Lecture 14: Empirical orthogonal functions

Recap

Last time we went through the mathematics underlying the singular value decomposition. Today, we'll put that framework to work to develop compact descriptions of data through empirical orthogonal functions.

Foundations

Often in oceanography we collect large data sets that are time series at a group of locations. Moored current meter arrays do just this. We may want to come up with a simpler description of the data than N time series. This description may be an end in itself, or more interestingly, may be the input to a linear estimator.

Suppose we have a time series which we write as an N-vector $\mathbf{y}(t)$. It is always possible to write a decomposition of $\mathbf{y}$ as

$$\mathbf{y}(t) = \sum_{i=1}^{N} \alpha_i(t)\mathbf{b}_i, \tag{1}$$

where the set of vectors $\mathbf{b}_i$ is orthonormal,

$$\mathbf{b}_i^T\mathbf{b}_j = \delta_{ij}, \tag{2}$$

and the temporal functions α_i are given by

$$\alpha_i = \mathbf{b}_i^T\mathbf{y}. \tag{3}$$

A simple example of such a decomposition is the case where the basis vector $\mathbf{b}_i$ has a one at position i and zeros elsewhere. The α_i are then the time series at those locations.

Our goal is to come up with a set of basis vectors such that the α_i are uncorrelated. In this new coordinate system,

$$\langle\alpha_i\alpha_j\rangle = \mathbf{b}_i^T\langle\mathbf{y}\mathbf{y}^T\rangle\mathbf{b}_j = \delta_{ij}\langle\alpha_i^2\rangle, \tag{4}$$

the covariance matrix of the α_i would be diagonal,

$$\mathbf{B}^T\langle\mathbf{y}\mathbf{y}^T\rangle\mathbf{B} = \mathbf{D}, \tag{5}$$

where $\mathbf{B}$ is the orthogonal matrix whose columns are the basis vectors $\mathbf{b}_i$, and $\mathbf{D}$ is a diagonal matrix whose elements are the variances of each of the α_i. Premultiplying (5) by $\mathbf{B}$ results in the eigensystem

$$\langle\mathbf{y}\mathbf{y}^T\rangle\mathbf{B} = \mathbf{B}\mathbf{D}. \tag{6}$$

We already know how to solve eigensystems, so we recognize that the diagonal of $\mathbf{D}$ is made up of the eigenvalues and the columns of $\mathbf{B}$ are the eigenvectors. The eigenvectors are commonly called **empirical orthogonal functions**, the temporal functions α_i are known as **amplitudes**, and the eigenvalues are the variances of the amplitudes. What we have essentially accomplished is a coordinate transformation such that the eigenvectors indicate those linear combinations of the data that are uncorrelated.

It turns out that the decomposition (1) where the basis vectors are the EOFs is optimum in another way. Suppose we desire a set of $K < N$ vectors that best approximate the data $\mathbf{y}$ in the sense that the mean square error is minimized. Our estimate is then

$$\hat{\mathbf{y}} = \sum_{i=1}^{K} \alpha_i\mathbf{b}_i, \tag{7}$$

and the measure of error to be minimized is

$$\left\langle (\hat{\mathbf{y}} - \mathbf{y})^T (\hat{\mathbf{y}} - \mathbf{y}) \right\rangle = \left\langle \mathbf{y}^T \mathbf{y} \right\rangle - \sum_{i=1}^{K} \langle \alpha_i^2 \rangle. \tag{8}$$

The expression above relies on the fact that the basis vectors are orthonormal according to (2). So the goal in finding the basis vectors is that the variance in the estimate, the second term on the right-hand side of (8), be maximized subject to the constraint (2). Using the method of Lagrange multipliers, the cost function to be maximized is

$$\mathcal{L} = \sum_{\rangle=\infty}^{\mathcal{K}} \left[\mathbf{b}_\rangle^{\mathcal{T}} \langle \mathbf{y}\mathbf{y}^{\mathcal{T}} \rangle \mathbf{b}_\rangle - \lambda_\rangle \left(\mathbf{b}_\rangle^{\mathcal{T}} \mathbf{b}_\rangle - \infty \right) \right]. \tag{9}$$

Extremizing this cost function with respect to $\mathbf{b}_i$ results in the equation to be solved:

$$\langle \mathbf{y}\mathbf{y}^T \rangle \mathbf{b}_i = \lambda_i \mathbf{b}_i. \tag{10}$$

This is of course identical to the eigensystem (6), and we find that the best K functions are the first K EOFs where the ordering the of eigenvalues is from largest to smallest. The first $K < N$ EOFs describe as much or more variance as any other possible set of K vectors subject to the normalization (2). It follows that a representation of the data with a different set of K vectors cannot produce a smaller mean square error than the first K EOFs. In this sense we say that the EOFs are the most "efficient" descriptors of variance. There are other sets of K vectors that would be just as efficient, but they must be in the subspace defined by the first K EOFs. Any set of K vectors which has components in the subspace of the $N - K$ higher indexed EOFs must be less efficient than the first K EOFs.

We have been discussing the EOF decomposition using a collection of time series at different locations. It is worth noting that the decomposition described by (1) may be made using data that have any two independent parameters. So far we have said the ensemble average is over time and we have found the EOFs to be vectors whose components are values at different locations, and that both the EOFs and the amplitudes obey orthogonality relations. There is nothing special about the independent variables of time and location. Other sorts of EOFs are sometimes used in the literature when the independent variables are different; examples are complex EOFs and frequency-domain EOFs. The basic idea is exactly the same although the definition of the ensemble average and/or the normalization condition (2) may differ.

