



LOW DIMENSIONAL EMBEDDING OF ENVIRONMENTAL VARIABLES

EA MAP581

17 mars 2018

Flore Martin and Lorraine Roulier



Table des matières

1	Introduction	3
2	Dataset presentation	3
3	Methods	4
3.1	Principal Component Analysis - PCA	5
3.2	Kernel PCA	6
3.3	Multidimensional scaling	6
3.4	Isomap	7
4	Results	8
4.1	Graphic Interpretation	8
4.1.1	PCA	8
4.1.2	Three dimensions	9
4.1.3	Kernel PCA	9
4.1.4	Multidimensional Scaling - MDS	12
4.1.5	Isomap	12
4.2	Time comparison	14
4.3	Error and preserved information	17
4.3.1	PCA	17
4.3.2	MDS	18
4.3.3	isomap	19
5	Conclusion	21
6	Bibliography	21

1 Introduction

Dimension reduction refers to the process of reducing the number of random variables under consideration. It is an efficient way to analyze a large dataset of high dimension. For instance, dimensionality reduction can be used for **visualizing** or exploring structure in data. It provides a **simple geometric interpretation** by mapping in 2D or 3D an original dataset of dimension $d > 3$. A 2D or 3D mapping is thus much easy to interpret. In terms of performance, having data of high dimensionality is problematic because it can mean **high computational cost** to compute. Finally, dimensional reduction is often used for classification. Data samples are being clustered, it is a way to establish clusters of gene and of population for example. It is a current tool in machine learning.

In practice, dimension reduction is widely used for denoising or compressing data, and embedded image processing like facial recognition.

In this project, we explored various dimension reduction techniques. The main way to classify the different methods is to observe if they are linear or non-linear. The basic linear dimension method is a linear one, Principal Component Analysis (PCA). In order to familiarize ourselves with dimension reduction, we implemented this method manually. We explored PCA and its non-linear version, Kernel Principal Component Analysis. We also used Multidimensional scaling (MDS) which is a linear method and Isomap, a non-linear method.

2 Dataset presentation

We applied the methods listed above on the dataset we describe here. It is a dataset that contains climate data associated with its position. Our goal is to show that the position on the globe is embedded in the climate data.

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. 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. Those datasets can be found here [6]. As each variable (temperature, radiation, pressure, wind speed and humidity) was stored in a separate dataset, we first had to merge all of them in a final table presented below. We did not select every dataset given by the NASA because we considered that temperature related data for example were already very correlated and did not add much information.

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

FIGURE 1 – Variables of our dataset

The latitude parameter varies from -90 to 89 and the longitude parameter varies from -180 (south pole) to 179 (north pole). 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 differentiate climate data was the latitude. We will detail the subset considered in every method.

Since our goal was to see if the position of a location on earth was embedded in its climate data, we did not use the first two columns of our dataset when computing the dimension reduction methods.

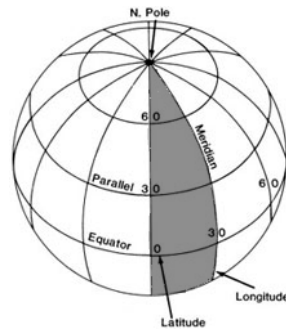


FIGURE 2 – An example of subset in grey

We classified the data according to the latitude, creating five classes listed in the table below. A more visual picture of our classification is given when we introduce the results of the different dimension reduction methods. The latitudes we chose are universally known parallels. The north temperate zone extends from the tropic of cancer to the arctic circle and the south temperate zone from the tropic of capricorne to the antarctic circle.

	North	Temperate North	Equator	Temperate South	South
Latitudes	90 to 66	66 to 23	23 to -23	-23 to -66	-66 to -90
	●	●	●	●	●
% of Earth surface	4	26	40	26	4

FIGURE 3 – Our classification

3 Methods

In all dimension reduction techniques, 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.

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. its latitude and longitude - were embedded in the climate data one could gather on it.

3.1 Principal Component Analysis - PCA

Principal Component Analysis detects tendencies in the data by maximizing the variance of the dataset matrix. There is an easy approach to understand PCA with a 2D dataset. On Figure 4, we can see the plot of a 2D-dataset. If we implement 2D-PCA, we will find two new axes that reflect at best the covariance of the samples. Those are the green axes. PCA just rotates the figure to find a more explicit "point of view". The yielded axes are a linear combination of the original axes.

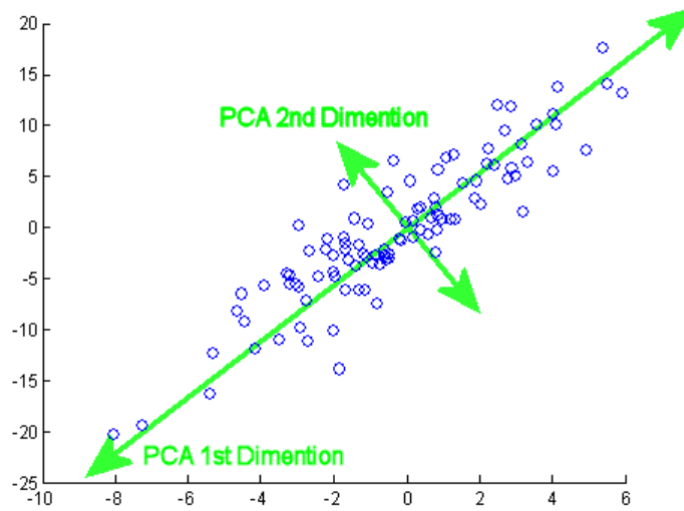


FIGURE 4 – role of PCA - interpretation

From a mathematical point of view, PCA yields an orthonormal matrix that can be diagonalized. The largest eigenvalues point to the eigenvectors that contain the most information about the dataset. The dataset can then be projected on a lower dimension using the right eigenvectors.

Let $X \in \mathbb{R}^{d \times n}$ be our dataset where d is the number of variables and n the sample size, PCA maximizes the following equation :

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

subject to $M \in \mathcal{O}^{d \times r}$ where $r < d$ and $\mathcal{O}^{d \times r}$ is the space of orthonormal matrix and we use the Frobenius norm.

The maximum of the previous function is reached by the covariance matrix of the dataset : the r greatest eigenvalues of $XX^T = PDP^T$ yield the maximum $M = P_r$ where the columns are the eigenvectors associated to the eigenvalues mentioned above. The largest eigenvalues point to the eigenvectors that contain the most information about the dataset.

