**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 are synthetic patterns that the EOF analysis creates that account for the most variance.

The principal components tell us exactly how well each sampling dimension (in this case each face) corresponds with the whichever eigenface we’re comparing it to. For example, PC1 would be a series of how all the faces in the sampling dimension compare to eigenface1 (or EOF1). A positive value would indicate that there is a positive correlation between them (some amount of variance is explained), and vice versa for a negative value.

In this case, the EOFs or Eigenfaces are “sort-of” human looking but incredibly distorted—they don’t look like an actual face. The first couple of EOFs look more male than female (in fact, all of the first 15 eigenvalues seem more male than female except for maybe #4 and #15)—likely because most of the faces in the dataset are male. In general, many of the eigenfaces look white as well for the same reason. Only Eigenface3 looks mostly black.

My main takeaways are that you get out what you put in; and to make sure that I give the EOF code enough data that could capture the variability that I’m trying to tease out. Simple example: if I give the EOF code only winter SST in the Pacific, it won’t be able to do a good job of predicting variability in the summer.

**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?**

It’s somewhat subjective when a photo is “good enough” but I tried to keep the same criteria each time.

White Male (face\_num = 150):

* Weights = 50 A close-up of a person

  Description automatically generated with medium confidence
* Weights = 100 A close-up of a person's face

  Description automatically generated with medium confidence
* Weights = 150 A close-up of a person

  Description automatically generated with medium confidence
* Weights = 200 Graphical user interface, application

  Description automatically generated

White Male w/ Glasses (face\_num = 10):

* Weights = 50 Graphical user interface, application

  Description automatically generated with medium confidence
* Weights = 100Graphical user interface

  Description automatically generated with low confidence
* Weights = 150 A close-up of a person

  Description automatically generated with low confidence
* Weights = 200 A close-up of a person's face

  Description automatically generated with medium confidence

White Female (face\_num = 333):

* Weights = 50 Graphical user interface, application

  Description automatically generated
* Weights = 100 Graphical user interface, application

  Description automatically generated
* Weights = 150 Graphical user interface, application

  Description automatically generated
* Weights = 200 Graphical user interface

  Description automatically generated
* Weights = 250 A close-up of two people

  Description automatically generated with low confidence

Black Male (face\_num = 206):

* Weights = 50 A picture containing calendar

  Description automatically generated
* Weights = 100 A close-up of two men

  Description automatically generated with medium confidence
* Weights = 150 A close-up of a person's face

  Description automatically generated with medium confidence
* Weights = 200 A picture containing calendar

  Description automatically generated
* Weights = 250 A picture containing calendar

  Description automatically generated

**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.***

I touched on this a little bit above, but it is evident when looking at the results above, especially in using the dataset to predict women. Arguably, the EOF analysis does a better job with black male faces than with white female faces. Face\_num # 150 above can be easily recognized even only using the first 50 PCs! Look at the difference between that and the woman (face\_num #333). Whether or not the person is smiling, or has their eyes open, or has glasses also matters quite a bit.

**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?.**

My initial thoughts at what EOF1 could be were that it looked potentially like a coastal kelvin wave propagation structure, or perhaps a reduced upwelling due to a reduction in southerly winds along the coast of North America. Perhaps also the North Pacific Gyre strength has been reduced, limiting the downwelling at the center of the gyre and producing a colder SST. Having taken a physical oceanography class, this also looks like the pattern of the PDO SST anomalies🡪 which is not 1 single mechanism, but a combination of many mechanisms. EOF2 looks like PDO-; or opposite forcings to those described above for EOF1. Again, these spatial patterns are the ones that the code finds that account for the most variance; they MAY not be anything physically, but given what we know about the PDO, it is likely that the code is picking up this dominant mode of variability in the north pacific.

The principal component time series (PC1 and PC2) tells us how well, over time, the sampling dataset correlates with EOFs 1 and 2. For PC1, in general there’s some back and forth, but sometimes the amplitude will remain positive or negative for several years. Also, it looks as if the amplitude is slightly increasing over time, but minimally. PC2 doesn’t look quite the same as PC1. While there is some randomness in the positive and negative directions, it also seems as if from 1950-1990 there is a negative trend, and after 1990, a positive trend of correlation. Perhaps there is a trend of SST warming that is also contributing to the variance explained by EOF2?

N\* in this case is calculated by averaging the entire SST dataset at each timestep and then calculating the lag-1 autocorrelation value of that averaging; of course the autocorrelation of averaged SST in the entire North Pacific is going to be fairly red. That being said, I don’t like how it is done, given that what EOF1 and EOF2 showed us was that there’s quite a bit of spatial variability in addition to temporal variability. We’re eliminating all of the spatial variability. Perhaps there is a way simultaneously do spatial and temporal autocorrelation.

For EOF1, SVD and Eigenanalysis provide the same results. For EOF2, they provide the same result but opposite sign. Having an opposite sign shouldn’t matter, as the PC will just be opposite; it still corresponds to the same amount of variability explained.

**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?**

I imagine that the dominant modes of variability captured by the EOF process will be the seasonal cycles rather than the PDO.

Turns out that the first EOF1 accounts for almost 90% of all variance explained 🡪 seasonal cycle.

**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?**

The EOFs and PC1 are fairly similar to the original dataset, although the amplitudes aren’t quite as high. PC2 is quite different, as expected because our initial hypothesis was that the original EOF2 had a lot of trends in it.

**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?**

The EOFs and PCs do look fairly similar, albeit with some minor differences at higher latitudes, where cosine weighting takes effect. I believe cosine weighting was still useful, even though the variance was stilled captured without it.

Cosine weight ins relevant whenever averaging information that spans large latitudes, especially nearest to the poles. You don’t want to over-weight the polar data by assuming that the amount of area between lat/lon lines is the same everywhere.

**4) 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?**

Standardizing the data make it so you can compare the relative importance of the variance of different parameters. We can also make it such that highly varying areas of the spatial distribution do not completely dominate all the variance found in our EOFs.

EOF1 looks fairly similar, while EOF2 has a similar pattern, but is much different along the coastal regions of North America. The PCs are fairly similar for both. The standardization likely allows the variance along the N.A. coast, and other lower magnitude variability areas to be more prevalent in the EOFs vs. the gyres.