**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 EOF’s are the “eigenfaces” that are output by the analysis. These spatial patterns explain the most variability in the facial data in descending order. The principle components tell us how much of each eigenface is represented in a particular face with a weight value. To reconstruct a face, we would use a linear combination of the eigenfaces weighted by their principle components.**

1. **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?**

**Some faces are easier to reconstruct than the others because they are more related to full dataset. Face number 150 is not particularly related to the rest of the faces (because it is a woman). The quality of a picture also is a factor. Faces with clear defined lighting, like 8, is much easier to represent.**

1. **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.***

**The eigenfaces created will not be as good at representing faces that are not white men. This means that people who are not white men will require more EOFs to accurately depict their face. Also, using these eigenfaces to represent a face not included the dataset would work much better for white men than other faces.**

**In atmospheric science, it is important to include situations in your data that are representative of the situation you are trying to study. For example, if you were doing EOF analysis on the glacial/interglacial scale, it’s important to think about the scenario that you are interested. If you are more interested in studying interglacials, you want to make sure that your data has a large amount of interglacial points.**

**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?.**

EOF1 corresponds to the strength of the North Pacific Gyre. Depending on how strong this is, the current induces upwelling that inversely impact ocean temperature along the North American Coast and the center of the northern pacific, so it makes sense that this would be strongly representated in the data. I have a harder time interpreting the second EOF, but it seems to correspond to another gye that is further north.

The PC time series tell us how relevant each EOF is in explaining that year of SST data. Each PC give the corresponding “weight” used in the linear summation of EOF’s to reproduce the data.

Assuming that the data is completely independent is not a good assumption. Temperature data in particular has a lot of memory, and I would guess that using a decrease in N star would give us a more accurate analysis. The lag-1 autocorrelation is 0.886, which tells us that the data has a large amount of memory.

1. **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? Discus 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?**

**The seasonal cycle will dominate the variance. I predict that the first EOF will just be a net temperature increase that is roughly even, and its PC will cycle with a period of 12 years. The first EOF explains roughly 90% of the variance. Removing the seasonal cycle allows us to view more subtle impacts than just the seasonal cycle.**

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

**Detrending would be helpful if we observed a trend in an EOF. We would isolate variance that is not due to just an increase or decrease over time. The resulting EOFs were very similar, but I did notice that detrending made the values less extreme. This may be because any gradual warming is removed, making less variation overall. The first EOF also explains 6% more variance in the detrended data, but I am not sure why this is.**

1. **Save a copy of the notebook, rename it. Repeat the analysis but this time do not apply the cosine weighting. Discus 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?**

**I predict that removing the cosine weighting will make more extreme values occur near the equation in the EOF. Weighting accounts for the fact that more northern grid points take up more space and should have more say in the variance. I was surprised that removing the cosine weighting had no impact on the first EOF. This was unintuitive to me and I am not sure why this occurred.**

1. **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). Discus your results. How do the EOFs and PC change? Was standardizing the data useful? What impacts does standardizing the data have on your analysis?**

**I predict that the values will be more extreme, but otherwise it will not alter the shape of the EOFs. It was as I predicted. The shape of EOF1 is pretty much identical, but the resulting values are definitely more extreme.**