Empirical Orthogonal Functions and their Applications to Climate Data

Kevin Leung
New York University
Kevin.Leung@stern.nyu.edu
Dynamics of Earth's Atmosphere and Climate

1. Introduction and Motivation

Earth's climate is a highly complex, nonlinear dynamic process with high dimensionality. Due to the complexity of climate systems, methods to reduce the dimensionality of obtained climate data, with hopes of finding the most important climate patterns, are extremely valuable to researchers. One popular dimensionality-reduction technique is the method of empirical orthogonal functions (EOFs), which leverages linear algebra and matrix properties to reduce the number of variables in the data while preserving (most of the) explained variance [1].

Originally, EOFs were used to reduce the large number of variables in climate data for computational and memory purposes, however, researchers have recently found that individual "modes of variability" (in other words, vectors in linear algebra terminology) extracted from EOF analysis on spatio-temporal climate data are consistent with physical climate phenomena [2]. One notable pattern that has been obtained from using EOFs on climate data is the Arctic Oscillation.

2. Climate Data

Before we can go into the mathematical details of EOFs, we must first understand the structure of climate data that EOFs can be applied to. EOFs are the method of choice for analyzing the variability and reducing the dimension of climate data of a single field (or scalar measurement) F, such as sea level pressure or sea-surface temperature. Climate data of a single field F generally comes as a 3-dimensional array, with the dimensions time (t), latitude (θ), and longitude (ϕ) [1], [3]. To be precise, the dimensions are discretized to yield the coordinates: time $t_i,\ \ i = \{1, \ldots, m\}$; latitude $\theta_j,\ j = \{1, \ldots, n\}$; and longitude $\phi_k,\ k = \{1, \ldots, p\}$. We can now denote our data matrix of a single climate field F as a function in 3-D space:

$$F_{ijk} = F(t_i, \theta_j, \phi_k), \text{ where } 1 \leq i \leq m,\ 1 \leq j \leq n,\ 1 \leq k \leq p \tag{1}$$

However, the method of EOFs is applied to 2-D matrix data so we must reshape our 3-dimensional array into two dimensions. This is also advantageous computationally, as working with a 2-D array is easier than processing the higher-dimensional, original array F.

We can convert F into two dimensions by concatenating the spatial coordinates of latitude and longitude. Our new data matrix X now has coordinates in terms of a time-dimension t and a spatial-dimension $s = \theta \cdot \phi$, so $X \in \mathbb{R}^{m \times (n \cdot p)}$.

With the reshaped array X, we now compute the time average of the field for each spatial coordinate s_j in our data matrix. We let $\bar{x}_{:,j}$ be the average-over-time of the field at the j^{th} column of our data:

$$\bar{x}_{:,j} = \frac{1}{n} \sum_{i=1}^{n} x_{i,j} \tag{2}$$

We then compute $\bar{x}_{:,j}$ for each location j along the spatial dimension s_j to get the column vector $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}} = [\bar{x}_{:,1}, \ldots, \bar{x}_{:,(n \cdot p)}] \tag{3}$$

We now subtract $\bar{\mathbf{x}}^T$ from each row of X to get a centered data matrix $\tilde{X}$, where each spatial location now has mean 0 over the time dimension. Mathematically, we can denote this as:

$$\tilde{X} = X - \mathbf{1}\bar{\mathbf{x}}^T \tag{4}$$

where $\mathbf{1} = (1, \ldots, 1)$, which is a column vector with size m consisting of all 1's.

Although our data matrix $\tilde{X}$ is both 2-dimensional and has mean 0 along the time dimension for each spatial coordinate, the observed data still exhibits a troubling property: due to the shape of the Earth, climate data is (generally) non-uniformly distributed over the spatial dimensions. For example, if the collected data is provided on a grid of $10°\text{lat} \times 10°\text{lon}$ resolution, then the distribution of data will be denser poleward. Non-uniformity in the data along the spatial dimension can potentially influence and bias the EOFs.

In order to mitigate potential issues from the geometric non-uniformity of the Earth, we can normalize each data point by the local area of its location. To accomplish this, we can weigh each point in our observed climate data using the cosine of its latitude. We denote θ_k as the latitude of the k^{th} grid point, $k = 1, \ldots, (n \cdot p)$ and D_θ as the diagonal matrix:

$$D_\theta = \begin{bmatrix} \cos(\theta_1) & 0 & \ldots & 0 \\ 0 & \cos(\theta_2) & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & \ldots & \cos(\theta_{n \cdot p}) \end{bmatrix} \tag{5}$$

So now we can weight our centered matrix $\tilde{X}$ using D_θ:

$$\mathbf{X} = \tilde{X} D_\theta \tag{6}$$

We can now apply the method of EOFs on the spatially-weighted, centered (across time), 2-dimensional climate data $\mathbf{X}^T$.

