University of Minnesota, Twin Cities

Masters Thesis

Department of Computer Science

A Study of Dimensionality Reduction Techniques and its Analysis on Climate Data

A THESIS SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA BY

Arjun Kumar

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Vipin Kumar

October 2015

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in graduate school. I would like to thank Jaya and Saurabh for their valuable inputs. Dr Stefan Leiss for his valuable guidance and insights. Professor Arindam and Professor Ansu for their guidance, my parents and brother for their constant encouragement and support and finally my advisor Professor Vipin for continuous support to my efforts without which this would not have been possible...

Dedicated to my parents and brother...

Abstract

Dimensionality reduction is a significant problem across a wide variety of domains such as pattern recognition, data compression, image segmentation and clustering. Different methods exploit different features in the data to reduce dimensionality. Principle component Analysis is one such method that exploits the variance in data to embed data onto a lower dimensional space called the principle component space. These are linear techniques which can be expressed in the form $B = TX$ where T is the transformation matrix that acts on the data matrix X to the reduced dimensionality representation B. Other linear techniques explored are Factor Analysis and Dictionary Learning. In many problems, the observations are high-dimensional but we may have reason to believe that the they lie near a lower-dimensional manifold. In other words, we may believe that high-dimensional data are multiple, indirect measurements of an underlying source, which typically cannot be directly measured. Learning a suitable low-dimensional manifold from high-dimensional data is essentially the same as learning this underlying source. Techniques such as ISOMAP, Locally Linear Embedding, Laplacian EigenMaps (LEMs) and many others try to embed the high-dimensional observations in the non-linear space onto a low dimensional manifold. We will explore these methods making comparative studies and their applications in the domain of climate science.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Dimensionality reduction is a significant problem across a wide variety of domains such as pattern recognition, data compression, image segmentation and clustering. Different methods exploit different features in the data to reduce dimensionality. Principle component Analysis is one such method that exploits the variance in data to embed data onto a lower dimensional space called the principle component space. These are linear techniques which can be expressed in the form $B = TX$ where T is the transformation matrix that acts on the data matrix X to get the reduced dimensionality representation B. Other linear techniques explored are Factor Analysis and Dictionary Learning. In many problems, the observations are high-dimensional but we may have reason to believe that the they lie near a lower-dimensional manifold. In other words, we may believe that high-dimensional data are multiple, indirect measurements of an underlying source, which typically cannot be directly measured. Learning a suitable low-dimensional manifold from high-dimensional data is essentially the same as learning this underlying source. Techniques such as ISOMAP, Locally Linear Embedding, Laplacian EigenMaps (LEMs) and many others try to embed the high-dimensional observations in the non-linear space onto a low dimensional manifold.

In climate data, we would like to find factors that have high interpretibility. Principle component analysis only establishes components that have high variability. But this does not tell us if there is a pattern in the data.

PCA finds patterns of high variability. Thus both dipoles and monopoles can be found using this technique. In fact any oscillatory patterns that show high variability are caught during EOF analysis and they become even more prominent when two regions are acting in opposition to each other. It is important to remove trends and seasonality in data as they are the first to be picked in EOF anaysis.

We will talk more on EOF analysis, its literature and its implementation along with rotation techniques such as varimax and promax in the following chapters. A detailed discussion will be made on other linear techniques of dimensionality reduction such as factor analysis, probabilistic PCA and Dictionary Learning. We will then introduce Non-linear methods of Dimensionality reduction based on the concept of manifold learning such as Isomaps, Laplacian EigenMaps, Locally Linear Embedding (LLE) to further explain the structure of climate data. Since climate processes are governed by a number of non-linear variables it is possible that a non-linear dimensional reduction can reduce the data onto a linear subspace that can help explain the underlying factors of climate teleconnections.

Many other well known dipoles such as the Southern Oscillation (SO), Western Pacific(WP), Pacific/North Atlantic (PNA) have been discovered by manually comparing time series data at these locations [6]. Manual analysis is not complete and may miss essential patterns that have not been known before. Dimensional reduction techniques to find patterns such as the Emperical Orthogonal Functions (EOFs) [7] have been used for a long time. The Arctic Oscillation (AO) and the Antarctic Oscillation (AAO) have been found using EOF analysis. EOF technique is thus a very useful technique in finding dipoles but these do not always necessarily give physically meaningful modes[8]. Unless a dipole is physcially present in the region it becomes uncertain if the time series associated with the first principle component is indeed a physically meaningful mode. One can speculate if techniques such as rotation of principle components leads to physically meaningful modes but this looks like an open ended question[9]. These problems can be overcome by doing EOF analysis in a small region where we know the dipole is present, but again we need to know where the dipole is present which is equivalent to discovering the dipole manually. Graph based techniques[10] help overcoming some of these challenges and tends to give regions that may possibly have a dipolic behaviour. Such techniques give many candidate dipoles which require further filtering to find the statistically significant ones among them.

Chapter 2

Data Description

2.1 Dataset

We use sea level pressure (SLP) data to find the dipoles because most of the important climate indices are based upon pressure variability. We analyze three reanalysis datasets. Reanalysis projects create gridded datasets for all the locations on the globe by assimilating remote and in situ sensor measurements using a numerical climate model to achieve physical consistency and interpolation for global coverage. In the absence of a global data of observations, the reanalysis datasets are considered the best available proxy for global observations. Such datasets are produced by modeling groups around the world, including the NCEP/NCAR Reanalysis project [5], the European Reanlysis project [11], and the Japanese Reanalysis project [12]. Table 2 shows the summary of the details of the three reanalysis datasets. We present most of our results using the NCEP data as it is the longest reanalysis dataset. The NCEP data spans 1948present and there are 10512 grid points in the 2.5 degree resolution data. We use monthly mean values for the 60 years of data (corresponding to 720 monthly values). The NCEP/NCAR reanalysis is provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA [5], available for public download at [13]

2.2 Seasonality Removal

An important component of Earth Science data is the seasonal variation in the time series. The change in seasons brings about annual changes in the climate of the Earth such as increase in temperature in the summer season and decrease in temperature in the winter season. The seasonality component is the most dominant component in the Earth science data. For example, consider the time series of monthly values of air temperature

at Minneapolis from 1948-1968 as shown in the Figure 3. From the figure, we see that there is a very strong annual cycle in the data. The peaks and valleys in the data correspond to the summer and winter season respectively and occur every year. The seasonal patterns even though important are generally known and hence uninteresting to study. Mostly, scientists are interested in finding non-seasonal patterns and long term variations in the data. As a result of the effect of seasonal patterns, other signals in the data like long term decadal oscillations, trends, etc. are suppressed and hence it is necessary to remove them. Climate scientists usually aim at studying deviations beyond the normal in the data.

Two methods are discussed here that remove seasonality. One is by obtaining the mean-scored anomalies and the other is by obtaining the z-scored anomalies. The mean scored anomalies are constructed as follows.

$$\mu_m = \frac{1}{end - start + 1} \sum_{y=start}^{end} x_y(m), \forall m \in 1 - 12 \tag{2.1}$$

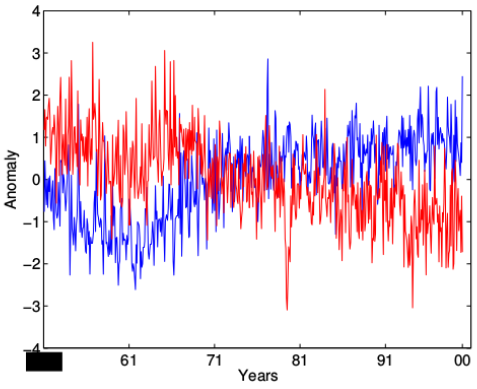
$$x_y(m) = x_y(m) - \mu_m \tag{2.2}$$

In this equation, start and end represent the start and end years to consider for the mean and define the base for computing the mean for subtraction (for example 1948 and 2009 for the NCEP data). mu_m is the mean of the month m and $x_y(m)$ represents the value of pressure for the month m and year y. Once we remove the monthly means, the resulting values are the anomaly time series for that location. Although, removal of monthly means is the most popularly used approach to construct anomalies, they can also be constructed by alternate measures like using the z-score.

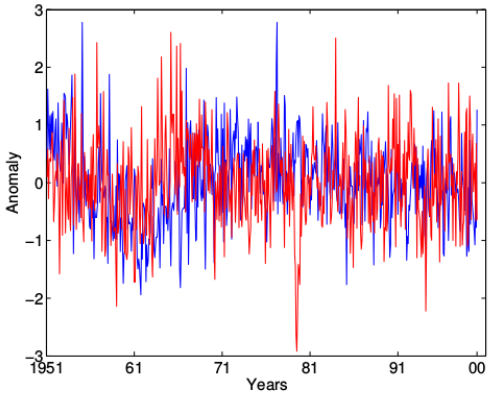
2.3 Detrending

Another important aspect of climate data is the presence of long term increasing or decreasing trends in the data. If there are two regions with strong trends in the opposite direction then it could result in spurious negative correlations between the two regions. For, example consider the pressure anomaly time series at two locations as shown in the figure (a) in the NCEP data during the time period 1951-2000. From the figure, we see that the two locations have trends in the opposite direction and hence as a result have a high negative correlation between their anomalies (-0.46 in this case). A possible approach to handle the trends in climate data that is widely used by climate scientists is to detrend the data before any possible analysis. If we detrend the data in the example

above, we no longer see the high negative correlation between the two time series as shown by the figure (b). Throughout this paper, we use detrended data to present our results. But, we acknowledge that detrending of non-stationary time series data itself has several issues and may result in removing connections or adding spurious ones, which might require a detailed investigation. Figures courtest of [14]



(a) Raw anomaly time series at the two locations showing trends in the opposite direction and a possible dipole having a correlation -0.4616



(b) Anomaly time series of the two locations after detrending has a correlation -0.0470

Chapter 3

EOF Analysis in Climate Science

EOF analysis (better known as Principle Component Analysis in Computer Science) is the most popular technique use in climate science to find teleconnections patterns in climate data in a given variable. Most teleconnection analysis is carried forth in pressure and temperature data. Climate being highly nonlinear and high dimensional presents a challenging task to find important patterns and ways to characterize them. EOF analysis, one of the most widely used methods in climate science is one such technique. Eofs have been used for analyzing the nature of climate oscillations all over the globe. It finds the variability in a field such as the Sea Level Pressure(SLP), Sea Surface Temperature. The pattern that an EOF represents when plotted on a map is a standing oscillation and the time evolution of the EOF shows how it oscillates in time [5]. The EOF method is thus a map series method of analysis that looks at the variability in the time evolving field of all spatial locations and breaks it into a few standing oscillations and a time series to go with each oscillation. EOF analysis is used to find important oscillations such as the Antarctic and Arctic Oscillations and other northern oscillations in SLP, the Madden Jullian oscillation, the Indian ocean dipole and tropical atlantic dipole in SST [1][6].EOF technique has been widely used in the climate domain to find well known teleconnections called dipoles. Dipoles are defined as a pair of regions such that locations within each region are highly positively correlated with each other and locations across these regions are negatively correlated to each other. EOF is a very simple technique and exploits the variability in the data to find such dipoles. It acts on the data on a global scale and gives a set of eigen vectors that correlates well with the major dipole present in that region. A simple method that climate scientists use to find dipoles are look at a small region where we know only one dipole exists and performing EOF on that region alone. This is done because it is very difficult to tell physically what information the 2nd eigen vector holds as it is constrained to be orthogonal to the 1st eigen vector. The 1st eigen vector captures the maximum variability in the data but

this generally does not correspond to the strongest dipole present in that region. EOF does not exploit the nature of the data to find dipoles more efficiently and accurately.

