

Empirical Orthogonal Functions

One of the most ubiquitous uses of eigenanalysis in data analysis is the construction of EOFs, the topic of this section. EOFs are a transform of the data; the original set of numbers is transformed into a different set with some desirable properties. In this sense the EOF transform is similar to other transforms such as the Fourier or Laplace transforms. In all these cases, we project the original data onto a set of orthogonal functions, thus replacing the original data with the set of projection coefficients on the basis vectors. However, the choice of the specific basis set varies from case to case.

In the Fourier case, for example, the choice is a set of sines and cosines of various frequencies. This is motivated by the desire to identify the principal modes of oscillation of the system. Thus if the signal projects strongly on sine waves of 2 frequencies, we will say that the signal is approximately the linear combination of these 2 frequencies. We will then attribute the remainder to other processes that are more weakly represented in the signal (the signal has low projection on them), and are thus assumed unimportant for the signal. Another important property for a basis is orthogonality (like sines or various frequencies); we would like to account for a certain component of the signal only once. An alternative to the sine/cosine set is a set of orthogonal polynomials, such as those named after Legendre. (Orthogonality often holds only over a specific interval, and sometime requires 'weighting functions'. These are related to the choice of metric, which we will talk about a bit.)

The representation of the signal in terms of the projection coefficients on a basis set is often very useful at separating cleanly various scales. For example, if our data is the sea surface temperature of a given ocean basin, we can think of the projection on the lowest frequency wave (the one which has one crest and one trough within the spatial extent of the domain) as representing the ocean's 'large-scale', while that on wavelengths of order 10-100 km as 'eddies'.

In EOF analysis we also project the original data on a set of orthogonal basis vectors. However, the choice of the basis is different. Here, the first EOF is chosen to be the pattern, without the constraint of a particular analytic form, on which the data project most strongly. In other words, the leading EOF (sometime called the 'gravest', or 'leading', mode) is the pattern most frequently realized. The second mode is the one most commonly realized under the constraint of orthogonality to the first one, the third is the most frequently realized pattern that is orthogonal to both higher modes, and so on. Hence the term 'empirical'; we still have an orthogonal basis, like the Fourier or Legendre bases, but whose members are not chosen based on analytic considerations, but based on maximization of the projection of the data on them.

Matrix and EOFs:

- [S-mode](#)

The EOFs are generally plotted as contour or vector maps, from which one can assess which regions are closely related, inversely related or unrelated, as well as identify centers of activity, regions with strong gradients, etc. The PCs, plotted as time series, quantify the overall strength of the associated EOF pattern over time. That is, the relative relationships between points on the grid (shown by the EOF) remain the same, but the absolute magnitude of the pattern changes with time (shown by the PC). The EOF/PC pairs (hereafter, modes) are ranked by their importance to the overall variability of the dataset. The relative importance of each mode is determined by its associated eigenvalue, which is used to calculate the variance attributable to that mode. In a dataset consisting of one or more strong signals (or physical forcing processes), most of the original variability is captured by the first few PC/EOF modes. In an S-mode analysis, the climatological mean is removed to generate t-centered data (i.e., time centered). At each grid point a separate climatological mean is calculated. If multiple fields contribute to the input data set (e.g., in a combined EOF) and have different units or levels of variability, some correction must be applied to the data to avoid having the field with greatest variance dominate the results. Normalizing each of the variables is one of the more common correction techniques, and is equivalent to using a correlation dispersion matrix.

- [T-mode*](#)

If the number of variables (grid points) is larger than the number of cases ($n > N$), the resultant covariance matrix ($n \times n$) will be even larger than the data matrix ($N \times n$), perhaps resulting in a problem too costly in computing terms. One option is to form the covariance matrix after swapping the variables and cases. So, instead of forming the covariance matrix from the time values at each station, one forms the matrix from the station values at each time interval. This is equivalent to a T-mode analysis. The resultant covariance matrix is of size $N \times N$, and the E matrix now represents the PCs and the Z matrix the EOFs.

EOF Examples:

- [Empirical Orthogonal Function analysis](#)
 - [Example of use: Spatial correlation maps based on EOF products*](#)
 - [An EOF Example with sinus functions](#)
 - [A manual for EOF and SVD analyses of climate data.](#)
-