The new dataset is thus $Y = P_r^T X$ of dimensions $r \times n$. In this example, there are r principal components for the PCA. Each of the M_{ij} coefficients of the matrix can be interpreted as weights given to the variable they are associated to. This enables us to interpret which variable of the dataset gives more significant information about the data.

3.2 Kernel PCA

Kernel Principal Component analysis is a non linear method for dimension reduction. Instead of directly maximizing the variance of X , we map it into a larger dimension space called the feature space, using a non linear function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^N$ where $N > d$.

Instead of directly mapping X into the feature space, we use a kernel, that represents a similarity function between pairs. Let $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be this kernel, then

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

where $(x_i, x_j) \in \mathbb{R}^d \times \mathbb{R}^d$ are vectors of the dataset X

This kernel will enable us to bypass the higher dimension calculation of the variance in the feature space. We only need to compute the pairwise values for ϕ , but there is no need to explicitly define ϕ .

The downside of this method is that since we don't compute directly the eigenvectors and eigenvalues of ϕ , the result is the projection of our dataset onto the eigenvectors. We thus don't have access to a simple interpretation of the principal components.

A classic kernel that is often used is the RBF function that is defined below :

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|}$$

3.3 Multidimensional scaling

Multidimensional scaling is a linear approach to reduce dimension of a dataset. The principle is quite different from PCA : given a matrix of distance or "dissimilarity", MDS aims at reconstructing a map preserving at best the distances between datas. In our example, the MDS algorithm aims to place each object in a 2-dimensional space such that the between-object distances are preserved as well as possible. Given a distance matrix D of dimension $n \times p$ ($p=5$ for us) , MDS attempts to find the datapoints y_1, \dots, y_t in dimension $d < p$ ($d=2$ for us) that minimizes :

$$\sum_{i=1}^t [\sum_{j=1}^t d_{ij}^{(X)} - d_{ij}^{(Y)}]$$

with $d_{ij}^{(X)}$ the euclidean distance between pairwise i and j in the original $n \times p$ matrix $D^{(X)}$ and $d_{ij}^{(Y)}$ the euclidean distance between pairwise i and j in the original $n \times p$ matrix $D^{(X)}$ and in the computed $n \times d$ matrix $D^{(Y)}$.

A basic approach to understand MDS is to use the example of cities. The input is the distance matrix between cities on the left on figure 5. MDS enables us to find the coordinates of each city based on this distance matrix.

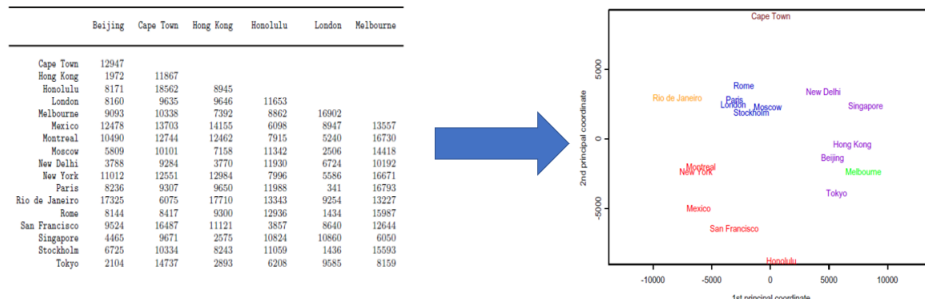


FIGURE 5 – role of MDS - interpretation

3.4 Isomap

Isomap is a low dimensional embedding method similar to MDS. The difference is that distance is not computed with euclidean norm but with geodesic distance. As the earth shape is round, isomap seems more relevant than MDS. On figure 6, we can see the difference between the geodesic distance (in red) used in isomap, and the euclidean distance used in MDS (in blue).

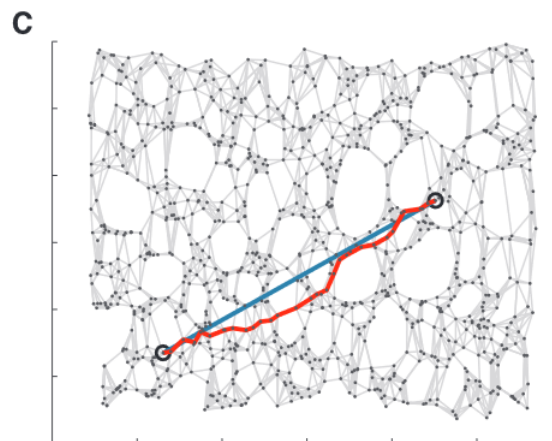


FIGURE 6 – geodesic and euclidean distance

4 Results

4.1 Graphic Interpretation

4.1.1 PCA

We first implemented PCA and ran it with only two principal components, which yielded the graph plotted in figure 7 for the whole dataset. We wrote the eigenvectors y_1 and y_2 that were used horizontally to make it easier to understand their meaning.

Temperature	Pressure	Relative Humidity	Wind Speed	Radiation
-------------	----------	-------------------	------------	-----------

$$y_1 = [-0.94 \quad -0.31 \quad 0.11 \quad -0.04 \quad 0.01]$$

$$y_2 = [0.03 \quad -0.40 \quad -0.91 \quad 0.04 \quad -0.05]$$

This enables us to understand the meaning of these vectors. y_1 is mostly related to a decreasing temperature and pressure, and y_2 represents decreasing humidity and pressure.