3.1 Drawbacks of EOF

This method despite its simplicity gives a plethora of information and hence is used by climate scientists. The EOF analysis, due to its simplicity fails on many aspects as well. The major constraint in EOF analysis is that the principle components need to be orthogonal to each other. Thus it is uncertain if the PCs of 3rd order or more have any physical significance attached to them or not [3][4]. The method also fails when you have prominent oscillations that are not spatially orthogonal to each other but are nearly statistically independent in the time domain. There are also variations to the EOF method. The varimax rotation of EOFs finds modes that are more localized in space than the standard EOF modes [1]. In varimax one needs to set more parameters than normal EOF analysis and hence is more subjective. We also have to be careful in interpreting the EOF or rotated EOF modes as physical modes as either one or both could be significant. [1] gives situations where the physical modes could be localized and they have nearly equal variances. Here EOFs may be misleading but by the study conducted by [2], EOF outperforms varimax and regression analysis. Thus the results of the EOF and rotated EOF can be complimentary rising eyebrows on the possible credibility of these methods and confusion in the climate community as to which method needs to be applied in a given scenario which requires a good domain knowledge leading to introduction of new parameters. It is thus important to introduce a non-parametric way of finding oscillations that do not rely on domain knowledge. EOF analysis may also result in spurious patterns and result in two camps arguing if the dipole is a manifestation of a physical process or an artifact.

3.2 Does EOF always work?

Climate oscillations can be broadly classified into two major categories. The first type are the dipoles that are a pair of regions whose intra cluster correlation is positive and inter cluster correlation is negative. The other type are the monopoles that are singular regions around the globe that are periodic in nature and are quite dominant in a particular spectral band. EOF can capture dipoles prominently in a given region only. If EOF analysis is done in a region that has only one dipole, then it would capture it very accurately considering noise in the data is minimum. If there are two dipoles and they are nearly independent (correlation 0), Then the dominant dipole with bigger clusters

or larger magnitude time series get captured in the 1st EOF pattern and the second less dominant dipole gets captured by the second EOF pattern. If the 2 dipole indices are not independent, i.e they have some correlation, then both the EOFs capture a mixed signal which is difficult to segragate and hence in climate science EOF is generally done in a region where we know for sure only one dipole exists. To overcome this challenge climate scientists use techniques such as varimax rotation (have to explain this properly). EOF technique in general is applied to the entire time series with no knowledge of its spectral properties. A monopole has a high postive intra cluster correlation only in a particular spectral band and this is hidden when applying EOF. Thus EOF cannot ideally find monopoles.

3.3 Comparison with Dynamic Dipole Discovery Framework

The dynamic discovery of dipoles framework [7] can capture multiple oscillations all over the globe which the EOF methodology cannot handle. Thus we cannot find global modes using EOF which can be captured by the dipole analysis as it is not dimension dependent and does not depend on the strength of the mode but simply on the presence of the mode. It also has problems in finding the dominant centers of interaction[1] which is not a problem in our algorithm as we also find the local attractors for all the clusters found. The EOF methodology can find standing oscillations only and cannot work on moving oscillations which the dynamic dipole discovery framework can handle. Again we have the problem that the modes of variability need to be orthogonal to each other in space and time. The modes are generally never orthogonal to each other and thus any mode we find is actually never a pure mode but a mixture of oscillations. This is not the case with our algorithm as it simply looks for clusters that are wholly negatively correlated with each other. The EOF method finds regions of high variations, thus not being streamlined to find monopoles or dipoles separately. A monopole can be looked at as a set of points lying in the 1st quadrant of a 2-d plane whereas a dipole has points lying in the 1st and 3rd quadrant. As EOF simply finds directions of maximum variance which is independent of the spatial distribution of points. It is thus a top-down approach. The bottom-up dipole discovery algorithm starts from the basic definition of a dipole and proceeds to find larges clusters that are negatively correlated, The physical interpretibilty of the dipoles found by this algorithm is very high.

Chapter 4

Linear Techniques for Dimensionality Reduction

4.1 Introduction

In this section we will cover some of the linear techniques of dimensionality reduction such as PCA,Factor Analysis and Dictionary Learning, their application on climate data and an introduction of rotation techniques and compare them with the modern frameworks available. There are other dimensionality reduction techniques such as Multi Dimensional Scaling, sparse PCA and Kernel PCA which will be mentioned within the confines of the above algorithm.

4.2 Factor Analysis - the Fundamental Equation

Let $\mathbf{Y}$ be an $N \times 1$ vector whose values are the observed random variables $Y_i \forall i = 1, 2..N$. Without loss of generality assume $E(\mathbf{Y}) = 0$ and that $E(YY^t) = S_{yy}$ is the covariance matrix. We could take a correlation matrix as well in which case $E(YY^t) = R_{yy}$, Y_i are normalized in which case the diagonal elements will be 1. Let $\mathbf{X}$ be an $r \times 1$ vector whose variables are the factors $X_i \forall i = 1..r$. Let R_{xx} be the correlation matrix. Let E be an $n \times 1$ random vector whose variables are the unique factors $\varepsilon_i \forall i = 1..N$. We assume that the unique factors have 0 mean and unit variance and are mutually uncorrelated. Finally, let A be the factor pattern coefficients which is an $n \times r$ matrix and Ψ an $n \times n$ diagonal matrix which gives a magnitude to the unique factors ε_i.[15]

$$y_1 = a_{11}x_1 + a_{12}x_2.....a_{1n}x_n + \Psi_1 e_1 \tag{4.1}$$

$$y_2 = a_{21}x_1 + a_{22}x_2.....a_{2n}x_n + \Psi_2 e_1 \tag{4.2}$$

$$y_3 = a_{31}x_1 + a_{32}x_2.....a_{3n}x_n + \Psi_3 e_1 \tag{4.3}$$

$$y_i = a_{i1}x_1 + a_{i2}x_2.....a_{in}x_n + \Psi_i e_1 \tag{4.4}$$

$$Y = AX + \Psi E \tag{4.5}$$

$$\tag{4.6}$$

Through simple calculations, we get

$$R_{yy} = AR_{xx}A^t + \Psi^2 \tag{4.7}$$

$$\text{or} \tag{4.8}$$

$$S_{yy} = AR_{xx}A^t + \Psi^2 \tag{4.9}$$

If all factors are orthogonal to each other with unit variance, then R_{xx} will be an Identity matrix. Thus the factor equation reduces to

$$R_{yy} = AA^t + \Psi^2 \tag{4.10}$$

$$\tag{4.11}$$

In PCA, we use the correlation matrix R_{yy} to compute the principle components. Here in factor analysis we are given with $R_{AA} = AA^t$ and we need to compute the principle components in a manner of speaking for R_{AA}. Since Ψ^2 is diagonal, covariances are represented by A. Note that PCA does not allow a separate Ψ^2 and it tires to account for both the covarianes and the variances. When $\Psi_{ii} = \Psi_{jj} \forall i, j \in 1..n$, then we have probabilistic PCA[16] and the conventional PCA is when $\Psi_{ii} = 0$. Thus the factor matrix that one would be factorizing in order to get the factors for PCA would have 1 on the diagonal, whereas for factor analysis, the values will be slightly less than 1, remainder of the values being the same. Thus even though the two techniques were developed independently based on different assumptions, the resultant math boils down to be very similar.[15]

4.3 Principle component Analysis

In a few words, a principle component is a directional vector that explains the maximum separation of data when data is projected onto it. The percentage of variance contributed is contained in the eigen value. The remainder variance is contained in the subspace orthogonal to the first principle component u_1. The second principle component, u_2 lies in this orthogonal subspace (which implies $u_1 \perp u_2$) and is such that the projection of all points onto u_2 gives maximum separation of points. Mathematically it can be explained below as follows

Let p_1, p_2, p_3..... p_N be N points in a space spanning d-dimensions.Without loss of generality we will assume that the data is mean centric. i.e

$$\frac{1}{N}\sum_i p_i = 0 \tag{4.12}$$

The data covariance matrix is given by

$$S = \frac{1}{N}\sum_i p_i p_i^t \tag{4.13}$$

Thus to maximize the projected variance $u_1^t S u_1$ with respect to u_1 which becomes a constrained maximization problem that we are already familiar with

$$u_1^t S u_1 + \lambda(1 - u_1^t u_1 \tag{4.14}$$

Now we would like to project all points on the subspace to orthogonal to u_1 i.e new $p_2 i = p_i - (p_i^t u_1)u_1$.

Thus the new covariance matrix in this projected space will be

$$S_{new} = \frac{1}{N}\sum_i (p_i - (p_i^t u_1)u_1)(p_i - (p_i^t u_1)u_1)^t \tag{4.15}$$

$$\tag{4.16}$$

Now we need to find a vector in this space that maximizes the separation of the projection of $p_2 i$. Maximizing the new projected variance along u_2 under the known conditions we get

$$u_2^t S_{new} u_2 + \lambda_1(1 - u_2^t u_2) + \lambda_2(u_1^t u_2) \tag{4.17}$$

Simplifying $u_2^t S_{new} u_2$ gives

$$u_2^t(\frac{1}{N}\sum_i (p_i - (p_i^t u_1)u_1)(p_i - (p_i^t u_1)u_1)^t)u_2 \tag{4.18}$$

$$= \frac{1}{N}\sum_i u_2^t((p_i - (p_i^t u_1)u_1)(p_i - (p_i^t u_1)u_1)^t)u_2 \tag{4.19}$$

$$= \frac{1}{N}\sum_i (u_2^t(p_i - (p_i^t u_1)u_1)u_2^t(p_i - (p_i^t u_1)u_1)) \tag{4.20}$$

$$= \frac{1}{N}\sum_i (u_2^t p_i - u_2^t(p_i^t u_1)u_1)^2 \tag{4.21}$$

$$= \frac{1}{N}\sum_i (u_2^t p_i - (p_i^t u_1)u_2^t u_1)^2 \tag{4.22}$$

$$= \frac{1}{N}\sum_i (u_2^t p_i)^2 \because u_1 \perp u_2 \tag{4.23}$$

$$= \frac{1}{N}\sum_i u_2^t p_i p_i^t u_2 \tag{4.24}$$

$$= u_2^t S u_2 \tag{4.25}$$

4.4 Principle component Analysis vs Factor Analysis

There is a good deal of overlap in terminology and goals between Principal Components Analysis (PCA) and Factor Analysis (FA). Both are two completely different techniques and serve different purposes but are sometimes statistically mistaken to be the same. Much of the literature on the two methods does not distinguish between them, and some algorithms for fitting the Factor Analysis model involve PCA. Both are dimension-reduction techniques, in the sense that they can be used to replace a large set of observed variables with a smaller set of new variables. They also often give similar results. However, the two methods are different in their goals and in their underlying models. Roughly speaking, you should use PCA when you simply need to summarize or approximate your data using fewer dimensions (to visualize it, for example), and you should use FA when you need an explanatory model for the correlations among your data.

Let us look at the difference in the final results between Factor Analysis and PCA from a mathematical stand point as shown below.[16]

$$FA : R_{yy} = AA^t + \Psi^2 \tag{4.26}$$

$$PCA : R_{yy} = AA^t \tag{4.27}$$

$$FA = \begin{bmatrix} c_{11} - \Psi_1^2 & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} - \Psi_2^2 & \dots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{kk} - \Psi_k^2 \end{bmatrix}$$

$$PCA = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{kk} \end{bmatrix}$$

$$\text{Probabilistic PCA} = \begin{bmatrix} c_{11} - \Psi^2 & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} - \Psi^2 & \dots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{kk} - \Psi^2 \end{bmatrix}$$

The final adjacency matrix upon which spectral decomposition is being done is AA^t. Thus for FA we are doing a spectral decomposition on $R_{yy} - \Psi^2$ and for PCA we are working on the correlation matrix itself i.e R_{yy}. Thus if the unique factors for each factor component is small, the results of factor analysis and PCA are very much the same. The values of $\Psi_i \forall i = 1..N$ will be large only if each and every data point are more or less independent of each other. Climate data lies on a high dimensional manifold that has a smooth structure due to its continuity and following physical laws. On a global scale on large timescales, climate data is less chaotic leading to a smoother topology. Thus intuitively the $\Psi - i$ will be small due to large spatio-temporal autocorrelation.

