**ATOC5860 – Application Lab #3**

**Empirical Orthogonal Function (EOF) Analysis**

**Note: This application lab requires netcdf4 and cartopy packages. Use the culabenv2022clean environment. See included culabenv2022clean.yml file**

**A reminder of the EOF/PCA Analysis Recipe – 5 steps**

**1) Prepare your data for analysis. Examples might include:**

**a) sub-setting the global data to a smaller domain**

**b) subtract the mean**

**b) standardizing the data (divide by the standard deviation)**

**d) cosine weighting (Account for the decrease in grid-box area as one approaches the pole (i.e. weight your data by the cosine of latitude)**

**e) detrend the data**

**f) remove the seasonal or diurnal cycle**

**g) remove NaN – EOF analysis does not work with missing data.**

**2) Calculate the EOFs and PCs using one of the two methods discussed in class: a) Eigenanalysis of the covariance matrix**

**b) Singular Value Decomposition (SVD).**

**3) Plot the first 10 eigenvalues (scaled as the percent variance explained) in order of variance explained. Add error bars following North et al. 1982. Describe how you determined the effective degrees of freedom N\*. How many statistically significant EOFs are there?**

**4) Plot EOF patterns and PC timeseries (usually just the first three or so unless you want to look at more).**

**5) Regress the data (unweighted data if applicable) onto standardize values of the 3 leading PCs. In other words, project the standardized principal component onto the original anomaly data X to get the EOF in physical units. You should have one regression pattern for each PC – i.e., the EOF pattern associated with a 1 standard deviation anomaly of the PC. *Note: The resulting patterns will be similar to the EOFs but not identical.***

**Notebook #1 – EOF analysis using images of people**

**ATOC5860\_applicationlab3\_eigenfaces.ipynb**

**LEARNING GOALS:**

1) Complete an EOF analysis using Singular Value Decomposition (SVD).

2) Provide a qualitative description of the results. What are the eigenvalues, the eigenvectors, and the principal components? What do you learn from each one about the space-time structure of your underlying dataset?

**DATA and UNDERLYING SCIENCE:**

In this notebook, you apply EOF analysis to a standard database for facial recognition: the At&t database.

<https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>

*“Our Database of Faces, (formerly 'The ORL Database of Faces'), contains a set of face images taken between April 1992 and April 1994 at the lab. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department.*

*There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).”*

The goal is to think a bit “out of the box” of Atmospheric and Oceanic Sciences about potential applications for the methods you are learning in this class for other applications.

**Questions to guide your analysis of Notebook #1:**

**1) Execute all code without making any modifications. What do the EOFs (spatial patterns) tell you? What do the PCs tell you? How do you interpret what you are finding?**

The EOFs help explain the structural patterns in the dataset. Since the EOFs represent the structural dimension, these are the “eigenfaces,” or the shape-like features you see in each individual image. The PCs represent the variance in the sampling dimension, so they show how these structural patterns vary from face to face.

**2) Reconstruct a face. How many EOFs do you need to reconstruct a face from the database? Does it depend on the face that it used?**

To simply show the broad outline of a face, very few EOFs (basically just 1) are needed. However, to start building more of the finer features beyond the head shape, more EFOs are needed. Not all faces result in as good of a reconstruction with very few EOFs. Roughly 80 EOFs are necessary to show that the original and reconstructed face match.

**3) Food for thought: The database contains 75% white men (**[**https://www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html**](https://www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html)**). How do you think this database limitation impacts the utility of the database for subjects who are not white men? What are some parallels that you might draw when analyzing atmospheric and oceanic sciences datasets? *Hint: Think about the limitations of extrapolation beyond the domain where you have data.***

In terms of representing a diverse population, this database is very limited. Since the database contains primarily white men, it would be very good at representing other white men, but probably not women or people of color. Therefore, this is a very useful dataset for demonstrating EOFs, but not for actually using it in a larger population.

The EOF analysis is completely dependent on the data it contains, so it is very difficult to extend the conclusions beyond what is already included. For instance, if you only have temperature and pressure data over land, it would be very difficult to make any conclusions about the relationship between temperature and pressure over the ocean. The analysis and conclusions are limited by the dataset contents.

**Notebook #2 – EOF analysis of Observed North Pacific Sea Surface Temperatures**

**ATOC5860\_applicationlab3\_eof\_analysis\_cosineweighting\_cartopy.ipynb**

**LEARNING GOALS:**

1) Complete an EOF analysis using the two methods discussed in class: eigenanalysis of the covariance matrix, Singular Value Decomposition (SVD). Check that they give the same results (They Should!).

2) Assess the statistical significance of the results, including estimating the effective sample size. (Lots more to think about here for estimating the autocorrelation and N\* in data…)

3) Provide a qualitative description of the results. What are the eigenvalue, the eigenvector, and the principal component? What do you learn from each one about the space-time structure of your underlying dataset?