3. Singular Value Decomposition and Empirical Orthogonal Functions

The singular value decomposition is a fundamental property of any $m \times n$ matrix A of real numbers, $A \in \mathbb{R}^{m \times n}$, which forms the basis of EOFs. Reiterating the theorem from [4]:

Theorem 3.1. *Every rank r matrix $A \in \mathbb{R}^{m\times n}$, where $m \geq n$, has a singular-value decomposition of the form*

$$A = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \dots & \vec{u}_r & \vec{u}_{r+1} & \dots & \vec{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & & 0 & \dots & 0 \\ & & & & \ddots & & & \\ 0 & 0 & \dots & \sigma_r & & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & & 0 & \dots & 0 \\ & & & & \ddots & & & \\ 0 & 0 & \dots & 0 & & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \\ \vec{v}_{r+1}^T \\ \vdots \\ \vec{v}_n^T \end{bmatrix} = USV^T$$

where the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are positive real numbers, the left singular vectors $\vec{u}_1, \vec{u}_2, \dots, \vec{u}_m$ form an orthonormal set in $\mathbb{R}^m$, and the right singular vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ also form an orthonormal set in $\mathbb{R}^n$. $U \in \mathbb{R}^{m\times m}$, $S \in \mathbb{R}^{m\times n}$, and $V \in \mathbb{R}^{n\times n}$.

When taking the SVD of $\mathbf{X}^T \in \mathbb{R}^{(n\cdot p)\times m}$, the EOFs are the column vectors of U while the column vectors of V are known as principal components (PCs). One can think of the method of EOFs as projecting the centered, spatially-weighted climate data onto an orthonormal basis. Additionally, EOFs, which are in $\mathbb{R}^{n\cdot p}$ and have the same size as the spatial dimension of our transformed climate data, are stationary patterns that do not evolve with time. However, an EOF's corresponding principal component, which is in $\mathbb{R}^m$ and has the same size as the time dimension of the data, provides information about the sign and overall amplitude of an EOF as a function of time. Although EOFs are constant vectors with respect to time, they do change sign and amplitude in order to represent the changing state of the atmosphere.

The amount of variance in the original data that is explained by the k^{th} EOF (and PC) can be derived from the singular values $\sigma_1, \dots, \sigma_r$:

$$\text{Percent of Variance Explained by } k^{\text{th}} \text{ EOF} = \frac{\sigma_k}{\sum_{i=1}^{r} \sigma_i} \tag{7}$$

And since $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$, the amount of variance explained by the set of EOFs from U is in decreasing order; in other words, the first EOF explains the most variance in the climate data (among all the EOFs), the second EOF explains the second most, etc.

After applying the SVD on climate data of a single field, you now have a few choices on what to do with the obtained EOFs. Originally, EOFs were used to reduce the dimensions of high-dimensional climate data while retaining most of the explained variance in the data. One can accomplish this by using equation (7) and truncating the data after the first k EOFs, such that the amount of variance explained by the k EOFs that are kept is greater than a predetermined threshold (typically, people use 80%, 90%, 95%, or even 99% of the original variance):

$$\text{Retain } k \text{ EOFs such that: } \frac{\sum_{j=1}^{k} \sigma_j}{\sum_{i=1}^{r} \sigma_i} \geq 80\% \tag{8}$$

The k EOFs that are retained now serve as an *approximation* of the original climate data but with the added advantage of being lower dimensional.

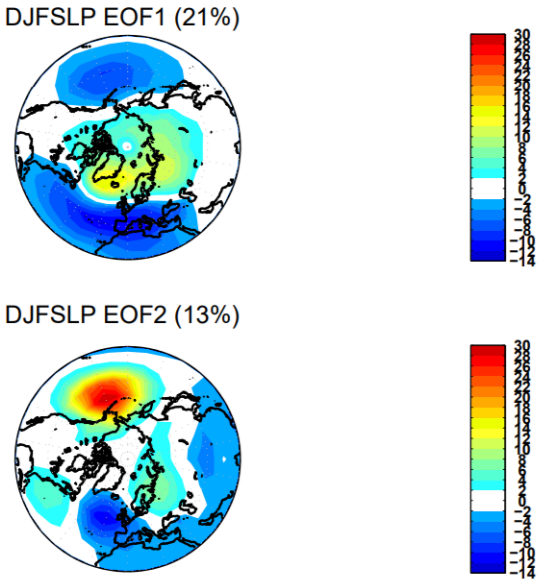

Figure 1: First 2 EOFs on SLP data from Hannachi.