4.5 PCA as a correlation preserving embedding

As will be discussed in the Non-linear dimensionality reduction section, where the techniques try to embed onto lower dimensions using a euclidean distance metric, we can draw parallels to PCA technique where the correlations are preserved in the lower dimensional manifold. Comparing the two techniques.

TABLE 4.1: Comparison of PCA and Laplacian eigenmaps

	Laplacian	PCA
Weight (W_{ij})	$\exp^{\|\|x_i - x_j\|\|_2^2}$	$cor(x_i, x_j)$
Objective	Minimize	Maximize
Objective function	$\sum_{ij} W_{ij} \|\|y_i - y_j\|\|_2^2$	$\sum_{ij} W_{ij} y_i^t y_j$
Matrix form	$y^t(D - L)y$	$y^t \Sigma y$

4.6 Rotation Techniques

Rotation Techniques have been known for many decades now and form a very important caveat of EOF analysis. Suggested procedures for analytic rotation vary from quartimax [17], which maximizes the sum of fourth powers of factor loadings, to varimax [18], which maximizes the variance of the factors squared loadings, to maxplane [19], which maximizes hyperplane count. Other procedures-such as covarimin [20], biquartimin [21], direct oblimin [22] and promax [23] have also been suggested. Differences in factor analytic technique can, of course, produce different results.

The purpose of rotation is to make the rotated factor loading matrix have some desirable properties. One of the methods used is to rotate the factor loading matrix such that the rotated matrix will have a simple structure.

L. Thurstone introduced the Principle of Simple Structure, as a general guide for factor rotation:

Simple Structure Criteria:

- Each row of the factor matrix should contain at least one zero
- If there are m common factors, each column of the factor matrix should have at least m zeros
- For every pair of columns in the factor matrix, there should be several variables for which entries approach zero in the one column but not in the other
- For every pair of columns in the factor matrix, a large proportion of the variables should have entries approaching zero in both columns when there are four or more factors
- For every pair of columns in the factor matrix, there should be only a small number of variables with nonzero entries in both columns

The ideal simple structure is such that:

- Each item has a high, or meaningful, loading on one factor only and
- Each factor have high, or meaningful, loadings for only some of the items

The problem is that, trying several combinations of rotation methods along with the parameters that each one accepts (especially for oblique ones), the number of candidate matrices increases and it is very difficult to see which one better meets the above criteria.

4.6.1 Orthogonal Analytical Rotation

In the 1950s and 1960s several factor analysts using different approaches came up techniques to solve the orthogonal rotation criterion to get meaningful factors. The methods that were developed were the Quartimax criterion, Orthomax, Varimax, Parsimax and Equamax which are all mathematically equivalent. Carroll (?) was the first who proposed the quartimax criterion. He proposed that the factor matrix A would represent the simple structure if the value of

$$f_1 = \sum_{s<t}^{d} \sum_{j}^{n} A_{js}^2 A_{jt}^2 = minimum \tag{4.28}$$

This draws parallels to Thurstone's third criterion.... should be minimum. To explain the intuition behind this cost function, let us assume that the data is normalized and there are only 2 factors. All $A_{ij} \in$ [-1,1]. We would want as little overlap as possible between the contributions of 2 factors to a given data point.

Shortly afterwards, Neuhaus and Wrigley proposed that the most interpretable factor loading matrix A would be such that the variance of the nxd squared loading was a maximum.

$$f_2 = \frac{nd \sum_{i=1}^{n} \sum_{j=1}^{d} A_{ij}^4 - (\sum_{i=1}^{n} \sum_{j=1}^{d} A_{ij}^2)^2}{n^2 r^2} = maximum \tag{4.29}$$

$$f_3 = \frac{nd \sum_{i=1}^{n} \sum_{j=1}^{d} A_{ij}^4}{(\sum_{i=1}^{n} \sum_{j=1}^{d} A_{ij}^2)^2} = maximum \tag{4.30}$$

Now we know that $\sum_j \sum_i A_{ji}^2$ is always a constant because the sum of squared contributions of the all factors for a given test point is always a constant given that the factors are orthonormal to each other. Thus for a given test point Summing over all factors

$$constant = \sum_{j} \sum_{i} A_{ji}^2)^2 - \sum_{j} \sum_{i} A_{ji}^4 + 2 \sum_{j} \sum_{i<t} A_{ji}^2 A_{jt}^2 \tag{4.31}$$

Thus minimizing $\sum_{s<t}\sum_j A_{js}^2 A_{jt}^2$ is equivalent to maximizing $\sum_j \sum_i A_{ji}^4$

Varimax is quite similar to the quartimax condition. The criteria used here is

$$q = \frac{r\sum_j \sum_i (A_{ji}^2)^2 - \sum_j (\sum_i A_{ji}^2)^2}{r^2} \tag{4.32}$$

In case of orthonormal bases, the second term is a constant which disappears on differentiating, which boils down to the quartimax condition. This is called the varimax criteria because we are summing the variance of the squared loadings for all data points. The goal is to minimize the value of q, the objective function. This can be solved using an iterative method using SVD decomposition as follows

Algorithm 1 Orthogonal Analytical Rotation algorithm

procedure ROTATE
$\lambda_D = 0$
$A = B$
  while $\frac{|\lambda_D - \lambda_{Dold}|}{\lambda_D} < threshold$ **do**
    $\lambda_{Dold} = \lambda_D$
    $[L, D, M] = SVD(A' * (d * B.^3 - \gamma B diag(sum(B.^2))))$
    $T = LM'$
    $\lambda_D = sum(diag(D))$
    $B = AT$
  end while
end procedure

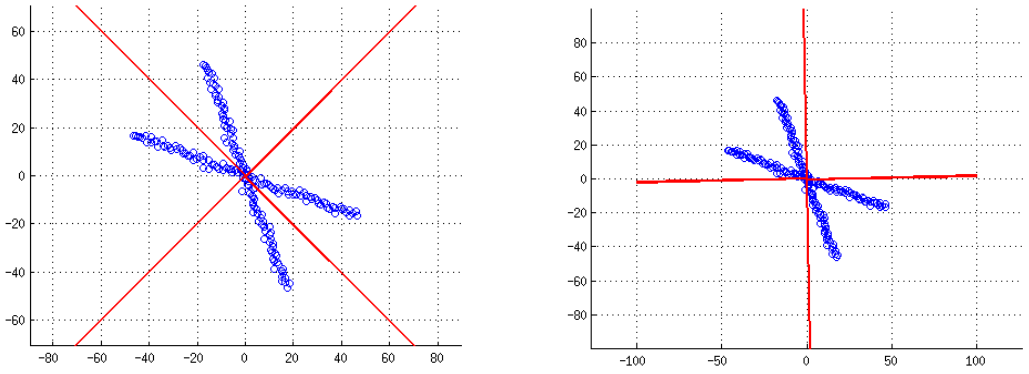


FIGURE 4.1: Comparison of PCA (left) with varimax rotated factors(Right)

4.6.2 Oblique Analytical rotation

If the restriction of orthogonality is relaxed, it is impossible to apply directly the quartimax criterion or the normal varimax criterion. This is because interfactor relationships are not considered when the criteria are in this form, and when applied all factors will

collapse into the same factor - that one factor which best meets the criterion. Thus we need a slightly more sophiticated method to get oblique factors. One of them is the promax rotation technique.This is an oblique rotation technique promax which seeks to revise varimax so that the solution can become oblique if such a construction is warranted by the data. We take the factor patterns obtained from varimax and raise it to powers of 2,4,or 6. This transformation drives down the values of all the loadings, with the smallest values from Varimax becoming much smaller with the Promax solution, while the larger loadings are not reduced as much. The result of this oblique rotation is a set of loadings that typically reflect simple structure better than do those from the Varimax solution, particularly when the latent traits are highly correlated.

Let A^{nxr} and B^{nxr} be two matrices such that $n \geq r$. Consider the class of rxr transformation matrices T in the equation

$$B = AT + E \tag{4.33}$$

B is the resultant matrix after rotation and translation, and A is the factor loading matrix that represents the data. We wish to find the transformation matrix that minimizes

$$tr(E^tE) = tr(B - AT)^t(B - AT) \tag{4.34}$$

This ensures that we have as little translation as possible. Differentiating with respect to T, we get

$$\frac{\partial tr(E^tE)}{\partial T} = -2A^tB + 2A^tAT = 0 \tag{4.35}$$

$$\Rightarrow T = (A^tA)^{-1}A^tB \tag{4.36}$$

How do we determine the resultant approximate factor matrix B? Hendickson and White[23] came up with an ingenious way of determining A and $B0$. Now $B0$ is not the true representation of the rotated factor loadings but just an initial assessment of what we would like it to be. A was taken to be the rotated factors obtained from the varimax solution as it can be assumed that the varimax solutions is close to the optimum oblique simple-structure solution. The desired features of $B0$ would be to have small valued loadings that are close to 0 approach 0 more rapidly than those distant from 0. One way to achieve this as explained above would be to take the matrix A to the m^{th} power preserving the signs. Thus $B0_{ij} = sign(A_{ij})|A_{ij}^m|$. We can thus obtain the

transformation matrix T from the equation 4.35. The final factor loading matrix B is obtained as[15]

$$B = AT \tag{4.37}$$

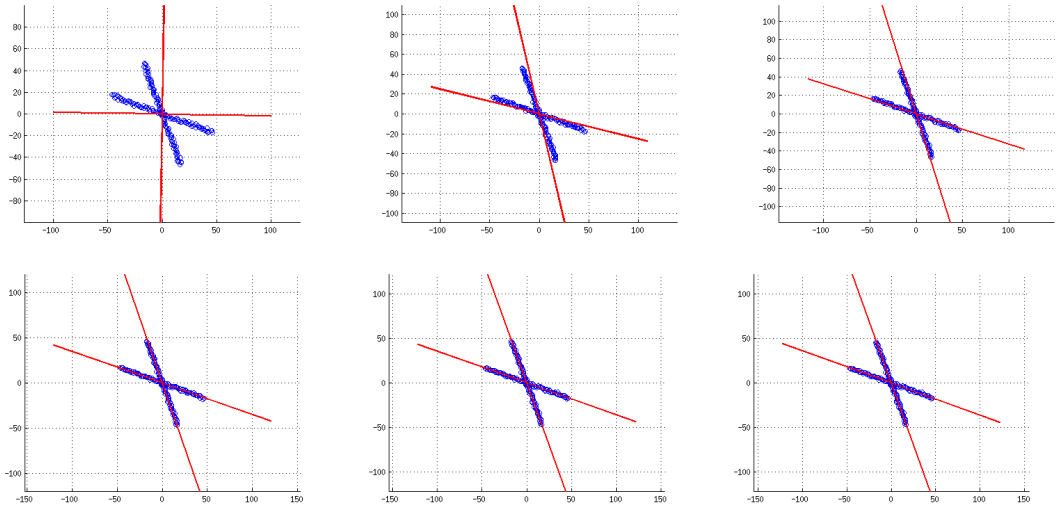


FIGURE 4.2: From Top Left to Bottom Right : Final Factors for powers from 1 to 6. When A is raised to power=1, the solution does not change from the varimax solution. The algorithm works best for a power of 4.

