Autocorrelation Function in EOF Analysis

Sujata Goswami, Prof. Nico Sneeuw, Prof. Kamal Jain

***Abstract*—Empirical Orthogonal Function (EOF) analysis has been successfully applied in order to separate the noise from the time-series dataset. In EOF analysis, time-series is decomposed via singular value decomposition. Out of all these decomposed components some of them represent noise and some capture the dominant signals from it. Different rules has been defined to select the modes in order to recover noise free dataset as dominant variance rule, time-series based rules. Dominant variance rule was not able to select all the useful modes whereas time-series based rule sometimes were not able to distinguish clearly between the noise and signal. Then, the need to look into the autocorrelation function arises and finally it has been used as a mode selection tool in EOF analysis. Here, we have explained the use of autocorrelation function in EOF analysis and its advantage over other rules.**

***Index Terms*—autocorrelation, EOF analysis, Kolmogorov-Smirnov hypothesis, Bartlett hypothesis, Principal Component Analysis (PCA)**

I. Introduction:

Empirical Orthogonal Function (EOF) analysis or Principal Component Analysis (PCA) is used to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in it. Dimensionality reduction and retaining the dataset's variance is achieved by a linear coordinate transformation to a new set of basis vectors.

EOF analysis is used for capturing the dominant modes of a time-series of data in spatial and temporal domain. It is achieved by the method called as Singular Value Decomposition (SVD). EOF and PCA are the terms used interchangeably.

Any (m*n) matrix Z can be written as, the product of an (m*n) column-orthogonal matrix U, an (n*n) diagonal matrix S with positive or zero elements (the singular values), and the transpose of an (n*n) orthogonal matrix V. The decomposition looks like this,

$$Z_{(m*n)} = U_{(m*n)}\, S_{(n*n)}\, V^{T}_{(n*n)} \quad \text{---(1)}$$

'U' matrix is containing left singular vectors of Z; they are the Eigen vectors of matrix ZZ^T.
'V' matrix is containing right singular vectors of Z; they are the Eigen vectors of matrix Z^TZ.
'S' is a (n*n) diagonal matrix where, 'n' is the rank of the matrix. The decomposed components are also called as modes.

Data set is represented by the two formulas known as, Analysis and Synthesis, which are described as follows [8],

Having new basis vectors after calculation, EOF analysis is performed. Analysis means, projecting the data onto the new basis vectors.

The data matrix *Z* with dimensions (m*n) can be represented in terms of principal components as follows, Writing the identity,

$$EE^T = I \quad \text{--- (2)}$$

As follows,

$$Z=Z(EE^T) = (ZE)E^T \quad \text{--- (3)}$$

Where, *E* is the matrix representing Eigen vectors, and defines [8],

(Analysis): A = Z E; with dimensions (*m*n*) --- (4)

And,

Manuscript received April, 2014.
Sujata Goswami (MTech)*, Geomatics Engineering, Indian Institute of Technology Roorkee ., Roorkee, India,*
Prof. Nico Sneeuw*, Institute of Geodesy, University of Stuttgart, Stuttgart, Germany*
Prof. Kamal Jain*, Geomatics Engineering, Indian Institute of Technology Roorkee*

(Synthesis): $Z = A\,E^{T}$; with dimensions ($m*n$) --- (5)

II. Mode selection methods:

Now, there are different rules defined in order to select the modes out of total 'n' number of modes such that, the dataset reconstructed from these selected modes has reduced noise or striping error. Then, the dataset is reconstructed again from those selected modes. The different methods to select mode are:

- Dominant variance rule- It is defined on the basis of variance in the singular values.

The criteria of selecting the modes with signal content and throwing away all the noisy modes is based on the value of variance present in singular values. A cut-off is defined on the basis of high variance value in singular value matrix, then all the modes lower to that cut-off value are considered as signal and all other modes are considered as noise [8].

- Time-Series rules–

A time-series is a stochastic signal with chronologically ordered observations at regular interval. It describes the temporal behavior. In atmospheric science, where principal components can be very large in number, interest may be restricted only to the first few dominant and physically interpretable patterns of variation, even though their number is fewer than that associated with most PCA-based rules. To select a mode for reconstruction, frequency spectrum of the time-series also needs to be analyzed to study its behavior. Time-series rules are defined to differentiate it from white noise [6].

