1 PCA qualitative

Statistics - A) Joe's first job as a statistician was to analyse the countries that won medals in the 2012 London Olympic Games. He started his task by collecting data from 84 countries that had won at least one medal. The data's features (dimensions) for each country (sample/observation) consisted of the number of medals won, the population size (in hundred thousands) and the GDP (in millions of US dollars) all taken in 2012. To start his analysis Joe ran PCA on the data.

1. The first eigenvector explained 99.9% of the data's variance. Joe was very happy! He was already planning on winning rivers of money by betting on the medals distribution of the 2016 Olympic Games. Assuming that Joe is a competent analyst why did he think he could predict the medals distribution of the next olympic games?

2. Joe's boss was not as happy. He ran the PCA by himself and he found that the first eigenvector explained only 74.0% of the data's variance. They compared what they both had done but the only difference they found was that Joe's boss used the GDP values in billions of dollars while Joe's GDP values were in millions of dollars. Why did they obtain different results? How should Joe and his Boss have treated the data before the PCA? What was probably the first eigenvector found by Joe?

3. Joe has finally learned the correct way to handle the data before running PCA. He was now trying to run the analysis with extra features (dimensions). He added the countries call prefix code (e.g. 41 for Switzerland) to the last column of the data. Table 1 shows the results that he obtained. How would you interpret these results?

	Medals	**Population**	**GDP**	**Codes**	**Variance Explained**
1^{st} **Eigenvector**	-0.619	-0.478	-0.621	-0.051	56%
2^{nd} **Eigenvector**	0.132	-0.263	0.149	-0.944	26%

Table 1: Joe's results.

Dimensionality Reduction - B)

1. Jenny has a dataset of 10 000 images of faces. These are all black and white 100 x 200 px images where each pixel can have a value from 0 to 255 (1 byte) to represent the intensity. Because Jenny cares about reducing the dimension of each image but not to find common features across the images, she stores the image in a matrix X where each column is a different image and each row corresponds to each different pixel location[1].

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,10'000} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,10'000} \\ \vdots & \vdots & \vdots & \vdots \\ x_{20'000,1} & x_{20'000,2} & \cdots & x_{20'000,10'000} \end{bmatrix}$$

[1] The images are now represented as column vectors with 20′000 (width × height) dimensions.

The dataset occupies 200MBs (1 (byte/pixel)$\times 20'000$ (pixels/image) $\times 10'000$ (images) $= 200$ (MBs)) in memory. Jenny needs to compress them because she wants to make the dataset available on her internet server, and, if multiple users download it, she ends up paying a big internet bill. To find the optimal compression, she decides to run PCA on those images. She finds that the first 100 principal components explain 90% of the data's variance. She also observes that if she projects the images onto the first 100 eigenvectors and discards the other projections, her images lose resolution but remain readable. This project hence might pay off thanks to the generated decrease in memory space. Has Jenny really gained in memory? And if so, how much has she gained? How much memory does the dataset occupy now?

Hint: You can consider each float number takes 8 bytes (64bits) of space

Eigenvectors intuition for image data - C) Let's assume that we have the dataset of grayscale images of pens with either vertical or horizontal orientation (refer to Figure 1). We want to be able to classify whether a pen is rotated horizontally or vertically, using a PCA projection.

1. First step when doing PCA on any dataset is to center the data, via subtracting the mean. a) Try to sketch (i.e. draw approximately) the "mean" of the dataset presented in Figure 1. Assume that dataset is balanced, i.e. contains equal number of vertical and horizontal pens. b) How does "centered" image (i.e. after mean image subtraction) of a vertically (and horizontally) oriented pen looks?

   Hint: Assume that image is grayscale uint8, 0 is for black pixels, 255 is for white, all negative values are visualized as black. "Subtracting images" should be treated as pointwise pixel intensity v subtraction, i.e. if pixel is white ($v = 255$), then subtracting black ($v = 0$) pixel results in white pixel ($255 - 0 = 255$).