4.7 Dictionary Learning

A dictionary in language is a set of words that can be used to create any sentence. Similarly we can have a set of basis vectors or atoms that can be used to describe the data at hand and the process of learning these basis vectors is called dictionary learning. We can have all data points to be atoms of the dictionary but this defeats the purpose of finding patterns in data. The purpose of Dictionary Learning (DL) is to find a few atoms, a linear combination of which reconstructs the original data point with minimal error. We also impose that any data point is reconstructed from only a few atoms of the dictionary.[24][25][26]

$$x_i = \alpha_i^T D_i \qquad \text{where} |\alpha_i|_0 < K \tag{4.38}$$

The L0 norm gives us the cardinality of the vector, that is the number of non-zero terms in the vector thus making it a sparsity inducing norm. Due to a lack of mathematical representation of the L0 norm, it is an NP hard problem not suitable for optimization.

Thus for practical applications the L0 norm is relaxed to a L1(Lasso) or an L2(Ridge Regression) norm or a combination of both (Elastic net) to induce sparsity. More importantly L1 induces sparsity and L2 helps in reducing the overall magnitude of the weight vector.

DL is used in many applications such as image analysis, speech recognition where usage of predefined dictionaries such such as wavelets do not give good signal reconstruction using sparse basis. In such as scenario it is best to learn the dictionary from the data provided and adapt the dictionary as more signals are processed.

4.7.1 Problem Formulation

Consider a signal x in R^m . We say that it admits a sparse approximation over a dictionary D in R mk, with k columns referred to as atoms, when one can find a linear combination of a few atoms from D that best approximates the signal x. Let $x_i^{1xd} \forall i \in [1..N]$ be a signal,$X = [x_1 x_2 ..x_N] in R^{dxN}$ and D^{dxK} be the dictionary with K atoms. Let α_i be the sparse weight vector for a signal $x_i, \alpha = [\alpha_1 \alpha_2 ..\alpha_N]$, then the loss function $\ell(\alpha, D)$ is given by

$$\ell(\alpha, D) = \frac{1}{N} \sigma_{i=1}^{N} (\frac{1}{2} ||(x_i - D\alpha_i)||_2^2 + \lambda |\alpha_i|_1 \tag{4.39}$$

$$\tag{4.40}$$

Minimizing over the both D and α simultaneously is not a convex optimization problem but solving the problem by alternating between the two variables, minimizing over one while keeping the other fixed and vice versa makes it convex thus standard methods can be applied to find an optimal value of D and α. For more details on how to solve the above problem please refer to ...

4.8 Application of Dictionary Learning to Simple Datasets

Consider a simple dataset X^{Nxd} where $d = 2, d = 3$. The dimensions are chosen for ease of visualization. The sample points are distributed such that they lie approximately on a hyperplane $x_i \; w_1 d_1 + w_2 d_2 ... w_N d_N \forall i \in [1..n_1]$. This can be extended to another set of points $x_i \forall i \in [1..n_j]$ with a different weight vector. $n_1 + n_2 ... n_j = N$. We learn the a dictionary of atoms that best represents the data. The number of unique patterns must be known beforehand to achieve the best results. We can simply assume that the number

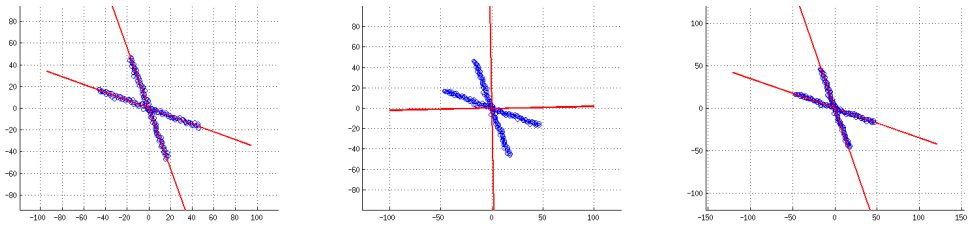


FIGURE 4.3: Data points in 2-d that indicate two distinct components. Varimax and promax rotated components for contrast.

of factors must be the degrees of freedom of the data. Since DL allows for correlated atoms, and as we know that the various unique processes that govern climate can be correlated, this is a reasonable assumption. In general, this is a hard problem and the number of basis vectors is either set by experience or coarsely evaluated empirically. A simple toy dataset in 2-d is presented below along with the learnt dictionary atoms. The points in blue are the initial dataset and the red lines indicate the atoms learnt by the above algorithm.

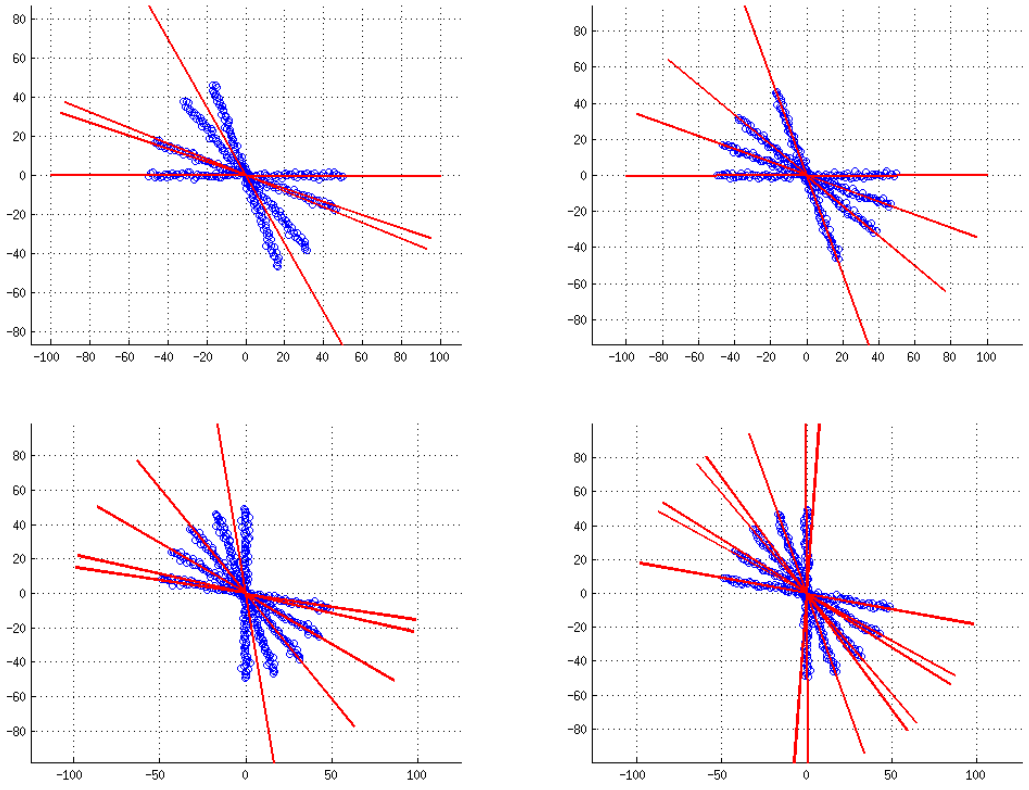


FIGURE 4.4: Top Left : DL algorithm exits at a local minimum that is not optimal; Top Right : DL find the optimal atoms; Bottom Left : DL again exiting with a non-optimal local minimum; Bottom Right : DL finds all the necessary components by overestimating

PCA \ FA	Factor 1	Factor 2	Factor 3
Factor 1	-0.9058	0.1875	0.9627
Factor 2	0.5501	0.8239	-0.0948
Factor 3	-0.0527	-0.9708	-0.0772

TABLE 4.2: Comparison of Components of PCA, Factor Analysis and Dictionary Learning over a small region

4.9 Application on climate data

Though the matrix structure looks similar for FA and PCA, the output eigen vectors are different. We express the similarity between the eigen vector components of the different methods by looking at their correlation and draw some conclusions from them. FA and PCA is applied on NCEP pressure data in a small region concentrated around the NAO index near Iceland. We take the top three non-rotated eigen vectors and make our comparisons.

PCA \ DL	Factor 1	Factor 2	Factor 3
Factor 1	0.9414	-0.1915	-0.7512
Factor 2	0.0648	-0.8788	0.7729
Factor 3	-0.8840	0.5413	0.5210

FA \ DL	Factor 1	Factor 2	Factor 3
Factor 1	0.9415	-0.6060	-0.4518
Factor 2	0.2525	0.6972	-0.8587
Factor 3	0.1721	-0.8648	0.7013

From the above results we see that the first eigen vector are correlated for all three methods indicating that the models are capturing the direction of maximum variance. This is intuitive because since we are applying the algorithm over a small region and this would result in the first component aligning itself with the dominant pattern in the region. Applying this on a large generic region gives poor relationship between the factors of different methods as shown below

PCA \ FA	Factor 1	Factor 2	Factor 3
Factor 1	-0.8842	0.3339	-0.6235
Factor 2	-0.3405	0.4613	0.7163
Factor 3	0.2486	0.7914	-0.2574

TABLE 4.3: Comparison of Components of PCA, Factor Analysis and Dictionary Learning over a big region

PCA \ DL	Factor 1	Factor 2	Factor 3
Factor 1	0.7425	0.2825	-0.5229
Factor 2	-0.5223	-0.2460	-0.8368
Factor 3	0.3917	-0.8790	0.1306

FA \ DL	Factor 1	Factor 2	Factor 3
Factor 1	-0.3857	-0.3717	0.7815
Factor 2	0.3118	-0.6887	-0.4531
Factor 3	-0.9387	-0.1371	-0.3116

PCA \ FA	Factor 1	Factor 2	Factor 3
Factor 1	-0.9349	0.1755	-0.1722
Factor 2	0.0375	0.1843	0.9672
Factor 3	0.2936	0.9403	0.0402

PCA \ DL	Factor 1	Factor 2	Factor 3
Factor 1	0.2150	-0.2221	-0.9279
Factor 2	-0.9631	0.0301	-0.2619
Factor 3	-0.0857	-0.9286	0.2388

FA \ DL	Factor 1	Factor 2	Factor 3
Factor 1	-0.2645	-0.0618	0.9283
Factor 2	-0.2270	-0.8795	0.0179
Factor 3	-0.9721	0.0115	-0.0890

The above values do not indicate the orthogonality but rather the correlation. The former is simply dot product of the eigen vectors and the latter is the dot product after subtracting the individual means. We see that the components of the three methods are aligned indicating that doing a rotation afterwards gives more meaning to the components.

4.10 Discussion and Conclusions

Rotated factor analysis and rotated PCA generally give different results as can be seen from the tables. The initial results of PCA and FA are different due to the difference in their covariance matrices. Thus the eigen vector are different. On rotation using varimax we see that the rotated vectors are close but are not exactly the same indicating that varimax converges to a local minima for the initial set of factor loadings. This leads to a very old problem of factor indeterminacy. Different methods give different results. There are various way of carrying out factor analysis such as Maximum likelihood estimation which gives a closed form solution in case of probabilistic PCA but requires an iterative methodology for FA where the unique variances for every data point is different. Iterative methods such as Newton-Raphson, Fletcher-Powell and Expectation maximization algorithm can be applied to get the maximum likelihood estimate. An issue with the Fletcher-Powell algorithm is that the covariance matrix needs to be positive definite. In case of climate data that has a high spatial-temporal auto-correlation getting a covariance matrix that is positive definite is next to impossible. In such cases we can apply the Expectation Maximization (EM) algorithm where we get the expected values of mean and variance of the factor loadings and use that to maximize the factor loadings but this generally does not yield orthonormal basis vectors (in case of PCA, the basis vectors or the eigen vectors of the correlation/covariance matrix are orthonormal). Alternatively, the EM algorithm can be modified in such a way as to yield orthonormal principal directions, sorted in descending order of the corresponding eigenvalues, directly[27]. Unfortunately, EM algorithm is sensitive to the initialization parameters and may not converge to the global optimum. Probabalistic PCA is a simplified version of the factor analysis model where it is assumed that the unique variance for each variable is the same and the maximum likelihood estimate to this gives a closed form solution which evaluates to finding the top eigen vectors of the data covariance matrix.