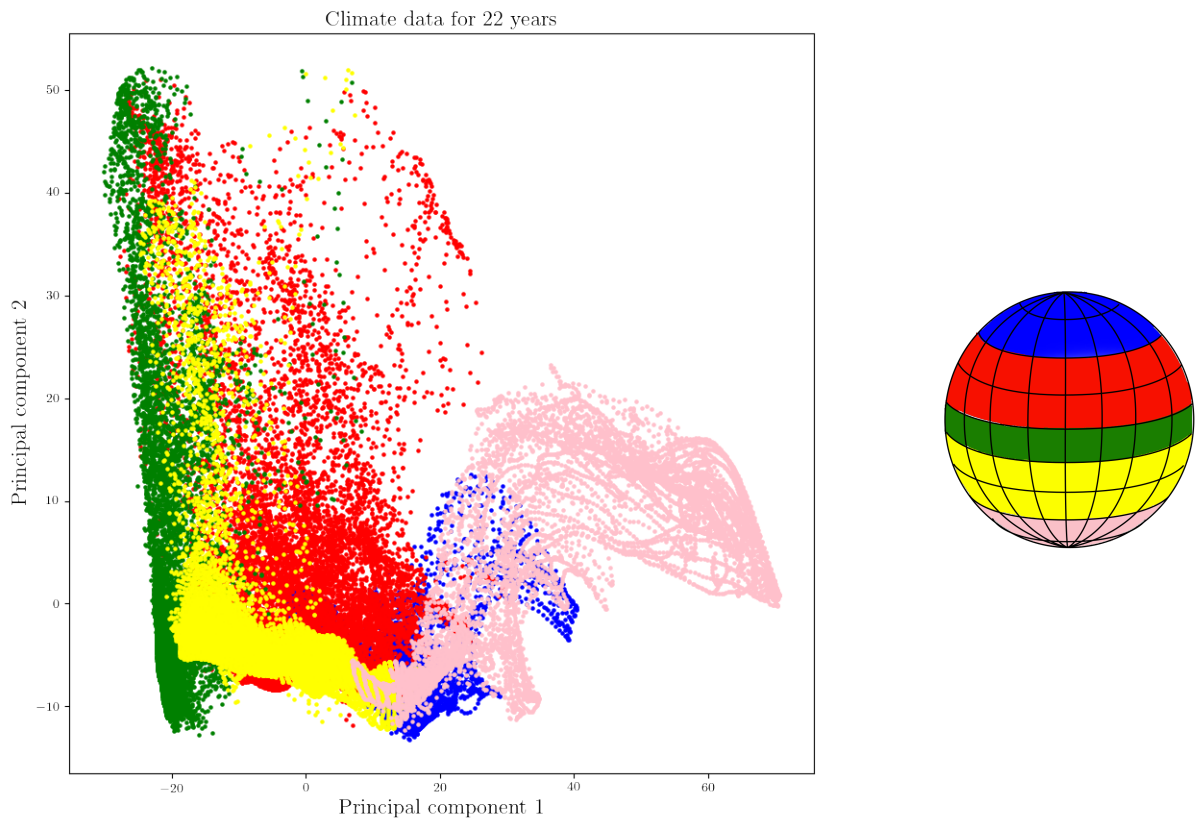


FIGURE 7 – PCA with two components

North	Temperate north	Equator	Temperate South	South
●	●	●	●	●

We can see on the graph that even if there is a strong dispersion, equator values are located at higher temperatures. On the contrary, north and south pole values are located at lower temperatures. Green and red classes overlap as these two categories have similar climate conditions.

Although it is an understandable figure, this is not satisfying.

4.1.2 Three dimensions

We attempted to run the PCA method with three principal components, it yielded similar results as linear PCA, but some areas with particular climate conditions stood out more strikingly. For example, deserts are located at the bottom of the peak on the third figure below, that maps the dry and hot areas.

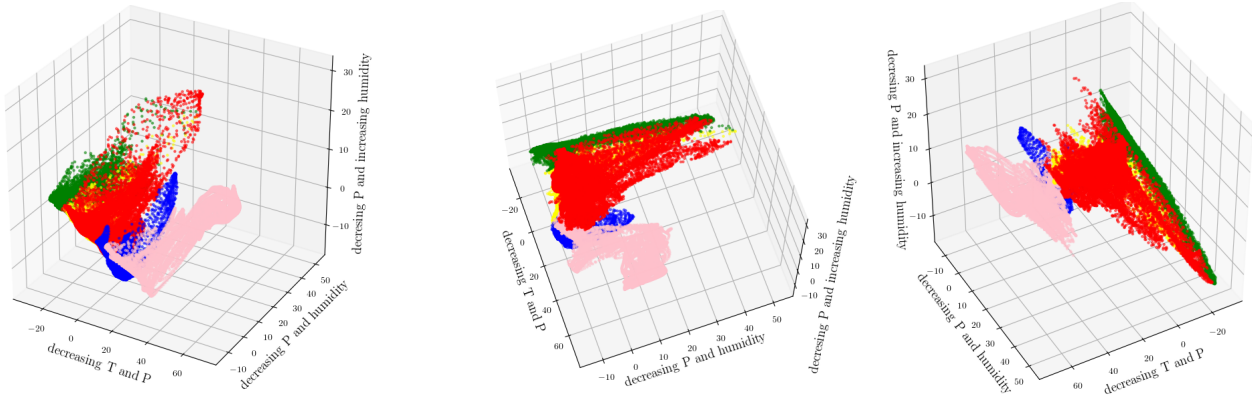


FIGURE 8 – 3D PCA

4.1.3 Kernel PCA

Kernel PCA has a complexity of $O(n^3)$ so we decided to run it on a subset of our dataset. Here is the two dimension result for all latitudes between longitudes -15 and +15.

