LOW DIMENSIONAL EMBEDDING OF ENVIRONMENTAL VARIABLES

EA MAP581

6 mars 2018

Flore Martin and Lorraine Roulier

ÉCOLE
POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY
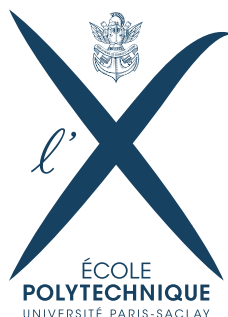
Table des matières

1   Introduction

Climate data amounts very quickly to a lot of unused data. In a day, we can collect temperature, pressure, wind data all over the world with satellites, even hourly. Our project was two sided. First, we familiarized with various dimension reduction techniques, then we attempted to show that the geographical position of a point on the planet - e.g. it's latitude and longitude - were embedded in the climate data one could gather on it.

Dimension reduction techniques can be divided in two classes, linear dimension reduction and non linear dimension reduction. However, in all methods, the main goal is to figure out a similarity function between vectors. Such a function will then enable to sort the dataset into classes of vectors with similar features, which would have been more intricate with the initial dataset. We used a set of datasets we found on the NASA website, that gathered various means on climate variables over 22 years at every given latitude and longitude. These variables are gathered in the table below

Latitude	Longitude	Temperature °C	Pressure kPa	Relative Humidity %	Wind Speed m/s	Radiation $kWh/m^2/day$

Figure 1 – First lign of our dataset

The latitude parameter varies from -90 to 89 and the longitude parameter varies from -180 to 179. the negative values are for the south hemisphere, the positive ones for the north. Depending on the running time of the method, we did not compute the dimension reduction with the 64800 lines, but with a subset. The subset is often a slice of longitudes containing all latitudes, as we assumed that the critical parameter to differenciate climate data was the latitude.
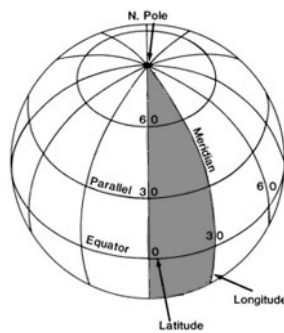

Figure 2 – An example of subset in grey

2 Principal Component Analysis - PCA

2.1 Method

Principal Component Analysis detects tendencies in the data by maximizing the variance of the dataset matrix. This yields an orthonormal matrix that can be diagonalized. The largest eigenvalues point to the eigenvectors that contain the most information about the dataset.

Let $X \in \mathbb{R}^{d \times n}$ be our dataset, PCA maximizes the following equation :

$$\|X - MM^T X\|^2$$

subject to $M \in \mathcal{O}^{d \times r}$ where $r < d$.

2.2 Results

2.2.1 Two dimensions

We first implemented PCA and ran it with only two principal components

3 Kernel Principal Component Analysis

3.1 Method

3.2 Results

4 Multidimensional Scaling - MDS

4.1 Method

4.2 Results

5 Isomap

5.1 Method

5.2 Results

6 Comparing the different methods

7 Conclusion

8 Bibliography