## Details and Variations on Empirical Orthogonal Functions

*Uncertainty and EOFs: How Many Modes Can We Use?*

Assigning uncertainties to EOFs can be a little tricky. There's no simple formula for the error bar on your EOFs, and the calculations for EOFs don't lend themselves to simple error propagation. One of the biggest questions one faces with EOFs involves deciding how many modes we actually believe.

Here are two rules of thumb for selecting EOF modes:

- If adjacent EOF modes explain nearly the same fraction of variance, then the modes aren't really distinguishable. That doesn't mean that they're meaningless, but it does mean that the modes should be treated as a coupled pair rather than as individual modes.
- North et al. (1982) suggest that the 95% confidence limit for an eigenvalue $\lambda$ is $\Delta\lambda = \lambda\sqrt{s/N^*}$, where $N^*$ is the number of degrees of freedom in the data set. Thus one strategy is to argue that if adjacent eigenvalues (or singular values) differ by less than $\Delta\lambda$, then the modes are unlikely to represent significant and unique spatio-temporal patterns.
- Another way to think about the number of usable EOF modes is to estimate which modes explain less variance than we might expect EOF modes computed from random noise to explain. You can evaluate this using an N-test. Generate 100 or 1000 Gaussian noise data sets of the same size as the original data set. Compute EOFs for each of the noise data sets. Now for each mode number, sort the 100 or 1000 singular values by size and identify 5th and 95th percentiles as cutoffs. Retain only the EOFs that exceed the noise floor. Figure 1 shows the N-test results for Pacific Equatorial temperature EOFs.

  Keep in mind that an N-test is an imperfect test; in particular the fraction of variance explained by mode 2 depends strongly on the fraction of variance explained by mode 1 (and so forth). Thus if you really want to know the probability that mode 8 is pure noise, you should be removing the impact of the first 7 modes. We don't normally do that. Nonetheless, the N-test is OK as a ballpark estimate to decide whether you should be trying to interpret 2, 4, or 15 EOF modes.

*Propagating Signals*

As we've discussed, basic Empirical Orthogonal Functions are good for identifying signals that involve simultaneous changes at multiple locations, but they are less effective at identifying propagating modes. A propagating signal might be represented by a single frequency/wavenumber pair in a two-dimensional spectrum. The energy associated with the propagating signal will be equally divided between two EOF modes. What are your options?

- *Frequency domain or complex EOFs* are designed to capture propagating modes. They are based on the notion that a propagating signal should contain signals that are orthogonal to each other, with a cosine/sine relationship. In other words, signals in the data matrix $\mathbf{X}$ also exist in the first (time) derivative of the data. A Hilbert transform provides a means to represent the first derivative of the data. To compute the Hilbert transform, we Fourier transform the data, multiple positive frequencies by $i$ and negative frequencies by $-i$, and then inverse Fourier transform to obtain $H(\mathbf{X})$. We then compute EOFs for our augmented data, $\mathbf{X} + iH(\mathbf{X})$. The resulting spatial and temporal modes will both be complex: the real and imaginary parts can be interpreted as the propagating signal at two stages separated by a quarter cycle.
- *Extended EOFs* are conceptually easier than complex EOFs, but they require you to make some prejudgments about likely timescales in the data. In extended EOFs, you assume that spatial patterns that occur at one point in time in your data are linked to spatial patterns that occur at a later point in time. Thus if $\mathbf{X}$ is an $N \times M$ matrix where $N$ is the time dimension, for extended EOFs, you define a new augmented data matrix $\mathbf{X}' = [\mathbf{X}(1 : N - n, :); \mathbf{X}(2 : N - n + 1, :); ...\mathbf{X}(n - 1 : N, :)]$. The EOFs of $\mathbf{X}'$ then provide an $M$-element spatial mode and an $(n - 1) \times (N - n)$-element temporal mode. A useful way to interpret these is to reconstruct the data corresponding to mode-1 or mode-2 variability to create a series of snapshots showing how a leading order signal propagates through the domain.

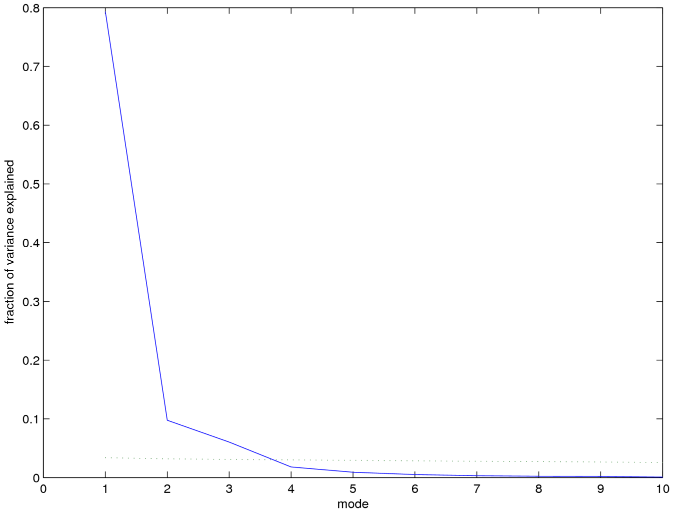

Figure 1: (Solid line) Fraction of variance explained by EOF modes 1 through 10 for Equatorial Pacific temperatures (considered in the MAE 127 notes). Dashed line shows 95th percentile results obtained with random noise.

*Canonical correlation and 'SVD'*

Sometimes we want to use EOF methods to determine how two different data fields are related to each other. Although we usually compute EOFs using a singular value decomposition of the original data matrix, another strategy is to compute EOFs using the eigenmode or singular value decomposition of the data covariance matrix $\langle \mathbf{X}^T\mathbf{X}\rangle$. A natural extension of this approach is to consider the singular value decomposition of the covariance matrix linking two different data sets: $\langle \mathbf{X}^T\mathbf{Y}\rangle$. In the atmospheric literature, this is often called 'SVD'. The singular vectors in this case represent the corresponding spatial (or temporal) modes for the two data sets, which together explain the maximum amount of the data covariance. This is sometimes called Maximum Covariance Analysis (MCA).

A variant of this is canonical correlation analysis, which is performed by using the leading order EOF modes for the two data sets $\mathbf{X}$ and $\mathbf{Y}$ to construct smoothed data sets $\mathbf{X}'$ and $\mathbf{Y}'$, which are then used to compute a covariance matrix, which is used for SVD/MCA.

Dennis Hartmann's notes on this topic are helpful:
http://www.atmos.washington.edu dennis/552_Notes_4.pdf

*PIPs, POPs, and rotated EOFs*

Finally, it's worth noting that there are a number of other EOF methods that are sometimes employed to analyze variability in data. Principal Oscillation Patterns (POPs) are good for explaining propagating patterns. Principal Interaction Patterns (PIPs) are a more general form of POPs that make use of a state space model that approximates the complex data field. Rotated EOFs are a strategy for shifting EOFs to a non-orthogonal basis with the intent of simplifying the interpretation. All of these methods are discussed in the textbook by von Storch and Zwiers.