Along with using EOFs as an approximation of your data, you can also analyze the individual EOF vectors to search for "modes of variability" that correspond to physical phenomena. In the next section, I explain two applications of EOF analysis on sea level pressure (SLP) data to help illustrate this use-case.

4. Application of Empirical Orthogonal Functions

4.1. Abdel Hannachi's (University of Reading) EOF Analysis

In Hannachi's primer on EOF Analysis [1], he applies EOFs to winter monthly (December to February) SLP data over the Northern Hemisphere. The data is taken from 1948 to 2000, with a horizontal grid of $2.5° \times 2.5°$ resolution. Hannachi follows the preprocessing procedure outlined in section 2: he first centers the data by removing monthly averaged SLP values and then weights the data by the cosine of the corresponding latitude.

In his analysis, Hannachi finds that the first 2 EOFs explain 21% and 13% of the total variance, respectively. When overlayed over a map, the first 2 EOFs exhibit physical patterns as depicted in Figure 1. The first EOF, which shows a high over the North Pole and low centers over the Mediterranean-North East Atlantic and the North Pacific, is consistent with the Arctic Oscillation. The Arctic Oscillation is a ringlike-pattern of SLP anomalies which is centered at the poles; climatologists believe the Arctic Oscillation is causally related to global weather patterns. The second EOF is less informative and illustrates a high center over the Northern Pacific with a low center over the North East Atlantic.

4.2. My Attempt at EOF Analysis

For my application of EOF analysis, I worked with daily-average SLP data taken from 2007 to 2017 on $3° \times 3°$ resolution. To be consistent with Hannachi, I only use the data from

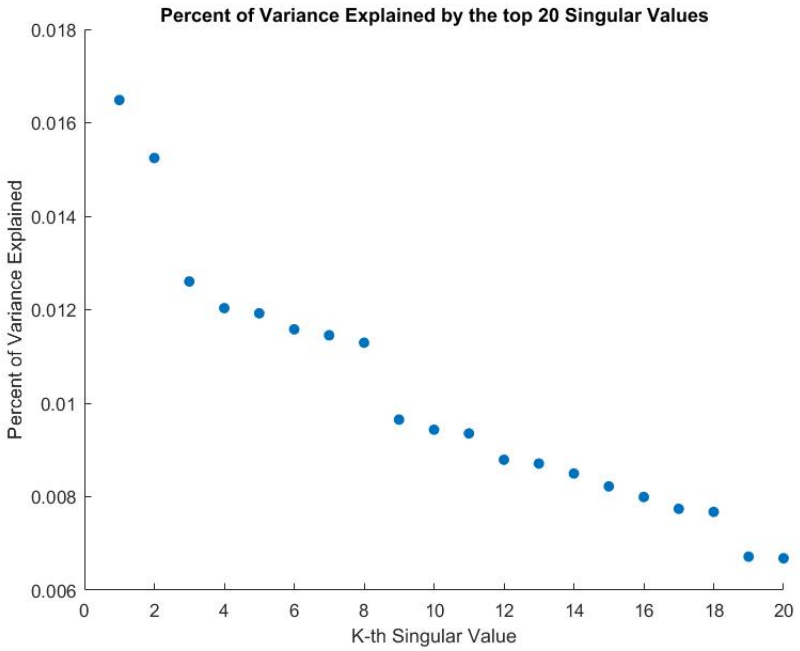

Figure 2: Plot of the percent of variance explained by the first 20 EOFs using the obtained singular value spectrum.

the Northern Hemisphere. I reshaped the array into 2-dimensions, subtracted the average-over-time for each spatial location, and then normalized each location by the cosine of its latitude. I then applied the SVD on the SLP data and extracted the first 2 EOFs.

Unlike the results from Hannachi, the first 2 EOFs from my daily-averaged SLP data explain much less variance: 1.65% and 1.52% respectively as we see in Figure 2. This indicates that the first 2 EOFs (and all subsequent EOFs) do not explain much of the data's variance by themselves, which immediately indicates that the EOFs may not exhibit any useful patterns. This claim is supported in Figures 3 and 4, which show the magnitude of the EOF values when flattened onto a horizontal latitude-longitude grid. We immediately see that the EOFs look like noise and are basically random. We do make note that the values of EOF 2 look, in general, to be greater than the corresponding values of EOF 1.

