Optimization and Computational Linear Algebra for Data Science
Homework 7: Principal component analysis

Due on November 5, 2019

- Unless otherwise stated, all answers must be mathematically justified.
- Partial answers will be graded.
- You can work in groups but each student must write his/her own solution based on his/her 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 (`lm4271@nyu.edu`) or to stop at the office hours.

Problem 7.1 (3 points). *Let $A \in \mathbb{R}^{n\times m}$. The Singular Values Decomposition (SVD) tells us that there exists two orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ and a matrix $\Sigma \in \mathbb{R}^{n\times m}$ such that $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq 0$ and $\Sigma_{i,j} = 0$ for $i \neq j$*

$$A = U\Sigma V^{\mathsf{T}}.$$

The columns $u_1, \dots, u_n$ of U (respectively the columns $v_1, \dots, v_m$ of V) are called the left (resp. right) singular vectors of A. The non-negative numbers $\sigma_i \stackrel{\text{def}}{=} \Sigma_{i,i}$ are the singular values of A. Moreover we also know that $r \stackrel{\text{def}}{=} \mathrm{rank}(A) = \#\{i \,|\, \Sigma_{i,i} \neq 0\}$.

(a) *Let $\widetilde{U} = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{pmatrix} \in \mathbb{R}^{n\times r}$, $\widetilde{V} = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{pmatrix} \in \mathbb{R}^{m\times r}$ and $\widetilde{\Sigma} = \mathrm{Diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r\times r}$. Show that $A = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathsf{T}}$.*

(b) *Give orthonormal bases of $\mathrm{Ker}(A)$ and $\mathrm{Im}(A)$ in terms of the singular vectors $u_1, \dots, u_n, v_1, \dots, v_m$.*

Problem 7.2 (3 points). *We say that a symmetric matrix $M \in \mathbb{R}^{n\times n}$ is positive definite if for all* ***non-zero*** *$x \in \mathbb{R}^n$,*

$$x^{\mathsf{T}} M x > 0.$$

If a matrix M is positive definite, then M is also positive semi-definite, but the converse is not true. One of the goals of this problem is to prove a part of Proposition 1.2 in the notes (Lecture 7). You are of course not allowed to use this proposition to solve this problem.

(a) *Let $M \in \mathbb{R}^{n\times n}$ be a positive definite matrix. Show that its eigenvalues are all strictly positive and that M is invertible.*

(b) *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 7.3 (4 points). *You have been given a mysterious dataset that may contain important informations! This dataset is a collection of $n = 3000$ 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* 3000×1000 *data matrix. The* Jupyter notebook `mysterious_data.ipynb` *contains a function to read the text file. 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 $M \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of M. Show that for all $d \leq n$:*

$$\max_{\substack{U \in \mathbb{R}^{n \times d} \\ U^\mathsf{T} U = \mathrm{Id}_d}} \mathrm{Tr}(U^\mathsf{T} M U) = \sum_{i=1}^{d} \lambda_i.$$