Essential Mathematics for Data Scientists

Assignment 4: Principal Component Analysis

This assignment asks you to implement, a principal component analysis (PCA) using a singular value decomposition on a real data set.

A real dataset

We now ask that you investigate the provided dataset which has come from the State of the Tropics report. (State of the Tropics (2017) Sustainable Infrastructure for the Tropics. James Cook University, Townsville, Australia, available at stateofthetropics.org). The data that we have used from this report is available at the Tropical Data Hub (https://research.jcu.edu.au/researchdata/default/search?query=State+of+the+Tropics&sort-field=score&sort-order=desc)

We have consolidated much of the data for the year 2010 into a file SotTCombined2010.xlsx. We ask that you analyse this data using the PCA and obtain the task requirements listed below. You will need to use MATLAB to do the computation. Write your analysis in a report, using Word, Latex, or any program you would like to use. You must submit your report as a PDF file.

The task

You are to implement the PCA using SVD, obtaining:

1. The principle component vectors
2. The proportion of variation explained by the principal components
3. The matrix of scores
4. A dimensionally reduced representation of the dataset
5. The residuals of the reduced representation
6. Any outliers

Your report must determine the relationships between the variables reported in the spreadsheet and the strength of those relationships. You should also identify the outlier countries, those for which the relationships the PCA identifies are not present.

In analysing the data there are several things to keep in mind:

- The data contains many missing values. You should exclude countries which contain missing values.