Michelle Maclennan

5 October 2020

**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?**

The EOF tells you the dominant modes of variability among the different faces. For example, it looks like the first EOF corresponds to the shape of the forehead and hair, and the second corresponds to the eyes and eyebrows. PCs explain the degree to which every individual face follows a certain mode of variability corresponding to an eigenvalue.

**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 takes about 150 weights to make a face recognizable, but it also really depends on the picture. For example, for picture index 359 the face is already recognizable with 100 weights. Conversely, a picture of a woman with her eyes closed doesn’t become clear no matter how many weights you pick, because the eyes are such a dominant mode of variability in the eofs. It seems like the positioning of the face relative to the camera is influential in determining the number of weights needed to make the faces visible, along with gender and skin color given most of the training photos are of white men.

**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.***

Yes, I absolutely think that the diversity of a database will impact an algorithm’s ability to predict or reconstruct a dataset. For example, it may be harder to reconstruct photographs of non-white, non-male faces. This is an indicator of how hard it is to extrapolate accurately from the domain of a dataset in atmospheric and oceanic sciences. Your sampling dataset may be missing important features of variability that you may see in an adjoining region.

**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?.**

These EOFs are examining monthly sea surface temperature over the North Pacific over 67 years. The first two EOF explain 21% and 13% of the variance in monthly SST, respectively.

Here are the PC time series and EOF for the first eigenvalue:

Chart

Description automatically generated

Here are the PC time series and EOF for the second eigenvalue:

Chart

Description automatically generated

The EOFs show the spatial modes of the Pacific Decadal Oscillation (PDO). The first EOF is showing the pattern with the eigenvalue that explains the highest amount of variance in the monthly SST time series, and the second shows the “alternate” phase of the PDO. The PC time series show the variability amplitude in variation of the SST. When the PC time series is positive, the monthly SST represents a similar pattern to the spatial EOF for that eigenvalue, and when it is negative the monthly SST has the opposite spatial pattern. For the EOF analysis corresponding to the second eigenvalue, the SVD method showed the opposite pattern of the eigenanalysis method. In multiplying the SVD results (spatial EOF and PC time series) by negative one, they yield the exact same result as the eigenanalysis. I don’t expect the sign difference to reach matter, because I think it’s more that there is a convention for plotting specific positive and negative phases of known modes of variability.

If we assume all samples are independent, then the sample size is 804. However, it seems unrealistic to assume that there is no month-to-month memory in SST anomalies, so it’s appropriate to calculate sample size assuming that not all samples are independent. The sample size is later calculated to be only 49 using the Wilks method (Barnes Ch. 2 Eq’n 88), which is a strong indicator that the ocean has a long memory for SST. The effective sample size could also be calculated using the Leith method. Elina mentioned a great point in our breakout room- it’s likely that since this dataset covers such a large area of the ocean, the sample size would probably be better off calculated regionally. It’s likely that different regions of the ocean experience different temperature memory.

**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?**

This time the EOF analysis is repeated without removing the seasonal cycle in temperature from the dataset. I expect that this will mean that the first EOF we see will be a spatial representation of the seasonal cycle in SST instead of the PDO. I also expect that the PC time series will show a bi-annual oscillation corresponding to seasonal variability.

As anticipated, about 90% of the variance is explained by the first eigenvalue. There are two positive phases with this mode – one of which corresponds to the SST in the summer months, and the other corresponds to SST in the winter months. Not removing the seasonal cycle is pretty unhelpful – it overshadows any other modes of variability we might see.

**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 are different from the non-detrended EOFs, noticeably in the spatial patterns corresponding to the first two eigenvalues. They still seem to identify the PDO, but the strength is slightly more muted even though the variance explained is higher. I think the decision to detrend data depends on what you’re looking at, because if you keep the trend in you may be seeing what is actually happening with the dataset in time. If the trend in a certain mode of variability is different from the long-term trend due to climate change, it may be helpful to detrend the dataset.

**4) 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?**

Because grid cells close to the poles have a smaller area than grid cells near the equator (due to the fact that we are applying a rectangular grid to a sphere), any features present in regions close to the north pole in this dataset may be undervalued in the EOF when compared to features near the equator. That said, there aren’t any huge visual differences in the special pattern of the EOF and the variance explained is about the same as well. Area weight is particularly important is you’re looking at the concentration of solutes in the ocean or the frequency of occurrence of certain events (like cyclones) over a polar region.

**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). 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?**

Standardizing the data did not make a huge difference on the EOFs. The variance explained by the first eigenvalue was slightly higher than in the original notebook but the spatial EOF did not really change. In general, I think it can be quite helpful to standardize datasets for EOFs because your PC time series can be more interpretable if it’s in units of standard deviations.