HOMEWORK 1
COMPUTATIONAL METHODS FOR DATA SCIENCE
FALL SEMESTER 2023

Climate.gov records climate statistics at individual stations in USA. In this problem, we analyze a subset of the original full data that focuses on the monthly record-breaking high temperature of California. The file CAmaxTemp.txt can be downloaded from the following link:
http://staff.stat.sinica.edu.tw/fredphoa/HW/HW1/CAmaxTemp.txt
In this dataset, the first column is the station name, the second column is the investigation period, and the last column is the yearly high-temperature record. The remaining 12 columns forms a full data (a 12×12 matrix) of the monthly high-temperature records during the investigation period.
NOTE: Please include your own iterative code how you obtain your results. DO NOT copy and paste any library or function from existing programs.

0. **A Practice on the Randomization** *(5 points)* This is a data preparation step for the rest of this homework. Instead of considering the full data (a 12×12 matrix), we will randomly pick six stations (rows) among all 12 stations and six months (columns) among all 12 months for investigation. This means you only need to work on a reduced data (a 6×6 matrix) and we denote this data as X.

   (a) Write a simple code (any programming language of your choice) to randomly pick 6 objects out of 12, where all 12 possible objects have the same probability to be picked (i.e. uniform probability).

   (b) Run the code on both the stations and the months, and state which stations and months you obtain as the reduced data X.

   (c) Do a quick check if X has at least four REAL eigenvalues. You can do this question by using existing library or packages implemented in the software of your choice.

1. **Monthly Record-Breaking Temperature in California I: Matrix Calculation** *(25 points)* In this problem, we try to do some matrix decompositions and solve the eigenvalue problems on your 6×6 dataset.

   (a) Run the LU factorization on X to obtain a lower triangular matrix L and an upper triangular matrix U.

   (b) Use Power Iteration method to find the largest eigenvalue-eigenvector pair of X.

   (c) Use QR factorization to find all eigenvectors with REAL eigenvalues of X.

   (d) Find the inverse of X.

2. **Monthly Record-Breaking Temperature in California II: PCA and SVD** *(35 points)* We continue to use the same 6×6 dataset in this problem.

   (a) Center the data and compute the variance-covariance matrix.

   (b) Find the top three principal components using power iteration. Calculate the cumulative percentage of the total eigenvalues that these three principal components cover.

   (c) Plot the data on a 3D space with three principal component axes. Provide the coordinates of the recast data.

(d) Find all principal components with their eigenvalues using SVD.

(e) SVD provides an extra information on U that PCA does not usually have. Is it any interpretation of this U matrix? If yes, please state it.

(f) Conduct a rank-3 approximation (SVD version) of X.

3. **Monthly Record-Breaking Temperature in California III: ICA** *(35 points)* We consider the full 12×12 dataset, NOT your subdata in the previous problems. in this problem, but we only consider the data of February, June and October, i.e. a 12×3 subdata X'. NOTE: Please include your own iterative code how you obtain your results. DO NOT copy and paste any library or function from existing programs.

   (a) Explain why the monthly data is unlikely to be Gaussian.

   (b) Run the three preprocessing steps of ICA on X'.

   (c) Provide a graphical illustration on the transformation, like the one in Lecture 03-2.

   (d) Run the Fast ICA on Kurtosis Maximization to find the three independent components.

B. **Determinant and Parallogram.** *(Bonus 10 points)* In Lecture 02-1, we know that a determinant is equal to the area of the parallelogram. Please prove it.