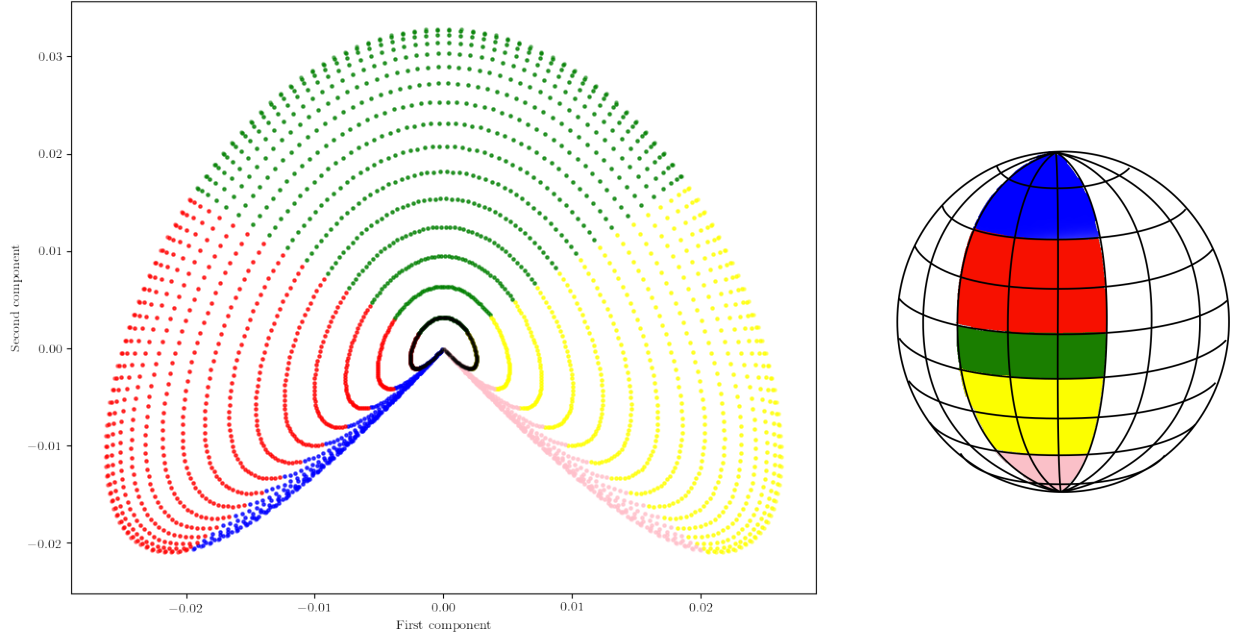


FIGURE 9 – Kernel PCA with two components and the RBF kernel

The black dots are all dots along one longitude, here 15. We see however that there is a degenerescence : -15 and +15 longitude are projected on the same positions. When changing the γ , some lines grow appart but most of them stay degenerated. Another flow of this representation is the space that is taken by the equator area. When representing the temperature in this area, we can see it does not vary very much compared to other regions : the variables should be represented closer together, as it was the case for example in PCA.

The peak present in the center of the figure is a consequence of the fact that data from each side of the peak even if they are close numerically are not close spatially : they are the data from both north and south pole.

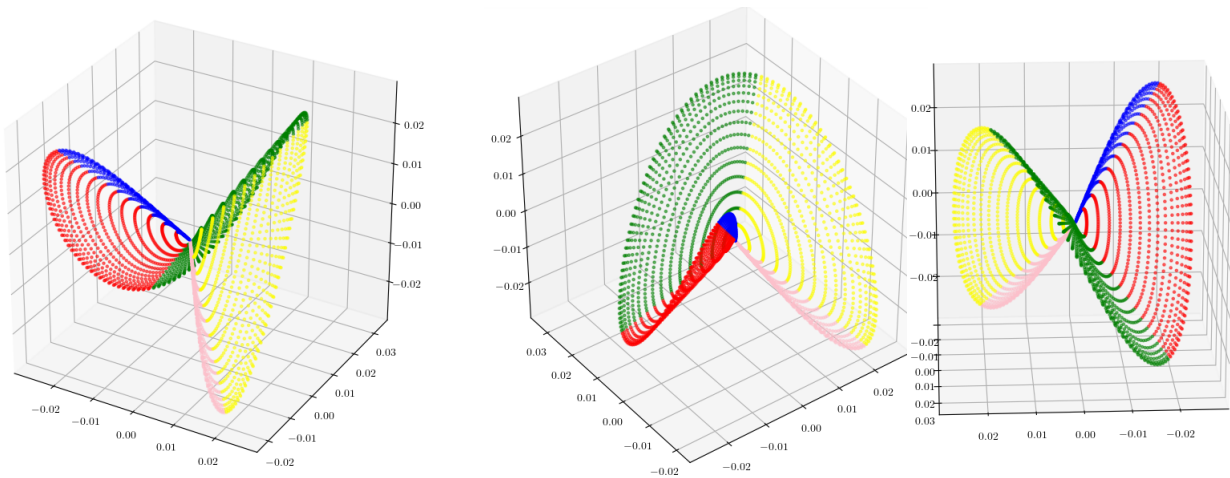


FIGURE 10 – 3D Kernel PCA

The dots representing the south and north poles are still closer together however, which might come from the smaller range of temperatures and climate conditions reached in these areas.

Changing the kernel function to a polynomial one however destroyed the degenerescence, but created a node on equatorial values as it can be observed just below. The purple and black lines represent the -15 and 15 longitudes : they are clearly different.

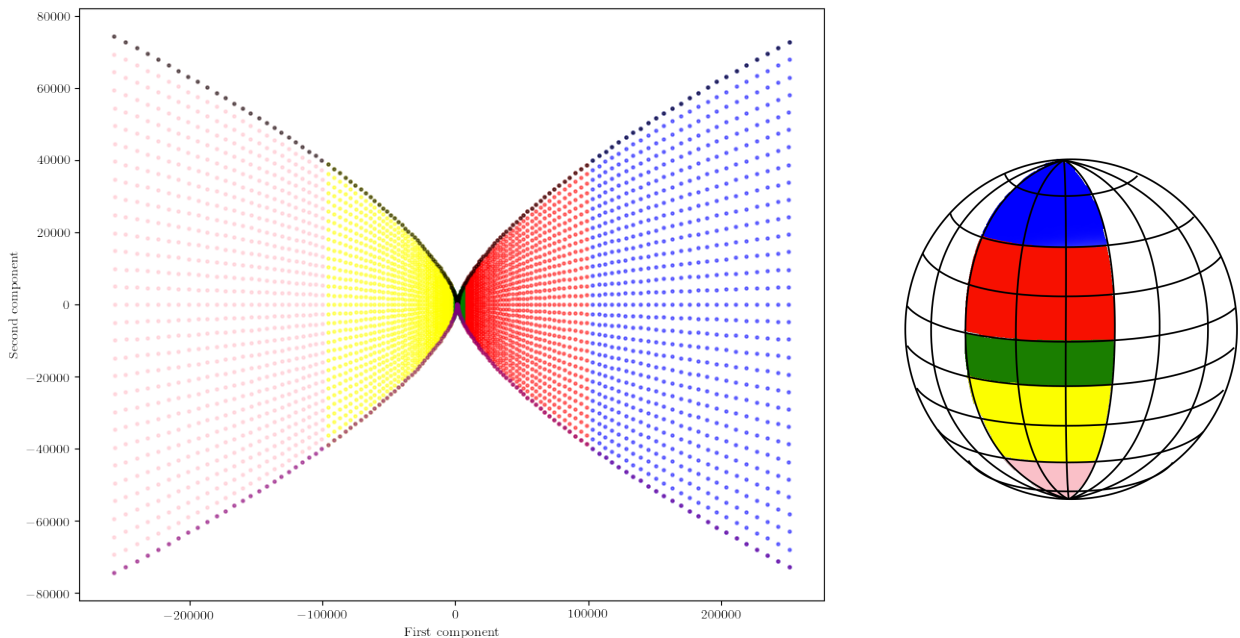


FIGURE 11 – Kernel PCA with two components and polynomial kernel

4.1.4 Multidimensional Scaling - MDS

As MDS was extremeley low and seemed to use a lot of memory, we could not run the programm for more than 600 datapoints (ie 600 rows).Therefore, we used a new subset of longitudes between 0° and $+5^\circ$, reducind the datasize to 546 datapoints. Figure 6 shows the results of MDS.