Historically, factor analysis has been the subject of controversy when attempts have been made to place an interpretation on the individual factors, which has proven problematic due to the nonidentifiability of factor analysis associated with rotations in the principle component space[16]. From our perspective, however, we view factor analysis as a form of latent variable density model and once this latent model is discovered we apply ay of the rotation techniques to establish some form of interpretability on our factor model. It must be noted that factor analysis models aims to establish the number for degrees of freedom of the dataset. The number can be determined by restricting the unique variance to a small value and gradually increasing the number for factors in the FA model to meet the unique variance constraint [15].

Due to the varied number of solutions present for solving the factor analysis model, which results in factor indeterminacy, it is difficult to cross-verify results. Thurstone argued that as long as interpretations could be drawn from the factor analysis model, the choice of the method was not of utmost concern[28]. But this can be always be argued otherwise and it is for this reason that PCA followed by rotation is used as it achieves the goal of interpreting the factors as stationary patterns in climate data.

Unlike principle component analysis and its variants, DL does not impose that the basis vectors be orthogonal to each other thus more effectively capturing the representation of data. We can actually draw parallels to rotation techniques such as varimax, where we would like to increase the variance of the squared loadings across factors for a variable i. The variance will be large when there are large loadings on few factors and near-zero loadings on the remaining factors. DL tries to achieve the same by "sparsifying" the weight vector α which in turn results in increasing the variance of the squared loadings.

What method to finally use in order to obtain the most interpretable factors is still up for debate. Linear methods have limited interpretability on climate data because if it being highly non-linear but over small regions they can be assumed to be almost linear and thus the above techniques can be applied. It is not advised to apply such techniques over larger regions as dominant patterns are in general correlated to each other and applying PCA with rotation reduces the interpretability of the factors. Non-orthogonal methods do not necessarily do a better job at interpreting factors either as can be seen from the table. One must exercise caution while applying the above techniques.

Chapter 5

Non Linear Techniques for Dimensionality Reduction

5.1 Introduction

Thus far we have discussed about linear methods of dimensionality reduction in order to obtain meaningful features in the dataset. Such methods have drawbacks that they ignore the topology of the dataset in higher dimensions. So on performing dimensionality reduction, the relative distances of the points are not preserved. Far away points may seem closer than nearby points in the reduced space. The distances or similarity index in most cases of non-linear dimensionality reduction is euclidean distance or the L2 norm. In climate data it is important to understand that similarity of two data points with respect to a third data point is not measured by their relative distances to the third point but rather the correlation of its values. Given the dataset, X^{nxd}, x_i is similar to x_j if they share similar behaviour in their values across the d time steps. The non-linear techniques such as ISOMAP[29][30], Locally Linear Embedding (LLE)[31][32], Laplacian EigenMaps[33][34][35] and Structure Preserving Embedding (SPE) use euclidean distances for creating a similarity matrix. But climate data requires correlation as its similarity index. In order to achieve this, we normalize the data. Normalization forces a linear relationship between squared L2 norm and correlation as follows

$$L2(x_i, x_j)^2 = ||x_i - x_j||_2^2 \tag{5.1}$$
$$= ||x_i||_2^2 + ||x_j||_2^2 - 2(x_i^t x_j) \tag{5.2}$$
$$= 1 + 1 - 2(correlation(x_i, x_j) \tag{5.3}$$
$$= 2(1 - correlation(x_i, x_j) \tag{5.4}$$
$$\tag{5.5}$$

We will use normalized z-score data for our analysis of the four Non Linear Techniques for Dimensionality Reduction (NLDR). From the analysis, we will determine if such non-linear techniques do find interesting patterns in data by taking it down to a low dimensional manifold. An interesting thought experiment to ascertain the usefulness of NLDR techniques is as follows. Consider climate data to have certain interesting patterns which are in the form of the swissroll dataset [Give hyperlink]. From our toy examples shown below, we see the NLDR techiques tend to unroll the data into a planar shape. Think of it as unrolling a mat onto the floor. If we are to visualize the data, we do not see any swissrolls but a set of points lying close to a plane preserving the local topology of the points. If no prior information is given, we would not know the original data consisted of swissrolls. Thus if the interesting pattern to be identified were swissrolls, then NLDR techniques would not be able to obtain the number as a result of the unrolling. These techniques preserve local topologies on the assumption that they are planar and thus only preserve geodesic distances.

5.2 Isomap

The Isomap algorithm is a two-step process that simultaneously attempts to find a low dimensional manifold in which a set of data points lies, and Euclidean coordinates for the points in this low-dimensional manifold. The first step in the algorithm uses a graph based approximation to the data manifold to calculate a similarity matrix based on approximate geodesic distances between data points. These geodesic distances are then analysed using multi dimensional scaling (MDS) to find an isometric embedding of the data onto a lower dimensional manifold.

5.2.1 Geodesic Approximation

In this step, we first determine the nearest neighbours of a given point. We can either have look at the K nearest neighbours of the point or looking at within an ε ball centred

at the point such that $d(x_i, x_j) = ||x_i - X_j||_2 <= \varepsilon \forall j = [1..n]$. These neighbourhood relations are captured in a graph matrix, G with the edge weights being the distances $d_G(x_i, x_j)$ in the previous step. Once the preliminary distance graph is computed, we compute the shortest path distance between any two unconnected nodes (they are not in each other's K nearest neighbours) using djikstra's algorithm or Floyd-Marshall Algorithm. With this we get the final similarity matrix G, such that $d_G(x_i, x_j)$ represents the approximate geodesic distance between any pair of points on the manifold where the true geodesic distance is given by $d_M(x_i, x_j)$ .Asymptotic convergence results exist showing that the difference between the approximation $d_G(x_i, x_j)$ and the true geodesic distance in the data manifold, $d_G(x_i, x_j)$, tends to zero in a probabilistic sense as the density of data point i.e $\lim_{n\to\infty} d_G(x_i, x_j) - d_M(x_i, x_j) = 0$ [36]

5.2.2 Dimensionality Reduction

Once the approximate geodesic distance function $d_G(x, y)$ has been found, a multidimensional scaling (MDS) procedure is applied. This procedure results in an eigenvalue spectrum that can be examined to determine the dimensionality of the data manifold. It also calculates embeddings of the data points into low-dimensional Euclidean spaces. MDS [Borg and Groenen, 1997] is a statistical technique that takes as input distance or dissimilarity measures for a set of data points and attempts to find points in Euclidean space such that the Euclidean distances between the output points correspond to the distance or dissimilarity values between the input points. Both PCA and Isomap can be considered within this framework. For PCA, the input distances are Euclidean distances in the input data, so that MDS leads to an orthogonal transformation of the data. For an idealisation of Isomap where the input distances are exact geodesic distances in the data manifold, MDS leads to an isometric transformation of the data.

5.3 Locally Linear Embedding

Suppose x_i^{1xd} is a data point sampled from a smooth manifold. If there are sufficiently many samples then we can assume that the data point and its neighbours lie on a locally linear patch. We characterize the local topology around a data point by writing it as a weighted linear combination of its neighbours. As in ISOMAP, we determine the K nearest neighbours or a set of points that like in an ϵ sized-ball. We can write x_i $\sigma_{j=1}^{k} w_{ij} x_j$. The reconstruction error will be

$$E(W) = \sigma_{i=1}^{n} |x_i - \sigma_{j=1}^{k} w_{ij} x_j|^2 \tag{5.6}$$

Following this we have two steps to compute the low dimensional isometric embedding. We first calculate the weights, and then we compute a reduced dimensionality embedding y_i^{1xp} of the point $x_i^{1xd}, p < d$

5.3.1 Compute Weights

We constrain the weights such that $\sigma_j = 1^k wij = 1$. This is to make it invariant to affine transformation of the data. Solving the weights as a constrained linear regression problem, we get the minimization error for a data point x_i as follows

$$E_i(W) = |x_i - \sigma_{j=1}^k w_{ij} x_j|^2 \tag{5.7}$$

$$= |\sigma_{j=1}^k w_{ij}(x_i - x_j)|^2 \because \sigma_{j=1}^k w_{ij} = 1 \tag{5.8}$$

$$= \begin{bmatrix} w_{i1} & w_{i2} & \dots & w_{ik} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & ckk \end{bmatrix} \begin{bmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{ik} \end{bmatrix}$$

$$= W_i^t C_i W_i \tag{5.9}$$

Here we have introduced the local covariance matrix C_i, where $c_{lm} = (x_i - x_l)(x_i - x_m)$. We introduce lagrangian λ to enforce $\sigma_j w_{ij} = 1 \forall i = [1..n]$. The loss function becomes

$$E_i(W) = W_i^t C_i W_i + \lambda(|W_i|_1 - 1) \tag{5.10}$$

Taking derivatives with respect to $w_{ij} \forall j$ and λ we get $k+1$ equations to solve for $k+1$ unknowns. Thus the weights are obtained.

5.3.2 Dimensionality Reduction

LLE now constructs a neighbourhood preserving mapping based on the weights discovered in the above step. In the final step of the algorithm, each high dimensional observation x_i is mapped to a low dimensional vector y_i representing global internal coordinates on the manifold. This is done by choosing dimensional coordinates to minimize the embedding cost function

$$L(W) = \sigma_i |y_i - \sigma_{j=1}^{k} w_{ij} y_j|^2 \tag{5.11}$$

$$L(W) = \sigma_{i,j} M_{ij} Y_i Y_j \tag{5.12}$$

Through simple calculations we get $M = (I - W)^t (I - W)$. We assume that the Y_i's are mean centered and have unit variance. Thus the Loss function can be rewritten as

$$L(W) = YMY^t \tag{5.13}$$

where,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

We now need to solve for the eigen vectors of M corresponding to the smallest non-zero eigen values and take as many as the number of dimensions onto which the data needs to be embedded.

The basic steps of this algorithm are:

1. Compute the neighbours of each data point
2. Compute the weights that best reconstruct each data point from its neighbours, minimizing the cost in eq. (1) by constrained linear fits.
3. Compute the vectors best reconstructed by the weights, minimizing the quadratic form in eq. (2) by its bottom non-zero eigen vectors.

5.4 Laplacian EigenMaps

This method is similar to LLEs where only local topologies are learnt and global topologies are ignored unlike ISOMAP where both the local and the global topologies are learnt. This method also starts with learning graph laplacian on which spectral decomposition is performed. The following steps are performed to get the eigen vectors of the laplacian.

5.4.1 Compute weights

This step is similar to LLE, where we determine edge weights W_{ij}. An edge is weighted only if x_j is in the nearest neighbour list of x_i or vice versa, else it is set to 0. The nearest neighbour can be obtained by either taking the KNN for a point or looking at an ϵ ball around the point. For the nearest neighbours we can construct W_{ij} as follows:

1. Heat Kernel : if i and j are connected, then

$$W_{ij} = \exp^{||x_i - x_j||_2^2} \tag{5.14}$$

   Otherwise, W_{ij} is set to 0.

