Principal Component Analysis & Whitening

This exercise requires to compute the Principal Components (PCs) for 3 different datasets. All three datasets are contained in the zip-file `PCAdata.zip` which can be downloaded from ISIS. Feel free to use available software packages for previously solved tasks (e.g. pca).

4.1 Preprocessing (2 points)

- Load the dataset `pca2.csv`. Compute the Principal Components PC1 and PC2 and plot the data in the coordinate system PC1 vs. PC2 – What do you observe?
- Remove Observations 17 and 157 and redo the first two steps. What is the difference?

4.2 Whitening (4 points)

- Load the dataset `pca4.csv` and check for outliers in the individual variables.
- Do PCA on a reasonable subset of this data. Use a scree plot to determine how many PCs represent the data well.
- "Whiten" the data, i.e. create a set of 4 *uncorrelated* variables with *mean 0* and *standard deviation equal to 1*. This can be done e.g. using the transformation

$$Z = \tilde{X} E D^{-1/2}$$

  The new variables z_i form the columns of Z, E is a matrix containing the Eigenvectors of the Covariance matrix Σ of the centered data $\tilde{X}$ and D is a diagonal matrix containing the corresponding eigenvalues.
- Make heat-plots (e.g. using `imshow` in Python or `imagesc` in Matlab) of the 4x4 covariance matrix Σ, the covariance matrix of the Principal Component values PC1-PC4, and of the whitened variables.

4.3 Rotation (4 points)

- Load the dataset `pca2b.csv` and plot estimates of the marginal densities of the two variables e.g. using a histogram or kernel estimator.
- Do a PCA and plot the data in the PC coordinate system.
- Whiten the data and rotate the whitened variables by 45 (or other angles you find informative) degrees using a rotation matrix, i.e.

$$Z^{\mathrm{rot}} = (RZ')' \qquad \text{with} \qquad R = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

- Plot the marginal densities of the whitened variables z_i and of the rotated variables z_i^{rot}. Compare the covariance matrices for the original variables, the whitened and the rotated variables.

Total points: 10