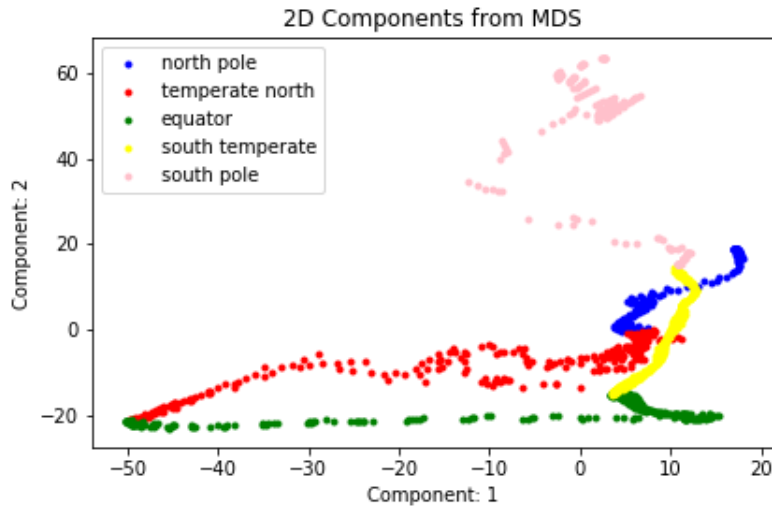


FIGURE 12 – Results of MDS in 2D

4.1.5 Isomap

As the main parameter of the method is the number of nearest neighbors taken into account, we present in figure ?? the result of isomap for number of neighbour between 1 to and neighbors.

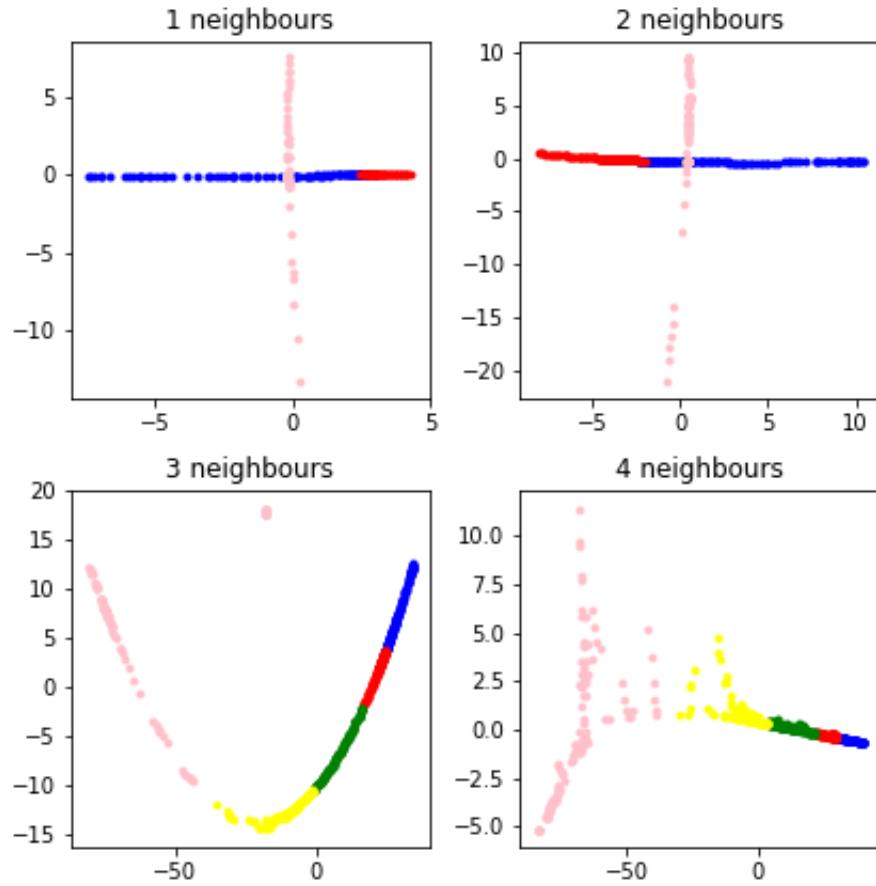


FIGURE 13 – results of isomap depending on the number of neighbors considered from 2 to 8

It is obvious that the number of neighbours considered strongly influence the 2D distribution of our dataset. To know which graph is the most accurate, we refer to the 'reconstruction error'. We computed this reconstruction error with respect to the number of neighbours considered. The results are presented in figure ??

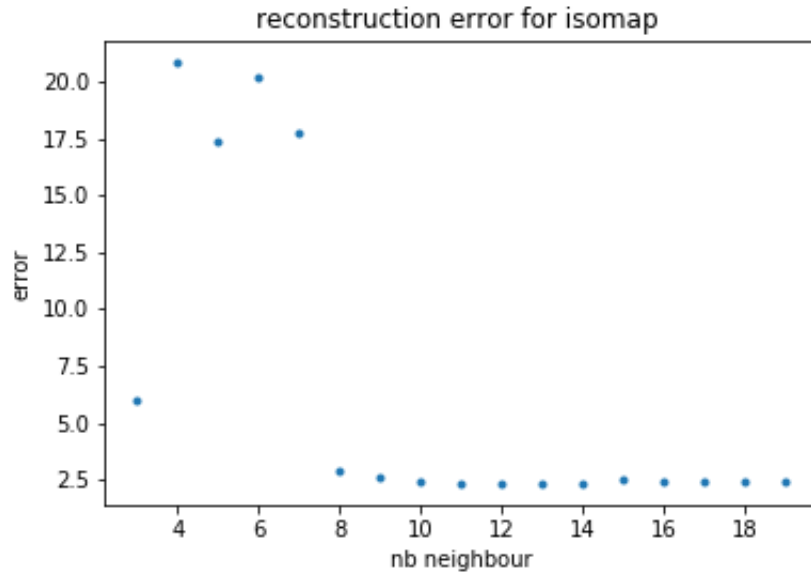


FIGURE 14 – reconstruction error as a function of number of neighbours considered

It is interesting to see that from eight neighbors, error seems to stabilize around 2,5. So the most 'accurate' representation should be with 8 neighbors.