2. We can set an adjacency matrix where the weights are $W_{ij} = 1$ if they are in the neighbourhood, else it is set to 0

5.4.2 Dimensionality Reduction

Compute the eigen vectors of the following eigen vector problem:

$$Lv = \lambda Dv \tag{5.15}$$

where D is the diagonal matrix such that $D_{ii} = \sigma_j W_{ij}$ and L is the graph laplacian where $L = D - W$. If our objective is to reduce the data to an embedding of p dimensions, then we take the bottom p eigen vectors corresponding to the smallest non-zero eigen values to be our new coordinates.

Consider the problem of embedding the points $X^{n\times d} = [x_1, x_2..x_n]$ onto a 1-d manifold $Y^{n\times 1} = [y_1, y_2....y_n]$. A reasonable loss function to make a good embedding will be

$$l(W, X) = \sigma_{ij} W_{ij} (y_i - y_j)^2 \tag{5.16}$$

This indicates that if the the points are close in d-dimensional space, then they should be close in the 1-d embedding. If they are not then they are heavily penalized because W_{ij} for such a pair of points will be large. This can be written in matrix form as L as explained 5.15.

5.5 Previous Application of NLDR techniques on Climate Data

Isomap has been applied to climate data analysis in the work of [37][38], where it was applied to observational SSTs to examine ENSO variability in the equatorial Pacific. Another recent application has been to understand the regime behaviour in atmosphere-ocean interactions during climate change[39]. [40] viewed 20th century intra-seasonal Asian monsoon dynamics using ISOMAP. ENSO dynamics in current climate models is also explored using nonlinear dimensionality reduction where they apply Isomap, one such technique, to the study of El Nino/Southern Oscillation variability in tropical Pacific sea surface temperatures, comparing observational data with simulations from a number of current coupled atmosphere-ocean general circulation models. Very little is known on LLE being applied on climate data. Laplacian Spectral Analysis is used to study the reemergence mechanisms for North Pacific Sea Ice[41]. [42] applied NLDR techniques such as Non-linear PCA, Hessian Locally Linear Embedding (HLLE) and Isomap to study pacific SSTs to understand the ENSO variabiity and to learn coupling modes in climate data.

5.6 Application on Toy Datasets

We will consider the swiss roll dataset to understand the application of the NLDR reduction techniques. We will also consider a case where NLDR techniques fail to do work as intended.

5.6.1 Swiss Roll Dataset

The dataset is as shown in the figure.

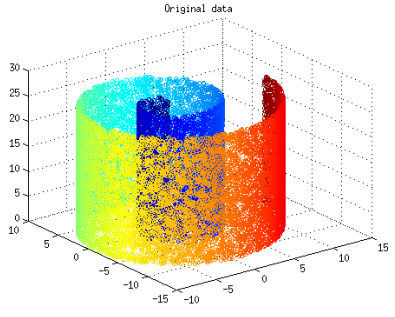


FIGURE 5.1: Data points lying in the form of a swiss roll in 3 dimensions

ISOMAP:

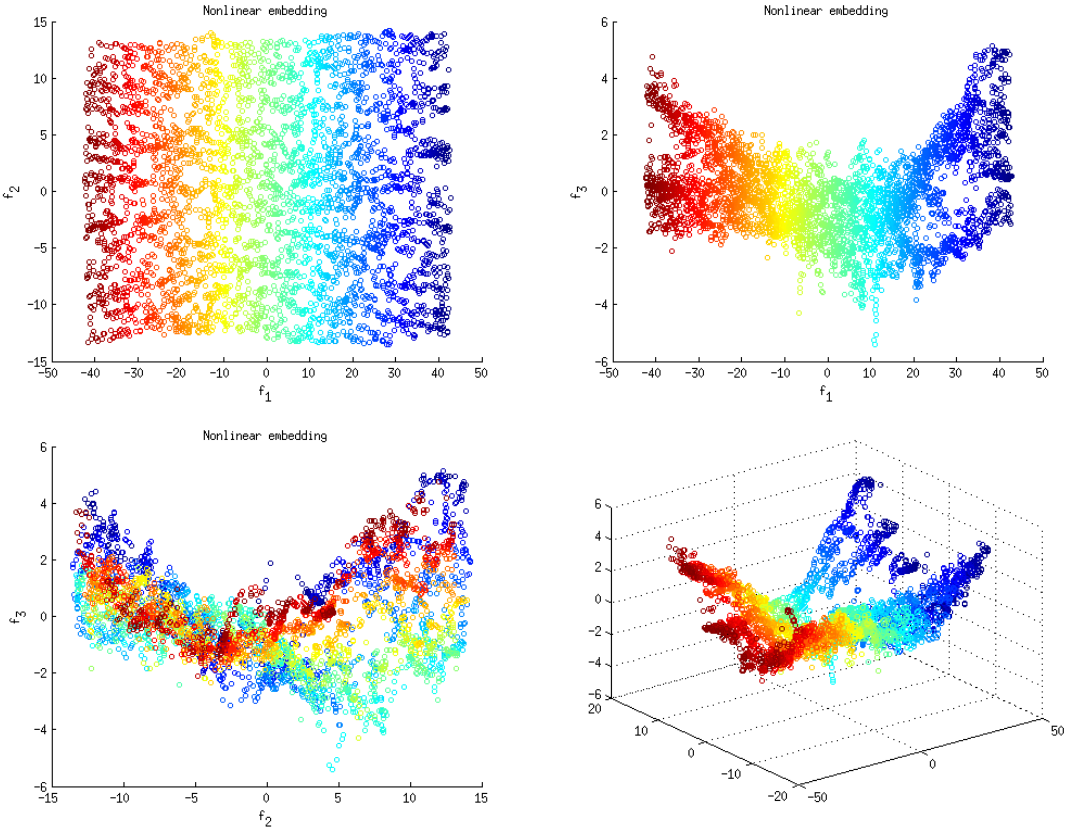


FIGURE 5.2: From Left : Spectral Decomposition along first and second Eigen vectors; first and third eigen vectors; second and third eigen vectors; 3-d embedding of swiss roll

Locally Linear Embedding:

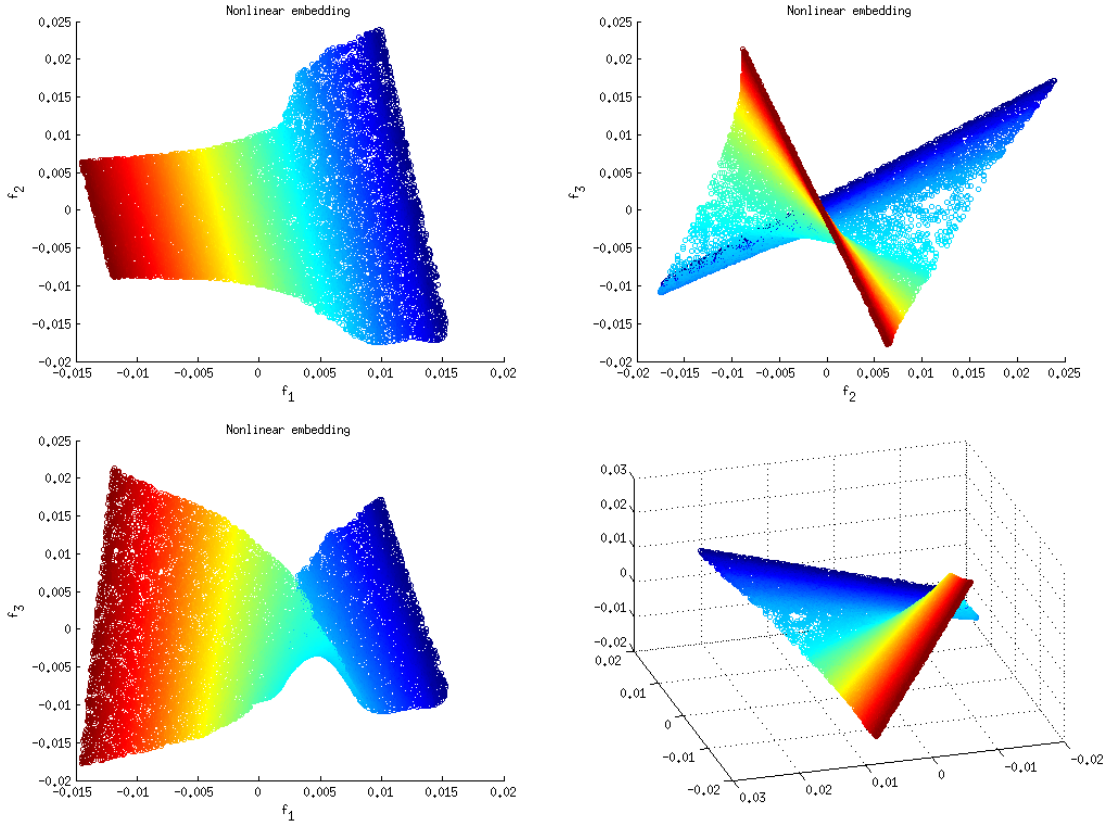


FIGURE 5.3: From Left : Spectral Decomposition along first and second Eigen vectors; first and third eigen vectors; second and third eigen vectors; 3-d embedding of swiss roll

	T(G)	T(S)	T(Spectral Decomposition)
Isomap	$O(N^2 \log N)$	$O(N^3)$	$O(min(N^3, d^3))$
LLE	$O(N^2 \log N)$	NA	$O(min(N^3, d^3))$
Laplacian	$O(N^2 \log N)$	NA	$O(min(N^3, d^3))$

TABLE 5.1: Time complexities of Isomap, LLE and Laplacian

Laplacian EigenMaps:

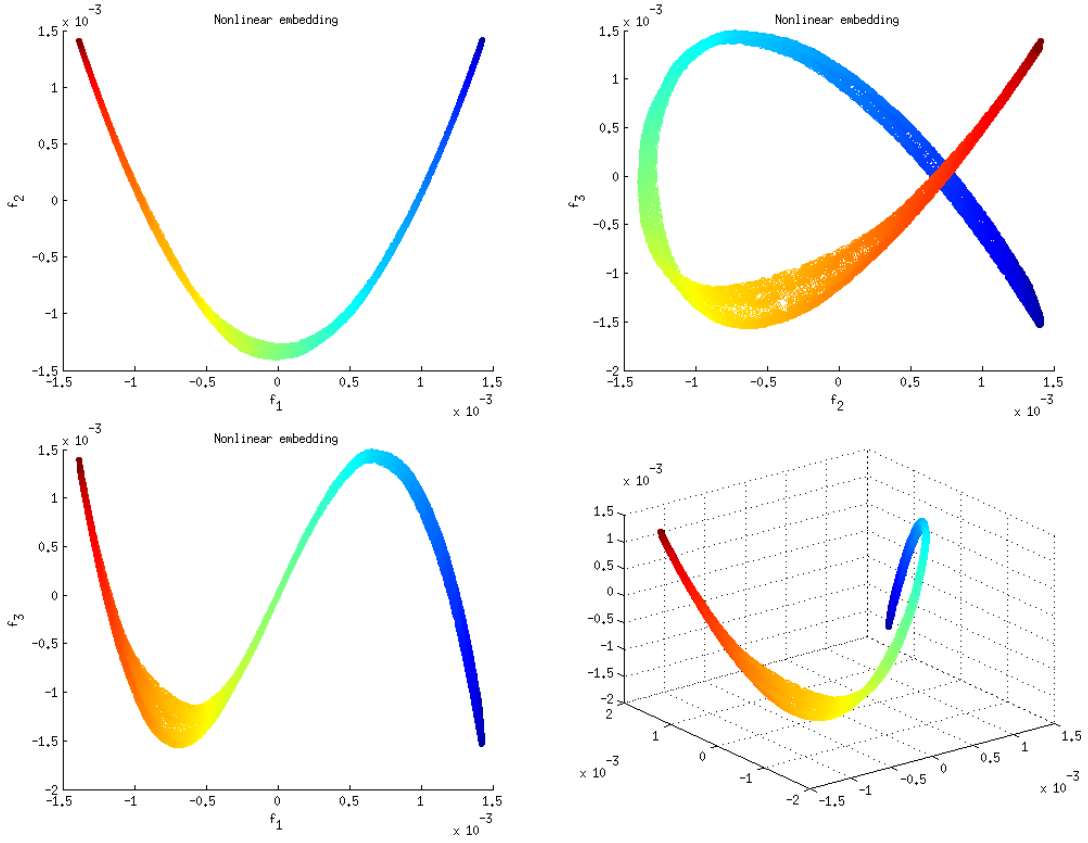


FIGURE 5.4: From Left : Spectral Decomposition along first and second Eigen vectors; first and third eigen vectors; second and third eigen vectors; 3-d embedding of swiss roll

Isomap takes the longest time to completion. Since it embeds not only local but also global topology by taking the geodesic distances, it has a larger time complexity. The net time complexity can be broken down into three steps based on the algorithm. The initial construction of the graph G, the computation of the spectral graph S (this is the L for a laplacian, M for LLE and G for isomap which includes geodesic distances as well), and spectral decomposition of S to get eigen vectors.

