correlation matrix entries are of the form

$$r_{i,j} = \sigma_{i,j} / (\sqrt{\sigma_{i,i}}\ \sqrt{\sigma_{j,j}})$$

2b. Calculate the eigenvalues λ_1 and λ_2 of *S* using the formula $\det(S - \lambda I) = 0$, where *I* is the 2x2 identity matrix.

2c. Calculate the eigenvectors ϕ_1 and ϕ_2 associated with these eigenvalues by solving the following equations:

$$S\phi_1 = \lambda_1\phi_1$$
$$S\phi_2 = \lambda_2\phi_2$$

where $\phi^T = [x_1, x_2]$.

2d. Compute the proportion of the total sample variance explained by the first principal component ϕ_1 of *S*. Is there any variable (x_1, x_2) that dominates ϕ_1? Explain.

2e. Analogously calculate the eigenvalue-eigenvector pairs of *R* (that is, repeat steps 2b and 2c).

2f. Compute the proportion of the total sample variance explained by the first principal component ξ_1 of *R*. Is there any variable (x_1, x_2) that dominates ξ_1? Explain.

2g. What have you learned?