4.2 Time comparison

We implemented manually PCA, but for the other methods we used the built in fonction of the scikit package on python. PCA's running time varies linearly with the number of rows as we can see on the graphic below

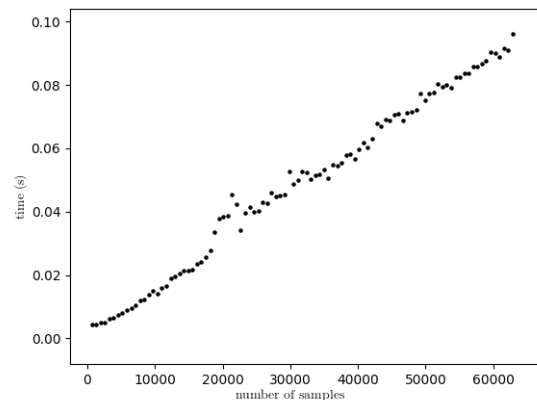


FIGURE 15 – running time for PCA

Kernel PCA was slower, we did not compute its running time for all the dataset, only until 12000 rows. It yields the following graph.

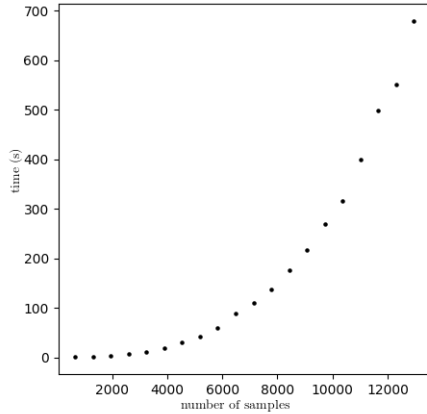


FIGURE 16 – running time for Kernel PCA

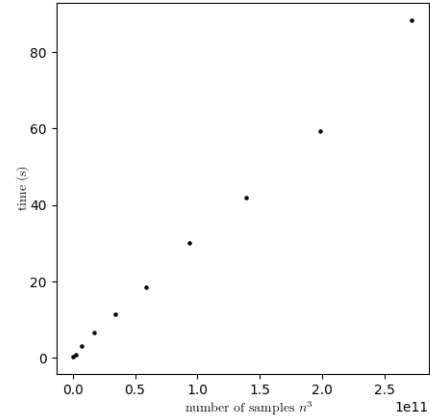


FIGURE 17 – Kernel PCA running time is a $O(n^3)$

In figure 10, we present computing time for different methods as a function of the sample size. From this we can conclude that the quickest method is PCA. Then comes kernel PCA, isomap and MDS. It is quite surprising to see that isomap is quicker than MDS for the same number of rows. Indeed, isomap uses the same algorithm than MDS but also needs to compute distance with neighbors first. It is thus expected to be slower. We found an explanation on the 'manifold guide' of scikit.

In Scikit-Learn implementation, Isomap algorithm runs faster than Multi Dimensional Scaling on the S-Curve dataset [...]. In the third stage of algorithm, the implementation uses Partial Eigen Value decomposition instead of MDS which is the version proposed by the researchers

We did not use the 'S curve dataset' as mentioned above, but we may think that due to the round shape of earth, our dataset is quite similar to the S-curve and thus Isomap runs faster than MDS. For a fixed dimension d , the following table present the complexity of every method as it appeared to us.

Method	PCA	Kernel PCA	MDS	Isomap
Complexity	$O(nr^2)$	$O(n^3)$	$O(n^3)$	

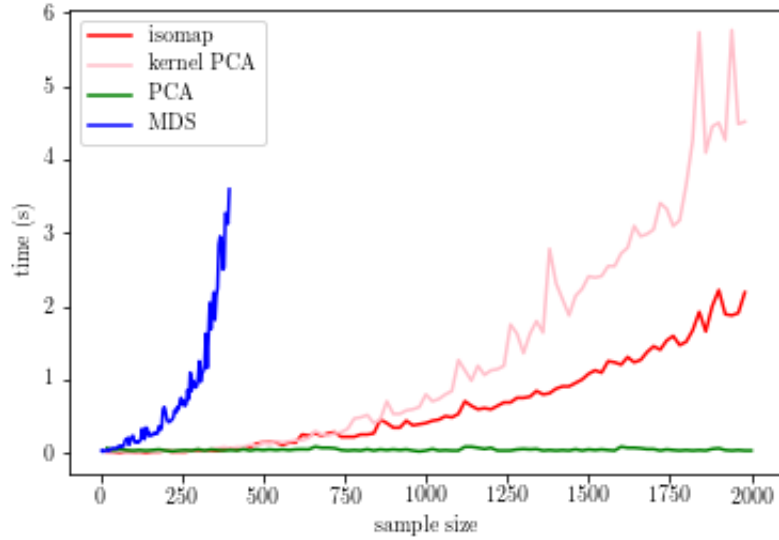


FIGURE 18 – computing time comparison

We also check how isomap computing time varies with the number of neighbor taken into account. Figure 11 shows that isomap efficiency time increases with the number of neighbors. So isomap efficiency of computing time may be limited by the number of neighbors.

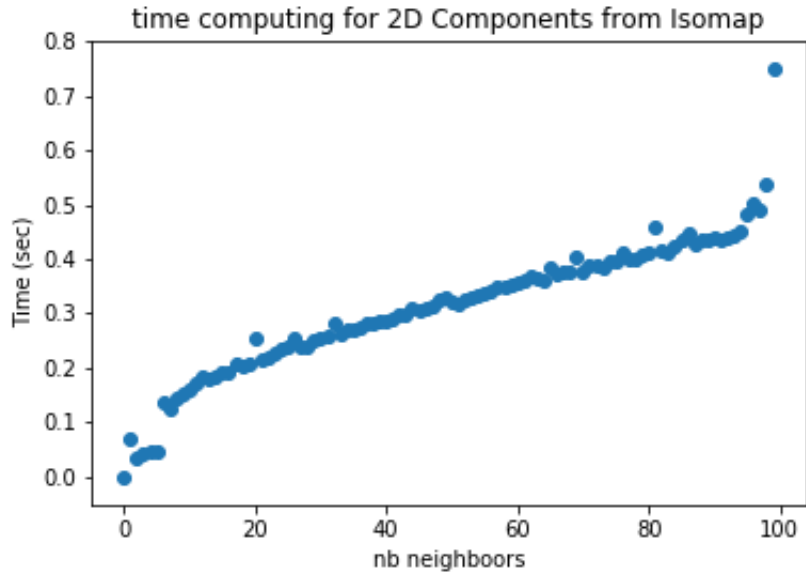


FIGURE 19 – computing time for isomap as a function of the number of neighbors considered

4.3 Error and preserved information

4.3.1 PCA

The main indicator of how much information has been preserved with PCA lies in eigenvalues. We plot the eigenvalues to see their relative importance in the dimension reduction.