5.6.2 Gaussian Dataset

In this dataset, the 3-d manifold looks like a gaussian distribution. There are such distributions of which one is inverted. Applying the NLDR techniques illustrated below Gaussian datasets perform poorly when NLDR techniques are applied on them. From

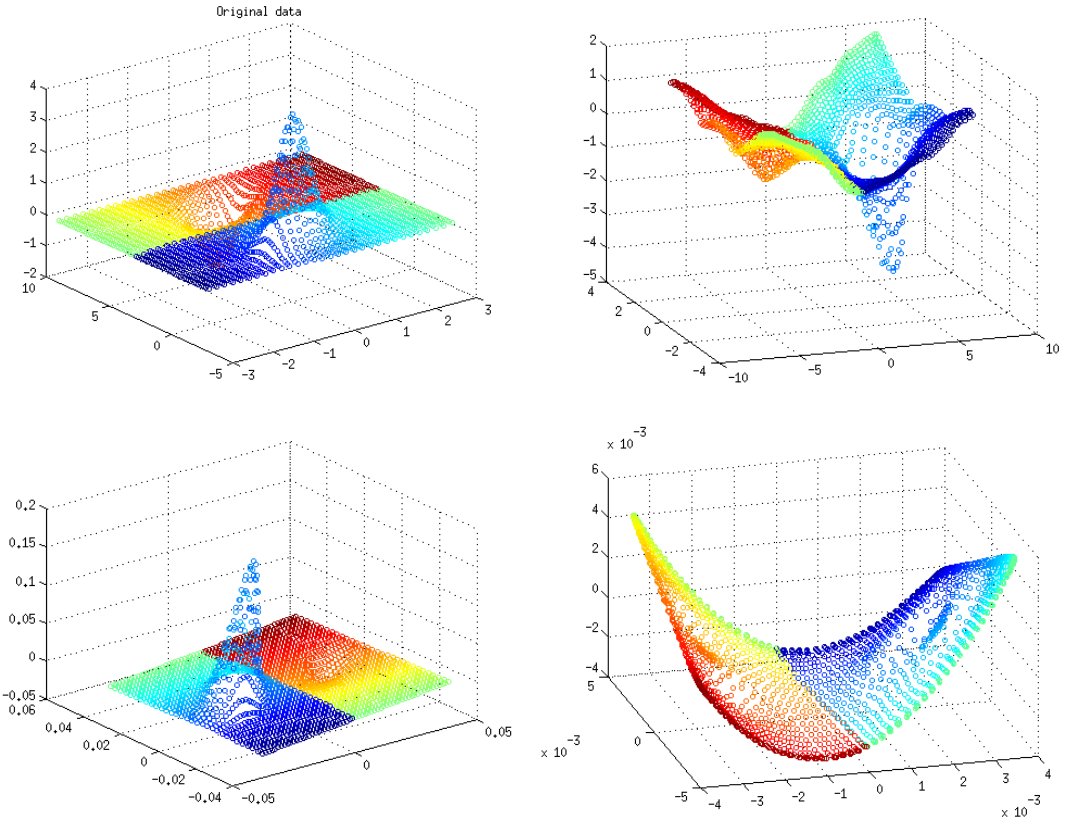


FIGURE 5.5: From Left : Original data with 2 gaussian distributions; 3d embedding from ISOMAP; 3d embedding from LLE; 3d embedding from Laplacian

the figures, we can observe that LLE preserved the topology, whereas Laplacian do not. In case of isomap, the two gaussians now face the same direction. This can give false indications that the two gaussians are positively related when they are not.

5.7 Application on climate Data

The above examples indicate that NLDR techniques give poor results on complex manifolds and one must exercise caution while applying them. For example one limitation of Laplacian eigenMaps is that the dimensionality of the graph Laplacian scales with the number of data points. Therefore, the associated eigenvalue problem can become intractable for large data sets. This problem can be overcome by a reduction of the effective number of data points or by taking a smaller spatially contiguous region of points.

For our experiments we take the region around North Atlantic Oscillation(NAO) and Southern Oscillation(SO) to further understand these oscillations. In building the nearest neighbour graph, the spatial neighbors are considered i.e points to the north, south, east and west of the current index and not the temporal neighbours based on temporal autocorrelation. This is because the points are not sampled uniformly in the spatial dimension and by taking such neighbours we are enforcing spatial connectivity across latitudes and longitude of a given point. One must note that a location that is spatially close is temporally close to i.e the correlation of the timeseries will be close to 1. We take $K = 4$ considering off-diagonal neighbours and $k = 8$ considering the diagonal neighbours as well. This is a common step in all manifold learning algorithms.

NLDR works best when applied on low dimensional data. For our experiments we will first reduce the dimensionality of the original data using PCA and we retain 98%(say) of the original variance. Since PCA is a correlation preserving embedding, it is simply a rotation of the original axes to the principle axes as determined by the eigen vectors.

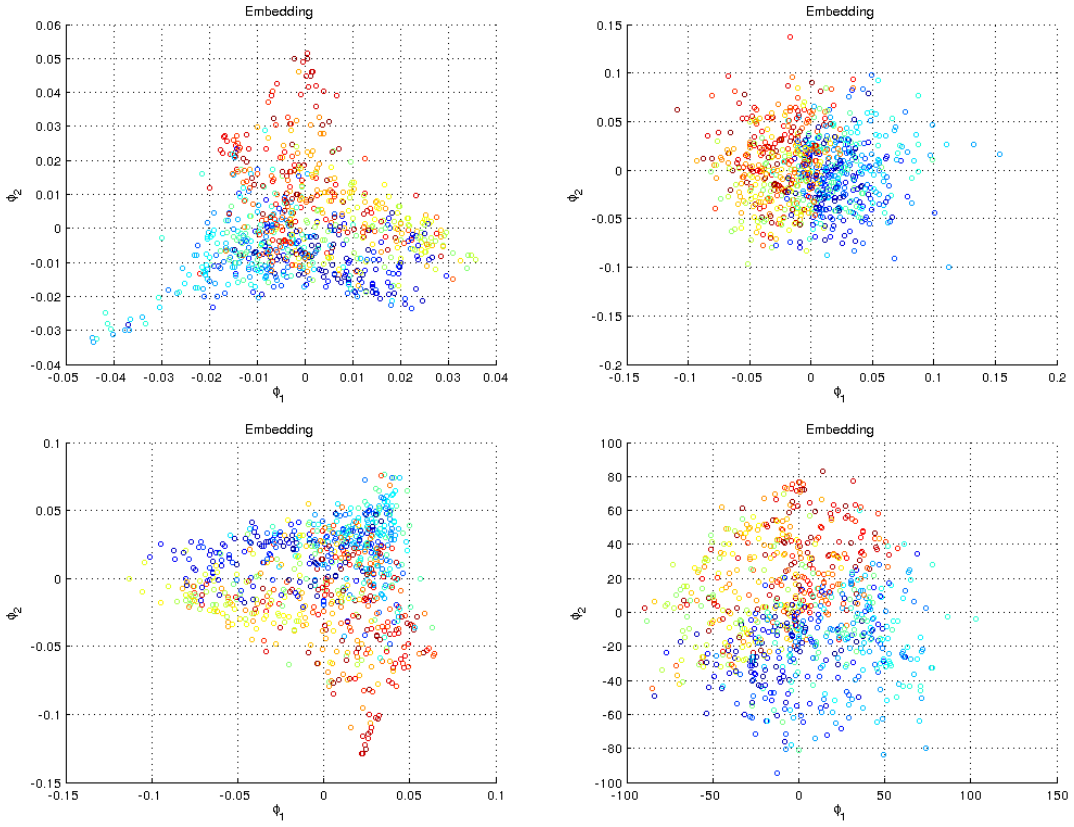


FIGURE 5.6: Top Left :Laplacian, Top Right : ISOMAP, Bottom Left : LLE, Bottom Right : PCA with varimax. Blue points indicate the la-nina phase and red points indicate the el-nino phase of Southern Oscillation

Comparison of the eigen vectors of linear and non-linear techniques do not make sense because the distribution of data is different due to the process of unfolding when applying NLDR techniques.

The plots indicate the distribution of points along the leading two principle components for the NLDR techniques and compared with rotated PCA component. The points in blue indicate the la-nina phase and red points indicate the el-nino phase of Southern Oscillation. The Isomap and PCA show pretty similar results and a good separation. LLE and Laplacian perform poorly and do not separate the 2 phases. Separation of these two phases is actually a trivial problem. Positive anomalies of the southern oscillation indicates the el-nino and the negative anomalies indicate the la-nina.

5.8 Discussions and Conclusions

NLDR techniques in general try to disentangle a tangled manifold. For example take a piece of paper and crunch it into a ball. if we sample enough points from this 3-d ball and the neighbours of any point lie along the plane of the paper, then applying any of the above techniques would get us back the isometric projection of the 3-d ball which is original flat paper that we started out with. Let us take another example of a mexican hat in the form of a gaussian distribution lying in 3-d space. Any amount of force applied at the brim of the hat in a radial direction is not going to deform the object (Our objective is not to tear it but simply deform it). Application of any NLDR technique is not going to give us a flatter manifold[43]. The distance preserving isometric projection of the mexican hat does not exist and the application of the NLDR techniques distorts the information doing more harm than good. Thuss if our manifold had important patterns in this case it is required to identify 2 gaussian distributions lying on the manifold, then this information will be lost. Let us take another example of a spherical distribution of points. An isometric projection for sphere is not possible[]. The 2-d embeddings of all techniques is shown here and as we can see though the local distances are being preserved. The global geodesic distances cannot be.

Notice that LLE performs poorly for a full sphere. The NLDR methods do not work well on a half sphere as well. But gives a reasonable output for LLE unlike its result for a full sphere.