4) Assess influence of data preparation on EOF results. What happens when you remove the seasonal cycle? What happens when you detrend? What happens when you cosine weight by latitude? What happens when you standardize your data (divide by standard deviation)? What happens when you compute anomalies?

**DATA and UNDERLYING SCIENCE:**

In this notebook, you will analyze observed monthly sea surface temperatures from HadISST (http://www.metoffice.gov.uk/hadobs/hadisst/data/download.html). The data are in netcdf format in a file called HadISST\_sst.nc. *Note that this file is ~500 MB so it might take a bit of time to download.* You will subset the data to only look at the North Pacific. Depending on how you prepare your data for analysis – you might expect to see different spatial patterns (eigenvectors) and different time series (principal components). Some things you might look for in your results are the Pacific Decadal Oscillation, “global warming”, the seasonal cycle, …. Depending on your data preparation – your hypothesis for what you should see in your EOF analysis should change. Note: In this dataset - land is NaN, sea ice is -999 – the notebook sets all values over land and sea ice to 0 for the EOF analysis.

**Questions to guide your analysis of Notebook #2:**

**1) Your first time through the notebook – Execute all code without making any modifications. Provide a physical interpretation for at least the first two EOFs and principal components (PC). What do the EOFs (spatial patterns) tell you? What do the PC time series for the EOFs tell you? What do you think of the method for estimating the effective sample size (Nstar)? Can you propose an alternative way to estimate Nstar? Do you get the same results using eigenanalysis and SVD? If you got a different sign do you think that is meaningful?.**

EOF 1: Pacific Decadal Oscillation (PDO)

EOF 2: North Pacific Gyre Oscillation (NPGO)

The EOFs show the spatial patterns, whereas the principal components show the variations in time.

The method of calculating N\* seems to be very conservative. While N is 804, this calculated N\* is only 49 (that’s just 6%!). That means that only 49 samples give independent information. Barnes suggests other ways to calculate N\* that vary in how conservative they are. However, it could be helpful to calculate N\* as one sample per season per year, or one sample per year.

The results for eigenanalysis and SVD are the same in magnitude, but for EOF 2, the signs are different. This does not have actual meaning, but is likely an artefact of the way the matrix was transposed.

Chart

Description automatically generated

**2) Save a copy of the notebook, rename it. Repeat the analysis but this time do not remove the seasonal cycle. What do you think you will see? Discuss your results with your neighbor. How do the EOFs and PC change? Was removing the seasonal cycle from the data useful? What impacts does removing the seasonal cycle have on your analysis?**

Since the seasonal cycle is so strong, I expect that the first EOF (which explains the most variance) will not be very instructive. However, subsequent EOFs will likely show interesting spatial patterns.

After removing the seasonal cycle, I have in fact found that the first EOF does not provide interesting information, but EOF 2 is rather instructive. Notably, it does not look like EOF 1 or 2 in the previous notebook.

**Chart

Description automatically generated**

**3) Save a copy of the notebook, rename it. Repeat the analysis but this time detrend the data. Discuss your results. How do the EOFs and PC change? Was detrending the data useful? What impacts does detrending have on your analysis?**

Detrending the data shows a similar pattern in EOF 1, but the signs are flipped. It’s also possible to see that PC 1 seems “flatter” than PC 1 in this method compared to the first method. In other words, the slope of the trendline in this method should be 0, whereas in the other methods, it’s greater than 0.

**Chart

Description automatically generated**

**4) Save a copy of the notebook, rename it. Repeat the analysis but this time do not apply the cosine weighting. Discuss your results. How do the EOFs and PC change? Was cosine weighting the data useful? What impacts does cosine weighting have on your analysis? What are examples of analyses where cosine weighting would be more/less important to do?**

Cosine weighting seemed to change the results fairly significantly. Not only is the pattern in EOF 1 shifted, and the sign is flipped, but the PC 1 has a very odd trend. The pattern in EOF 1 is “pulled” poleward, which seems to be a direct result of not applying the cosine weighting that takes into account changing area as latitudes increase.

**Graphical user interface, chart

Description automatically generated**

**5) Save a copy of the notebook, rename it. Repeat the analysis but this time do not standardize the data (i.e., comment out dividing by standard deviation). Discuss your results. How do the EOFs and PC change? Was standardizing the data useful? What impacts does standardizing the data have on your analysis?**

Surprisingly, not standardizing the data had very little effect on the outcome of EOF 1 and PC 1. The magnitude and shape of the EOF is nearly the same, and the PC shows a very change over time. Thus, we conclude that standardizing the data (in this case at least) is not as important to do compared to other preparatory tasks (e.g., cosine weighting).

**Graphical user interface, chart

Description automatically generated**