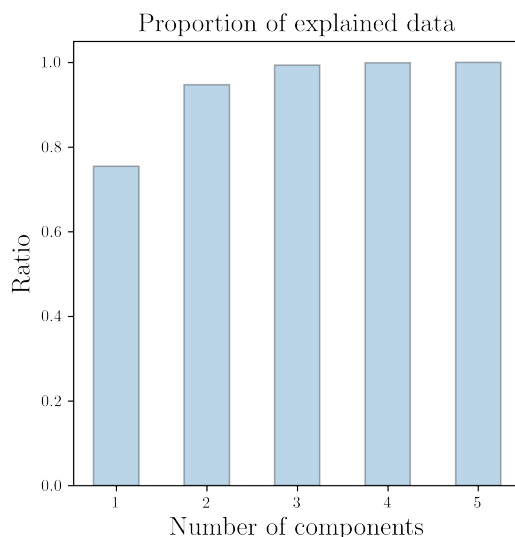


FIGURE 20 – PCA with two components

We can see that with only two components, we obtain an explained variance of 94%, which is quite enough to interpret our dataset. PCA seems to be a relevant technic to analyse data, and in particular to establish classification.

We also reconstructed the dataset from its projection with PCA. It is indeed possible to build an estimated \hat{X} dataset from the projected data of PCA. We plotted the first column of the dataset, the temperature for X and \hat{X} as a function of latitude and longitude. Even if the curve is reversed and the numerical values are not correct, the structure of the data is saved.

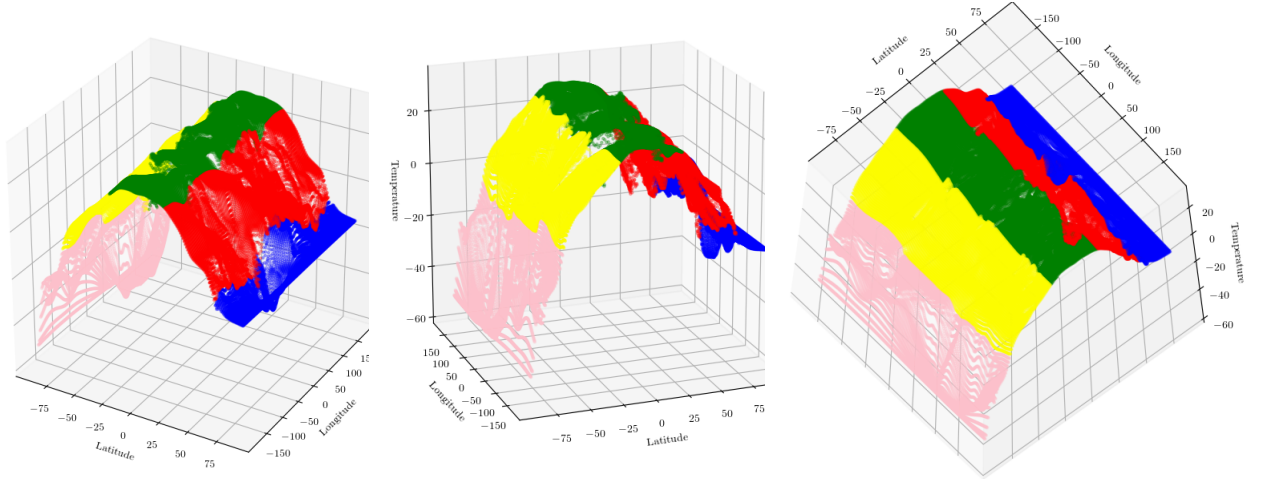


FIGURE 21 – Temperature as a function of latitude and longitude in X

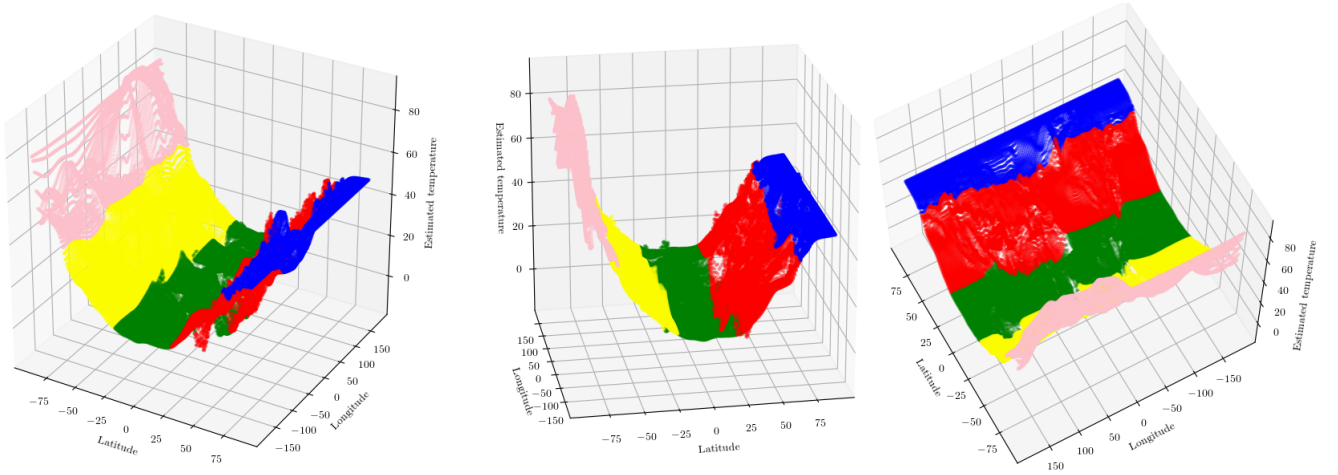


FIGURE 22 – Temperature as a function of latitude and longitude in \hat{X}

4.3.2 MDS

One indicator to estimate the reconstruction is accurate is the 'stress'. It is roughly the square difference between the final distance of two items in the MDS model and the true distance between them. The exact formula is :

$$s = \sqrt{\frac{\sum (d_{ij} - d(i, j))^2}{\sum d_{ij}^2}}$$

The aim of MDS is to minimize this stress. Figure 7 presents the value of stress depending on the number of lines. We can see it is relatively low ($<0,002$) with our dataset.