Because of the noisy EOF values, we cannot make any physical interpretations about the extracted EOFs. The insight that we can generate, however, is that using data taken on a shorter time scale (days instead of months) can lead to noisy EOF analysis. Intuitively, this makes sense because the first EOFs represent patterns that explain the most variance in the data; in general, these patterns would be larger-scale trends in the climate. However, if one uses daily data, there is a lot more noise and less of a pattern between data points taken in time than if one took data measured on a larger time frame. In other words, there are more obvious and profound climate patterns, such as seasons, on a monthly timescale than there are on a daily timescale (i.e. the weather between 2 consecutive days is going to be pretty similar on average).

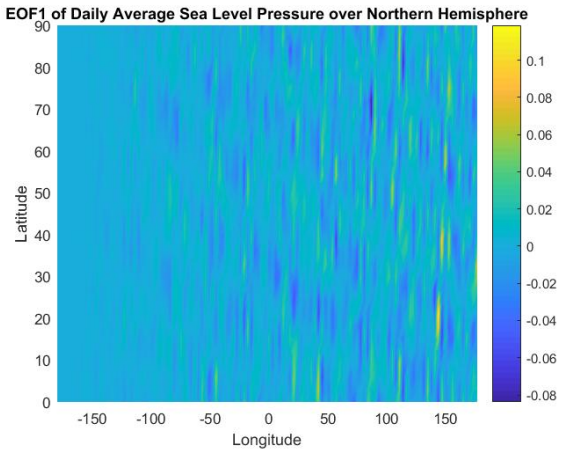

Figure 3: EOF 1 from my application of the EOF method on SLP data. We see no obvious, informative physical patterns, since the generated EOF values look very noisy.

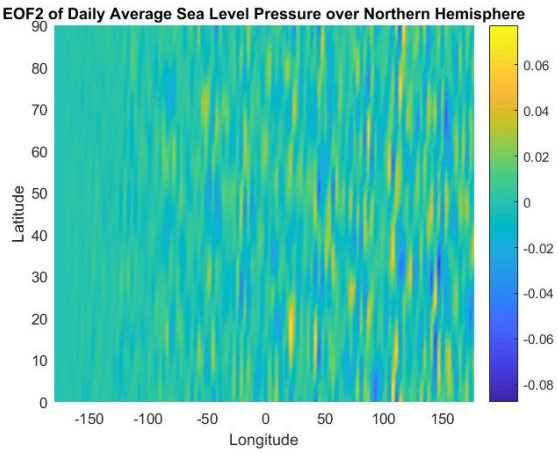

Figure 4: EOF 2 from my application of the EOF method on SLP data. Like EOF 1, there are no clear patterns. However, we the magnitude of EOF 2 values are generally higher than EOF 1.

5. Drawbacks of Empirical Orthogonal Functions

Although the method of EOFs is an extremely useful data analysis tool and has historically been able to extract useful insights when applied to climate data, the technique still has its flaws and limitations. First and foremost, the physical interpretation of EOFs can be difficult and hand-wavy. Due to the SVD, extracted EOFs are orthogonal to one another but what does it mean to be orthogonal in the physical space? Physical modes are not necessarily orthogonal so this constraint set by EOFs may be unrealistic [5]. In addition, EOFs are generally domain and geographic dependent which further limits their physical interpretability. One method to mitigate the limitations of EOFs is to use rotations through the rotated EOF: once you find an orthogonal basis, you can rotate the basis to one that can be better explained in terms of physical phenomena [6].

Furthermore, when using EOFs for dimensionality reduction, choosing the number of EOFs to discard may be difficult and, once again, hand-wavy. There is no rigorous method of selecting the number of EOFs to truncate, and people generally keep the number of EOFs in order to retain an arbitrary amount of variance, as I explained in section 3.

References

[1] http://www.met.rdg.ac.uk/~han/Monitor/eofprimer.pdf
[2] https://www.sciencedirect.com/science/article/pii/B9780128000663000061
[3] http://brunnur.vedur.is/pub/halldor/PICKUP/eof.pdf
[4] https://cims.nyu.edu/~cfgranda/pages/OBDA_fall17/notes/matrices.pdf
[5] https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/empirical-orthogona
[6] https://websites.pmc.ucsc.edu/~dmk/notes/EOFs/EOFs.html