Cheng Lu

ATOC7500 – Application Lab #3
Empirical Orthogonal Function (EOF) Analysis
in class October 5 and October 7, 2020

Note: This application lab requires netcdf4 and cartopy packages.

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

1) Prepare your data for analysis. Examples might include:
- **a) subsetting 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 pjysical 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
ATOC7500_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?

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?
100 is enough to get a blur one, 200 is pretty well. Slightly depends on the face used.

3) Food for thought: The database contains 75% white men (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.*

Yes. If one factor dominates the variance due to sample bias, it may not be able to show the pattern that actually contribute to the total variance but not showing nough in the samples.

Notebook #2 – EOF analysis of Observed North Pacific Sea Surface Temperatures
ATOC7500_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).
2) Assess the statistical significance of the results, including estimating the effective sample size.
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 #1:

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?.

Both of the first two EOFs show a pattern of positive phase of PDO, which is a cold pool at the center of the north Pacific. In time series, if the value is a large positive value, that means the real pattern contains a large portion of this EOF pattern. If it's a large negative pattern, then, it's still a large portion but the opposite phase. If the value is close to 0, then the real pattern is not like the EOF pattern. For Nstar, we averaged all the spatial data points first and then calculate the correlation. Another way can be calculate the correlation at all grid points and then pick up the lowest one. Averaging spatially may eliminate the real red information for example if the redness comes from the seasonal cycle but we averaged globally. But for the same case if we average a regional area that shares the same redness signal then it should be ok. The negative sign and positive sign are giving the same result, it's just randomly give a phase. We need to combine the sign of EOF with the sign of PC to get the real result.

2) <u>Save a copy of the notebook, rename it. Repeat the analysis but this time do not remove the seasonal cycle.</u> 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?

If not removing the seasonal cycle the most variance will be the seasonal cycle. The EOF is pretty even everywhere, and the PC changes from negative to positive once every year. If we don't want to find the seasonal cycle then removing it is very helpful.

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

The warm center is not that warm and more evenly distributed. I think without detrending, the pattern catches some variance from globe warming on the cold and warm centers.

4) <u>Save a copy of the notebook, rename it. Repeat the analysis but this time do not apply the cosine weighting.</u> 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?
Not much difference.

4) <u>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).</u> 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?

No obvious changes for the first EOF. I think without standardizing, the first EOF is similar to the real unit one. But the second EOF seems to change a lot. I don't quite understand it yet.