From the linear methods of dimensionality reduction discussed above it is natural to seek a trade-off between the two goals of statistical fidelity(explaining most of the variance in the data) and interpretability (making sure that the factors involve only a few coordinate axes). Solutions that have only a few nonzero coefficients in the principal components

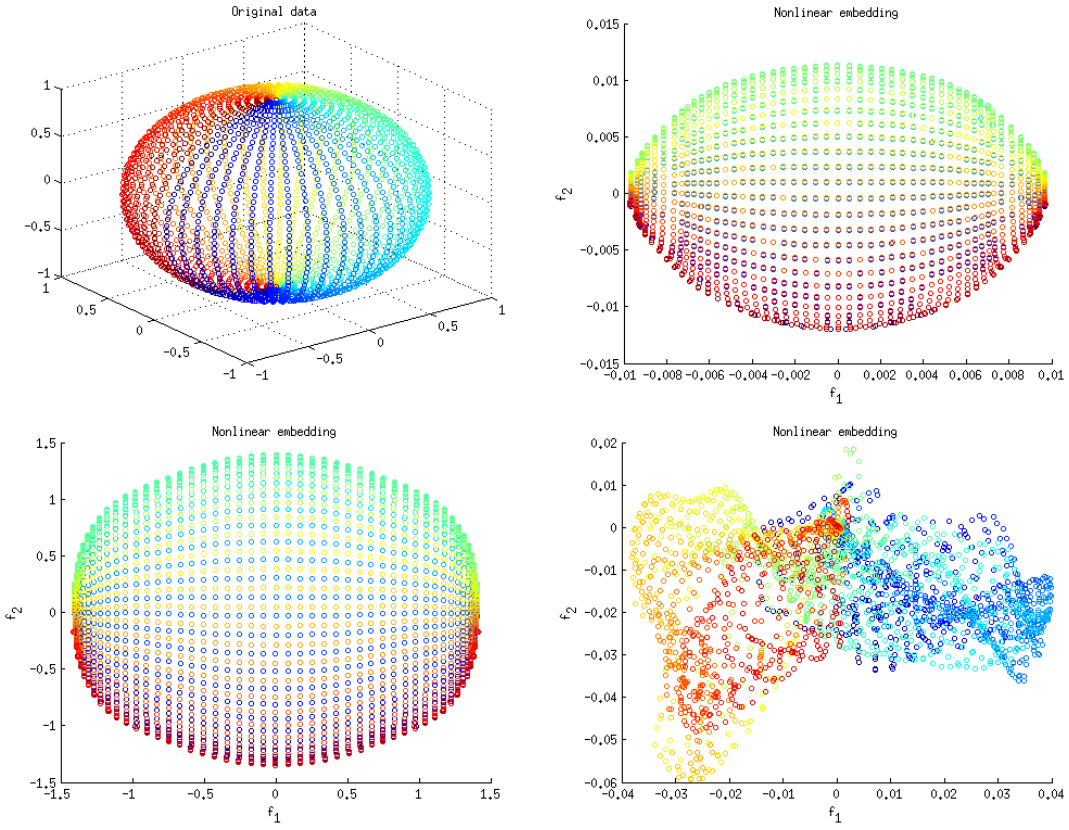


FIGURE 5.7: Top Left : Original data with spherical distribution of points;Top Right: 2d laplacian embedding. Bottom Left to Right : 2-d projection ISOMap, LLE

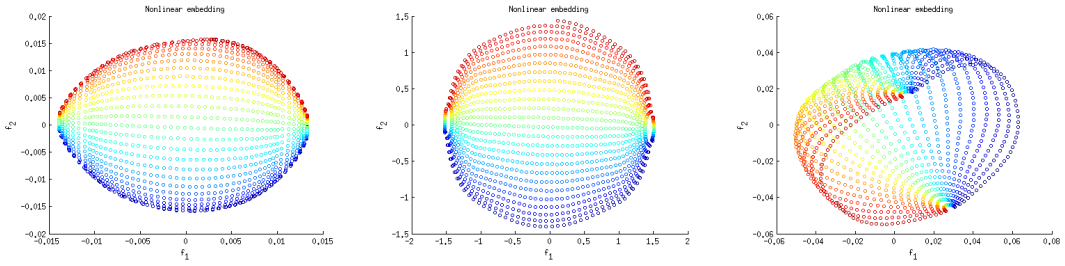


FIGURE 5.8: Left to Right : 2d laplacian embedding, 2-d projection ISOMap, 2-d projection LLE

are usually easier to interpret. When studying climate data, patterns are what we would like obtain and interpretibility is given higher importance. Interpretibility is a hard thing to do if the data is lying in some high dimensional manifold that is entangled. Again consider the above example of a paper mexican hat lying attached to a plane of paper. If this were wrapped up into a ball, then the mexican hat pattern is lost to us unless we perform any of the above NLDR techniques. Direct application of PCA on this would not give us anything useful to interpret.

For best results it would be best to apply PCA or a variation of PCA called the Robust PCA that is robust to outliers. We capture a percentage of variance i.e say around 95

to 99 percent. PCA is a topology preserving application because it rotates the original axes to principle component axes. After capturing a high percentage of variance (call it the topology smoothener), we apply NLDR techniques on this reduced space to untangle the embedding if present. This would open up many intricate small patterns that would not have been discovered by doing simple PCA/rotation. An important point to note here is that temporal information is lost in the above process but we have preserved the relationships of points based on temporal information only, thus for any two data points on the manifold, x_i and x_j are close only of they have a high correlation i.e are similar in their behaviour over all time steps.

Can we have a correlation and distance preserving dimensional reduction? My experiments have indicated that it seems unlikely. We can combine the two together to create a loss function that needs to be optimized in the following way (4.5).

$$l(C, W, X) = \sigma_{ij} W_{ij} (y_i - y_j)^2 - \lambda \sigma_{ij} C_{ij} y_i^t y_j \tag{5.17}$$

The second term is the correlation embedding term where if two points are negatively correlated, then their corresponding embeddings also need to be negatively correlated else it is penalized by bring the net sum lower. Thus the second terms needs to be maximized or the negative of it needs to be minimized. $\lambda \in [0, 1]$ is a tuning parameter.

Application on climate data does not seem to yield any patterns of interest when looking at a smaller region. This is perhaps due to the fact that climate data is smooth and linear on a smaller regions. It must be noted that most applications of NLDR techniques are use for clustering or for semi-supervised learning[44]. If patterns exist in a non-linear structure which are not separable then it may be possible to cluster them after applying any of the above techniques. Generally climate data does not come with labels thus making it an unsupervised learning problem where stationary patterns i.e teleconnections are learnt. The mechanics of the patter/teleconnection or rather the underlying structure of the pattern can be learnt.

Bibliography

[1] G.T. Walker. Correlation in seasonal variations of weather, ix. a preliminary study of world weather. *Memoirs of the India Meteorological Department*, 24:275–332, 1924.

[2] CL Pekeris. Atmospheric oscillations. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 158(895):650–671, 1937.

[3] S. Gadgil, PN Vinayachandran, PA Francis, and S. Gadgil. Extremes of the indian summer monsoon rainfall, enso and equatorial indian ocean oscillation. *Geophysical Research Letters*, 31(12):L12213–1, 2004.

[4] H.A. Bridgman and J.E. Oliver. *The global climate system: patterns, processes, and teleconnections*. Cambridge Univ Pr, 2006.

[5] R. Kistler, E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, et al. The ncep-ncar 50-year reanalysis: Monthly means cd-rom and documentation. *Bulletin-American Meteorological Society*, 82(2):247–268, 2001.

[6] G.T. Walker and MA CSL. Correlation in seasonal variations of weather, viii. 1910.

[7] D.S. Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

[8] D. Dommenget and M. Latif. A cautionary note on the interpretation of eofs. *Journal of Climate*, 15(2):216–225, 2002.

[9] HM Van den Dool, S. Saha, and Å. Johansson. Empirical orthogonal teleconnections. *Journal of Climate*, 13(8):1421–1435, 2000.

[10] J. Kawale, S. Liess, A. Kumar, M. Steinbach, A. Ganguly, N.F. Samatova, F. Semazzi, P. Snyder, and V. Kumar. Data guided discovery of dynamic climate dipoles. 2011.

[11] S.M. Uppala, PW Kållberg, AJ Simmons, U. Andrae, V. Bechtold, M. Fiorino, JK Gibson, J. Haseler, A. Hernandez, GA Kelly, et al. The era-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012, 2005.

[12] K. Onogi, H. Koide, M. Sakamoto, S. Kobayashi, J. Tsutsui, H. Hatsushika, T. Matsumoto, N. Yamazaki, H. Kamahori, K. Takahashi, et al. Jra-25: Japanese 25-year re-analysis projectprogress and status. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3259–3268, 2005.

[13] http://www.esrl.noaa.gov/psd/data/.

[14] Jaya Kawale, Stefan Liess, Arjun Kumar, Michael Steinbach, Peter Snyder, Vipin Kumar, Auroop R Ganguly, Nagiza F Samatova, and Fredrick Semazzi. A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(3):158–179, 2013.

[15] Stanley A Mulaik. *Foundations of factor analysis*. CRC press, 2009.

[16] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[17] Jack O Neuhaus and Charles Wrigley. The quartimax method. *British Journal of Statistical Psychology*, 7(2):81–91, 1954.

[18] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[19] Herbert W Eber. Toward oblique simple structure: Maxplane. *Multivariate Behavioral Research*, 1(1):112–125, 1966.

[20] Frank W Warburton. Analytic methods of factor rotation. *British Journal of Statistical Psychology*, 16(2):165–174, 1963.

[21] John B Carroll. Biquartimin criterion for rotation to oblique simple structure in factor analysis. *Science*, 1957.

[22] Robert I Jennrich and PF Sampson. Rotation for simple loadings. *Psychometrika*, 31(3):313–323, 1966.

[23] Alan E Hendrickson and Paul Owen White. Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1):65–70, 1964.

[24] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[25] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[26] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

[27] JeeWon Ahn and Ji-Young Oh. A constrained em algorithm for principal component analysis. *Neural Computation*, 15(1):57–65, 2003.

[28] Louis Leon Thurstone. Multiple factor analysis. 1947.

[29] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16:177–184, 2004.

[30] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[31] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[32] Dick De Ridder and Robert PW Duin. Locally linear embedding for classification. *Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology, Delft, The Netherlands, Tech. Rep. PH-2002-01*, pages 1–12, 2002.

[33] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[34] X Niyogi. Locality preserving projections. In *Neural information processing systems*, volume 16, page 153. MIT, 2004.

[35] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[36] Basil B Bernstein. *Pedagogy, symbolic control, and identity: Theory, research, critique*. Number 4. Rowman & Littlefield, 2000.

[37] Dimitrios Giannakis and Andrew J Majda. Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear laplacian spectral analysis. *Geophysical Research Letters*, 39(10), 2012.

[38] Arellano J Gámez, CS Zhou, Axel Timmermann, and Jürgen Kurths. Nonlinear dimensionality reduction in climate data. *Nonlinear Processes in Geophysics*, 11 (3):393–398, 2004.

[39] Bijan Fallah and Sahar Sodoudi. Bimodality and regime behavior in atmosphere–ocean interactions during the recent climate change. *Dynamics of Atmospheres and Oceans*, 70:1–11, 2015.

[40] Abdelwaheb Hannachi and AG Turner. Isomap nonlinear dimensionality reduction and bimodality of asian monsoon convection. *Geophysical Research Letters*, 40(8): 1653–1658, 2013.

[41] Mitchell Bushuk, Dimitrios Giannakis, and Andrew J Majda. Reemergence mechanisms for north pacific sea ice revealed through nonlinear laplacian spectral analysis*. *Journal of Climate*, 27(16):6265–6287, 2014.

[42] Ian Ross. Nonlinear dimensionality reduction methods in climate data analysis. *arXiv preprint arXiv:0901.0537*, 2009.

[43] PL Robinson. The sphere is not flat. *The American Mathematical Monthly*, pages 171–173, 2006.

[44] Henning Sprekeler. On the relation of slow feature analysis and laplacian eigenmaps. *Neural computation*, 23(12):3287–3302, 2011.