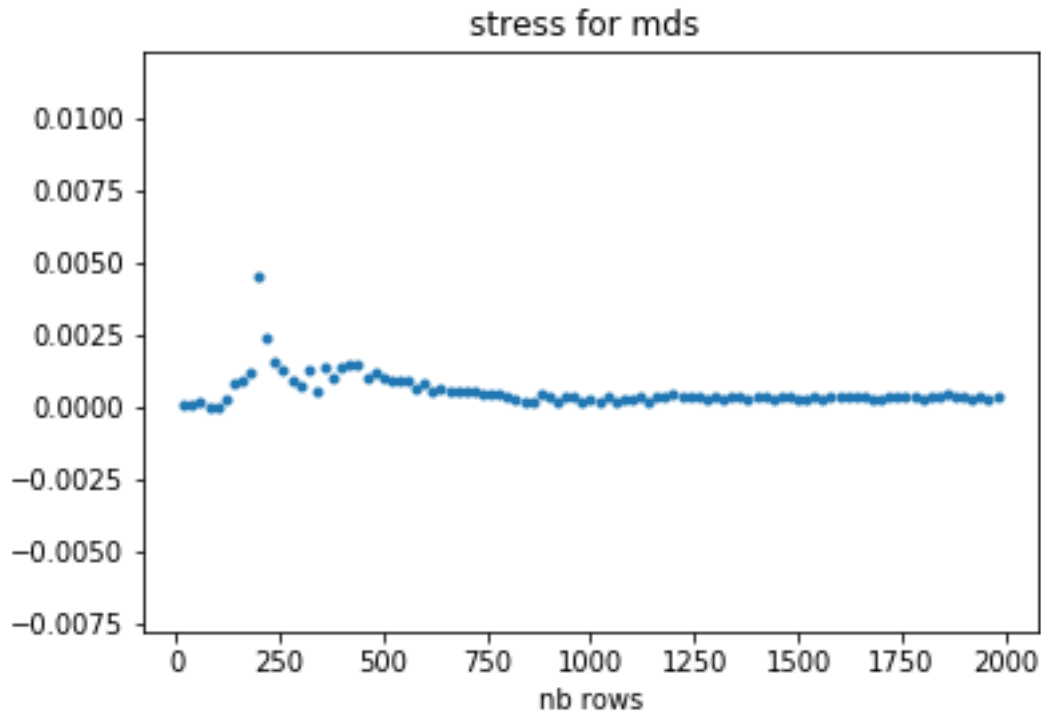


FIGURE 23 – stress as a function of sample size

4.3.3 isomap

We have already talked about error reconstruction of isomap, a function already embedded in scikit. It may be interesting to see how reconstruction error evolve with the number of dimension taken into account. On figure ?? we compare how reconstruction error evolve with dimension ? We make dimension vary between 2 and 3.

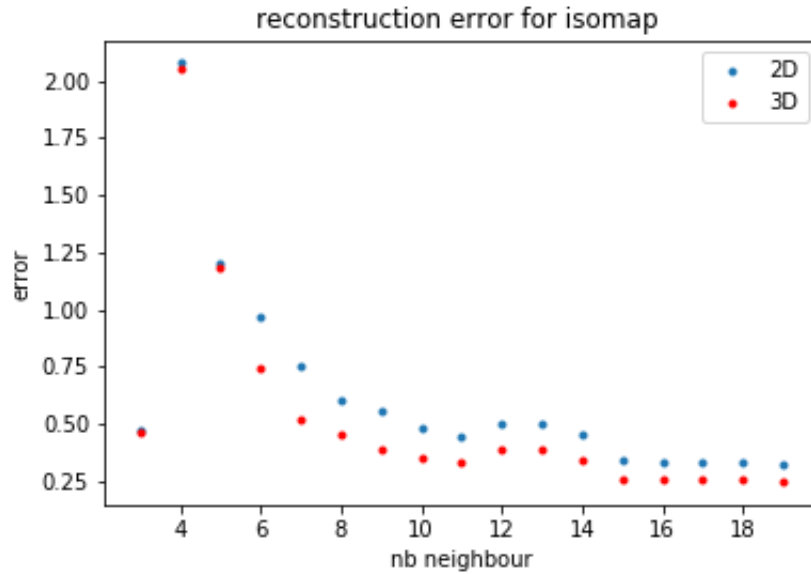


FIGURE 24 – reconstruction error as a function of the datasize

We can clearly see that error is slimmer in 3D than in 2D. It seems logic as in 3D we reduce less dimension than in 2D. But as the number of neighbors increases, both of 2D and 3D reconstruction error seem to stabilize around ?.

We also computed the error reconstruction as a function of the sample size in figure ??. This error keeps increasing with the dataset size, so isomap accuracy seems to be limited by the size of the dataset.

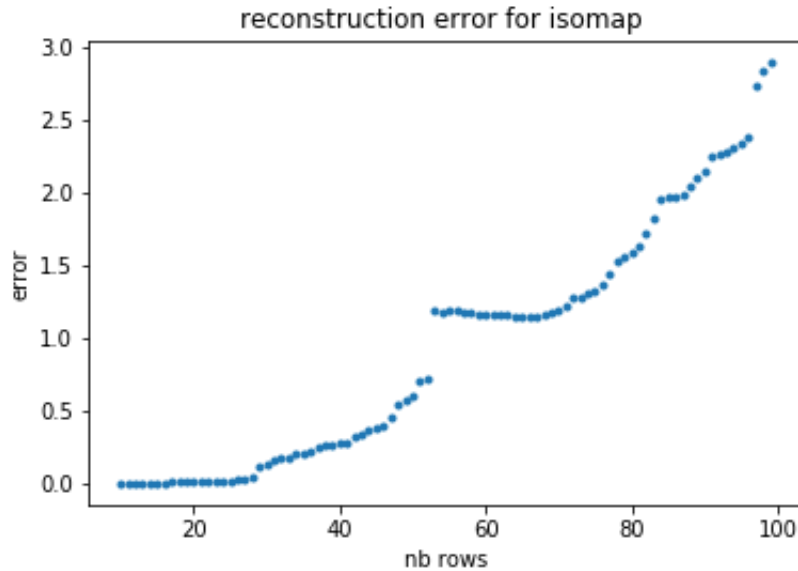


FIGURE 25 – reconstruction error as a function of the datasize

5 Conclusion

6 Bibliography

Datasets

<https://eosweb.larc.nasa.gov/cgi-bin/sse/global.cgi?email=skip@larc.nasa.gov> ↑[3]

John P. CUNNINGHAM, Zoubin GHAHRAMANI, Linear Dimensionality Reduction : Survey, Insights, and Generalizations, , *Journal of Machine Learning Research* 16 (2015) 2859-2900

Alon ORLITSKY, SAJAMA, Supervised dimensionality reduction using mixture models, Proceedings of the 22nd international conference on Machine learning, p.768-775, August 07-11, 2005, Bonn, Germany