2. After centering the data, and computing the covariance matrix, one should perform eigenvalue decomposition, and use resulting eigenvectors as a projection basis for the data of reduced dimensions. Since the eigenvectors have the same dimensionality as the original images, they can be visualized, highlighting filtering features. How does such an eigenvector looks like? Try sketching two first eigenvectors of the projection basis.

3. Using these two eigenvectors as a new basis, plot two-dimensional projections of the images of vertically and horizontally oriented pens. Can they be separated easily? How many eigenvectors is required for these two classes to be linearly separable? Can you think of a single eigenvector, that can be used to separate two classes (horizontally and vertically oriented pens)?

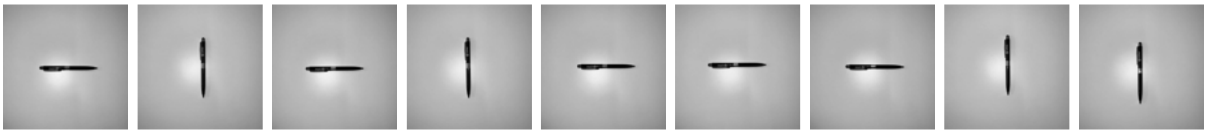
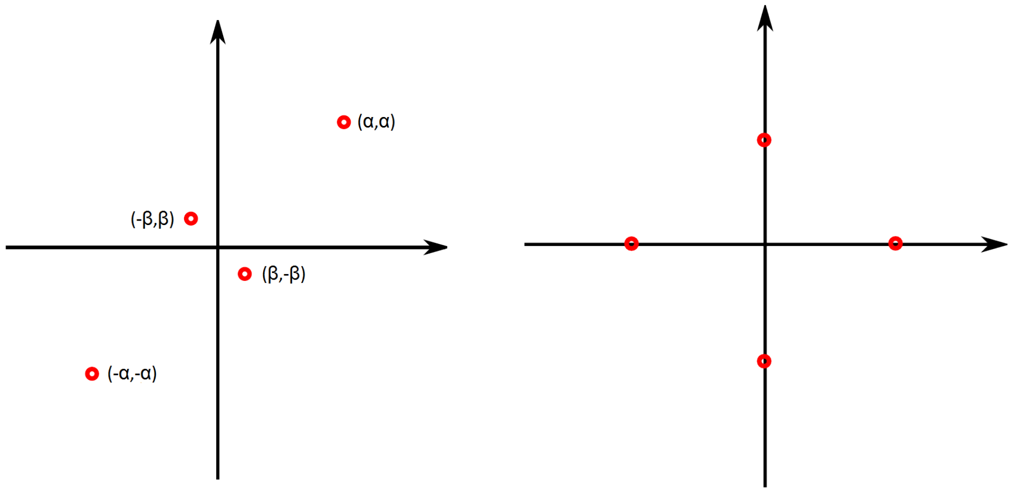


Figure 1: Pen dataset images $i = 1..9$

2 PCA quantitative

A) Let us go through through the important steps of PCA:

1. Compute the covariance matrix, C, of the dataset composed of the four points shown in the left figure ($(\alpha\ \alpha)$, $(-\alpha\ -\alpha)$, $(-\beta\ \beta)$, $(\beta\ -\beta)$).



(a) (b)

2. Compute the eigenvectors and eigenvalues of C. Discuss how the eigenvectors and eigenvalues are affected by α and β.

3. Compute the projection of the dataset on the principal directions. Then reduce the dimensionality of the dataset by discarding the less relevant dimension in the projected space (assuming $\alpha > \beta$).

4. From the reduced dataset, reconstruct an approximation of the dataset in the original referential. Compute the reconstruction error, and discuss its relation to the eigenvalues of C.

5. For the data from the right figure, what structure (isotropic, diagonal, full, symmetric) does the covariance matrix have?

B) Show that when an N-dim set of data points X is projected onto the eigenvectors $V = [\mathbf{e}^1, \cdots, \mathbf{e}^N]$ of its covariance matrix, $C = XX^{\mathrm{T}}$, the covariance matrix YY^{T} of the projected data Y is diagonal and hence that, in the space of the eigenvector decomposition, the distribution of X is uncorrelated.