1) KS2 rule: Kolmogorov-Smirnov2 (KS2) rule is applied on the time-series of principal component of dataset. Time-series is compared against white noise. The distance between the two vectors is calculated; if maximum then it is signal, else noise. This way, the modes which are having signal with non-dominant variance values and still are different from white noise, they cannot get avoided, cf.(Fig.1).

2) Autocorrelation: The normalized auto covariance function is termed as autocorrelation. The covariance between two observations y_t and y_{t+h} of a stationary stochastic process is given as [3],

$$C_h = \text{cov}\,(y_t, y_{t+h}) = E\,[(y_t)\,(y_{t+h})] \quad \text{---- (6)}$$

This C_h in equation (6) is defined for all integral values of 'h' and together it is called the auto covariance function of y_t. It measures the covariance between pair at a distance or lag 'h' for all different values of 'h'. The auto covariance function represents the stochastic process which is normally distributed. It specifies the joint probability distribution along with the mean (μ).

Time-series is analyzed for the type of autocorrelation present in it. From the different patterns such as, weak autocorrelation, strong autocorrelation, sinusoidal behavior of the time-series can be predicted, [5].

Applications of autocorrelation: Autocorrelation is used to look for stationarity and non-stationarity, seasonality, randomness in time-series respectively.

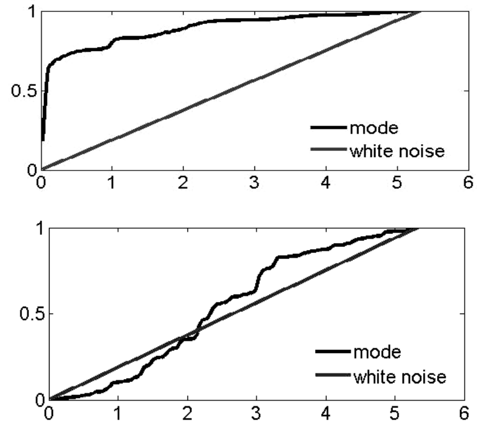

Fig.1 Kolmogorov-Smirnov hypothesis on the time-series (top – signal, bottom- noise)

III. EOF Analysis

When empirical orthogonal analysis is applied on the time-series dataset, it is decomposed into three components namely, spatial component, singular values and temporal component shown in figure, cf. (Fig. 2). The requirement of EOF analysis is that the data should be pre-smoothed and centered, [8]. EOF analysis is capable of capturing the dominant signals in the initial decomposed components which are interpreted to study the various types of signal. For interpretation purpose, spatial and temporal components are taken together and studied. For example, in figure 3 time-series shows the trend in positive direction with annual signal behavior corresponding to this are the Greenland region, Hudson Bay region in the spatial plot. To look for the increased or decreased water level, it is assumed that principal component time-series is of positive sign then corresponding regions in spatial plot shows decrease or increase in water level as it is there. The colors in the spatial map show the changes in the equivalent water height in various areas especially river basins on the earth.

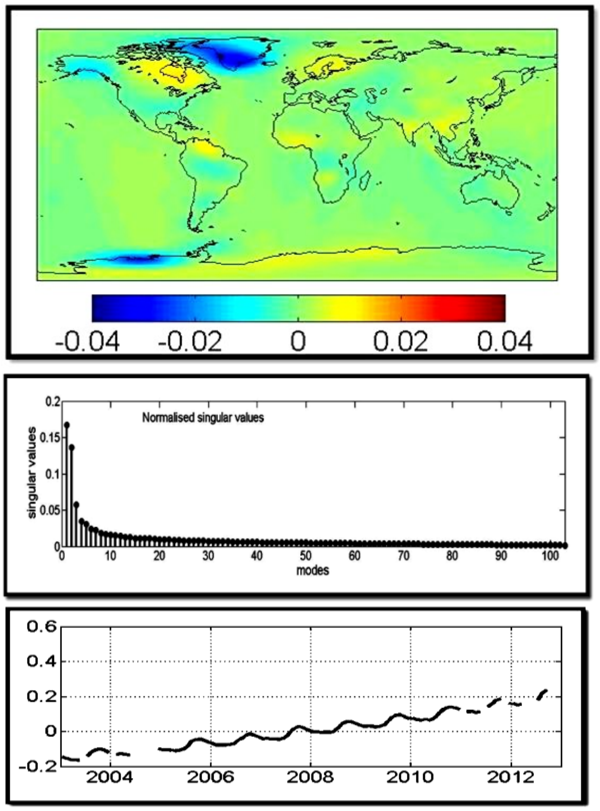

Fig. 2 Three decomposed components- spatial component, Singular values, temporal component (from top to bottom)

