Rules:

- Unless otherwise stated, all answers must be mathematically justified.
- Partial answers will be graded.
- Your submission has to be uploaded to Gradescope. In Gradescope, indicate the page on which each problem is written.
- You can work in groups but each student must write his/her/their own solution based on his/her/their own understanding of the problem. Please list on your submission the students you work with for the homework (this will not affect your grade).
- Problems with a ($\star$) are extra credit, they will not (directly) contribute to your score of this homework. However, for every 4 extra credit questions successfully answered your lowest homework score get replaced by a perfect score.
- If you have any questions, feel free to contact me (`mgabrie@nyu.edu`) or to stop at the office hours.

Problem 7.1 (3 points)**.** *We say that a symmetric matrix $M \in \mathbb{R}^{n\times n}$ is **positive semi-definite** if for all **non-zero** $x \in \mathbb{R}^n$, $x^\mathsf{T} M x \geq 0$. Furthermore, a symmetric matrix $M \in \mathbb{R}^{n\times n}$ is **positive definite** if for all **non-zero** $x \in \mathbb{R}^n$, $x^\mathsf{T} M x > 0$.*

(a) *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Show that M is positive semi-definite if and only if its eigenvalues are all non-negative.*

(b) *Consider J_n the $n \times n$ matrix of all ones (all entries equal to 1). Show that J_n is positive semi-definite using* **(a)**.

(c) Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Show that there exists $\alpha > 0$ such that the matrix $M + \alpha \mathrm{Id}_n$ is positive definite.

Problem 7.2 (3 points)**.** *Using PCA, we reduce the dimension of a dataset $a_1, \dots, a_n \in \mathbb{R}^d$ of mean zero, to get a «dimensionally reduced dataset» $b_1, \dots, b_n \in \mathbb{R}^k$, for some $1 \leq k \leq d$. We note A the $n \times d$ matrix*

$$A = \begin{pmatrix} - & a_1^\top & - \\ & \vdots & \\ - & a_n^\top & - \end{pmatrix}.$$

(a) *Show that the dataset $b_1, \dots, b_n$ is centered: $\sum_{i=1}^n b_i = 0$.*

(b) *Show that for all $i, j \in \{1, \dots, n\}$, we have*

$$\|b_i - b_j\| \leq \|a_i - a_j\|.$$

This means that PCA shrinks the distances.

(c) *For $i \in \{1, \dots, k\}$ we let*

$$f^{(i)} = (b_{1,i}, b_{2,i}, \dots, b_{n,i}) \in \mathbb{R}^n$$

be the vector made of all i^{th} components of the vectors $b_1, \dots, b_n$. Show that for $i \neq j$, $f^{(i)} \perp f^{(j)}$. This means that the new features computed using PCA are uncorrelated.

Problem 7.3 (3 points)**.** *You have been given a mysterious dataset that may contain important informations! This dataset is a collection of $n = 6344$ points of dimension $d = 1000$. Investigate the structure of this dataset using PCA/plots... , and find out if the dataset contains any information.*

The `zip` *file* `mysterious_data.zip` *contains a text file containing the* 6344×1000 *data matrix. The* Jupyter notebook `mysterious_data.ipynb` *contains a function to read the text file.*

You are not allowed to use any builtin PCA function: you have to do the all process by yourself (centering the data, computing the covariance matrix...). Of course, for computing eigenvalues/eigenvectors you will need to use the numpy library. The numpy function `numpy.linalg.eigh` *is great to compute eigenvalues and eigenvectors of a symmetric matrix.*

It is intended that you code in Python and use the provided Jupyter Notebook. Please only submit a pdf version of your notebook (right-click → 'print' → 'Save as pdf').

Problem 7.4 ($\star$)**.** *Let* $A \in \mathbb{R}^{n \times n}$ *be a symmetric positive semi-definite matrix. Prove that there exists* $B \in \mathbb{R}^{n \times n}$ *positive semi-definite such that* $A = B^2$.