Mode selection:

Rule 1: Dominant variance rule - The variance values in the singular value matrix, with high value of variance define a boundary between the noise and signal. Decomposition separates the highest variance modes as some of the initial modes. The modes contain most of the information and less noise. As the variance value goes on decreasing, the amount of noise increases in the modes.

Rule 2: Time-series rule - Time -series is generally tested against white noise. There are rules based on the study of time-series, then selecting modes if it is different from white noise. It is not easy to distinguish to differentiate noise from signal by looking at the time-series only. Sometimes, noisy time-series also contain important seasonal signal, losing which means loss of important information. So, its corresponding frequency spectrum can be seen. It makes easy to get a clear picture of the small cycles, but some high frequency noise, alias signal bring the confusion back between the noise and signal. This high frequency noise and aliasing is very dangerous if considered as signal.

- Kolmogorov Smirnov hypothesis: It compares the power spectral density of the time-series with that of the white noise curve. If they are close to each other, means time-series is nothing but noise leading to the acceptance of hypothesis. Since, it distinguish between signal and noise on the basis of distance between the two curves, it is not able to distinguish the high frequency noise also it can discard modes with good seasonal information signal. Thus, there arises the need to look into the autocorrelation of the time-series.

Autocorrelation: The autocorrelation function of a time-series shows the correlation at its lagged versions. If it has little bit of signal content, it is reflected in the plot in the form of autocorrelation at adjacent lag values. Thus, no signal is lost and enough information is reflected to separate the noise. For autocorrelation, in order to separate noise Bartlett's hypothesis is used ([3]). Bartlett's hypothesis defines an interval on the basis of standard deviation of the time-series vector. The values lying inside this interval are close to the white noise. Therefore, modes are rejected if all the autocorrelation coefficient values lie near to zero, otherwise they are kept. Thus, modes can be selected on the basis of autocorrelation only.

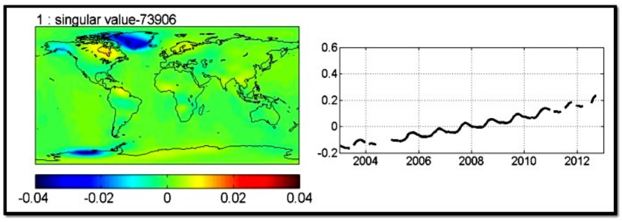

Fig. 3 Spatial (x axis- longitude, y-axis- latitude) and temporal components (x axis -year, y axis – principal component value)

IV. Future recommendations

Since, the autocorrelation plot is more detailed form for distinguishing the noise from signal. It can be used in atmospheric science in order to recover the signal from noisy time-series. Its future use as mode selection criteria in the Empirical orthogonal analysis can be recommended for global and regional studies. Its performance can be evaluated in terms of signal strength recovered in the noise-free dataset.

References:

[1] Bence. Analysis of short time series: Correcting for autocorrelation. 1995.

[2] Bentel. Empirical orthogonal function analysis of GRACE gravity data. Master’s thesis, University of Stuttgart, 2009.

[3] Piet M.T Broerson. *Automatic Autocorrelation and Spectral Analysis*. Springer, 2006.

[4] Jenkins. *Time-series and forecasting*. Tata McGraw hill.

[5] Jenkins. Exploratory data analysis. E-Handbook of Statistical Methods, 2012.

[6] Jollife. *Principal Component Analysis, Second Edition*. Springer Series, 2 edition, 2002.

[7] Jurgen Kusche. Filtering techniques and their potential application for generating new products for GRACE / GRACE-FO. *IAG-GEO workshop*, 2009.

[8] Mobley Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

[9] Wahr Swenson. Post-processing removal of correlated errors in GRACE data. *GEOPHYSICAL RESEARCH LETTERS*, VOL. 33, 2006