Relationship to singular value decomposition

It turns out that the representation (1), where the basis vectors are EOFs, is exactly equivalent to the singular value decomposition of the $N \times L$ matrix $\mathbf{Y}$ whose rows are the N time series:

$$\mathbf{Y} = \begin{bmatrix} y_1(t_1) & y_1(t_2) & \cdots & y_1(t_L) \\ y_2(t_1) & y_2(t_2) & & \\ \vdots & & \ddots & \\ y_N(t_1) & & & y_N(t_L) \end{bmatrix} \tag{11}$$

The covariance matrix is then simply:

$$\langle \mathbf{y}\mathbf{y}^T \rangle = \frac{1}{L} \mathbf{Y}\mathbf{Y}^T \tag{12}$$

We know that the matrix $\mathbf{Y}$ has a singular value decomposition

$$\mathbf{Y} = \mathbf{USV}^T, \tag{13}$$

where the number of nonzero singular values indicate the rank of $\mathbf{Y}$. If $N < L$ and the rows (that is, the data) are linearly independent, then the rank would be N. Now the covariance matrix is equivalent to:

$$\frac{1}{L}\mathbf{YY}^T = \frac{1}{L}\mathbf{USS}^T\mathbf{U}^T \tag{14}$$

The right hand side is just the eigenvalue decomposition of the covariance matrix where matrix $\mathbf{SS}^T$ is square and diagonal with elements equal to $L\lambda_i$, and the columns of $\mathbf{U}$ are the EOFs. The amplitudes are given by the rows of the matrix

$$\mathbf{U}^T\mathbf{Y} = \mathbf{SV}^T \tag{15}$$

associated with nonzero singular values.

So the EOF decomposition is mathematically equivalent to a singular value decomposition. The important physical issue is whether the implicit ensemble average and normalization are appropriate to your particular problem. This is the same sort of question considered when we used the singular value decomposition to solve least-square problems. In that case we asked whether the norms were reasonable. It is important not to simply find the EOFs for some data set, which you can always do, and assume that the answer will be physically meaningful. The answer has relevance only if the average and normalization are appropriate.

EOFs and physical modes, weighting

A common misconception about EOFs is to associate them with physical modes. An easy way to see the folly in this is to consider the orthogonality relations:

$$(e_i, e_j) = \delta_{ij} \text{ for physical modes} \tag{16}$$

$$\mathbf{b}_i^T\mathbf{b}_j = \delta_{ij} \text{ for EOFs} \tag{17}$$

These are not the same. The inner product for normal modes in the ocean, for instance, is a vertical integral weighted by the buoyancy frequency squared, while that for EOFs is a simple dot product. The normalization for a physical mode is generally an expression of some conserved quantity such as energy. The normalization for EOFs is somewhat arbitrary, depending on the locations of the data. While a dominant physical process ought to be represented in the first few EOFs, there is no one-to-one correspondence between physical modes and EOFs.

We can attempt to approximate the integral in (16), thus changing our estimate of energy, by weighting and using the orthogonality relation:

$$\mathbf{b}_i^T\mathbf{W}\mathbf{b}_j = \delta_{ij}. \tag{18}$$

Similarly we might use some sort of weighting in the definition of our ensemble average. In any case the problem is easily solved by transforming to new coordinates

$$\mathbf{b}' = \mathbf{W}^{1/2}\mathbf{b}, \tag{19}$$

and solving the eigensystem

$$\mathbf{W}^{-1/2}\langle\mathbf{yy}^T\rangle\mathbf{W}^{-1/2}\mathbf{b}' = \lambda\mathbf{b}' \tag{20}$$

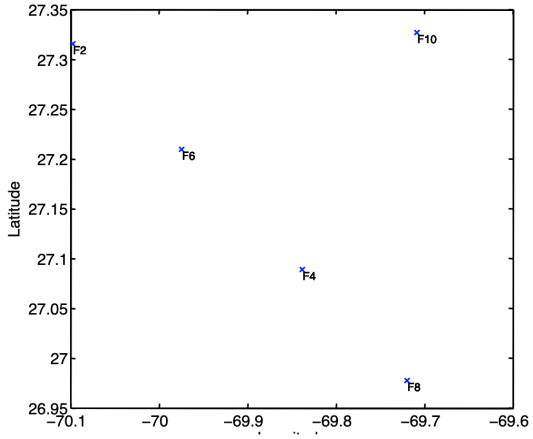

Figure 1: Mooring locations in the Frontal Air-Sea Interaction Experiment (FASINEX).

All of the results above immediately follow.

Moored temperature data, an example

As an example, consider moored temperature data from the North Atlantic, taken in the vicinity of the subtropical front. An array of five moorings was deployed to observe air-sea interaction in the region of the front (Figure 1). The array was in the form of a right triangle, with the hypotenuse oriented roughly parallel to the anticipated acrossfront direction. A total of 38 temperature sensors were divided roughly evenly among the 5 moorings, with each time series 2448 hours in length.

The temperature time series from one of the moorings reveals some of the processes occurring in the region (Figure 2). The upper 40 m are mostly uniform in temperature, as the curves for these time series nearly always overlap. The main exception to this patter is that temperature at 1 m is warmer on a daily cycle, especially near the end of the record as spring warming sets in. From 80-160 m, the ocean is stratified, although the stratification weakens during the last month of the record. The largest variability is at a period of nearly one month, and is present at all depths, although it is strongest at 160 m. This variability is caused by the passage of mesoscale fronts and eddies, which were a focus of the experiment.

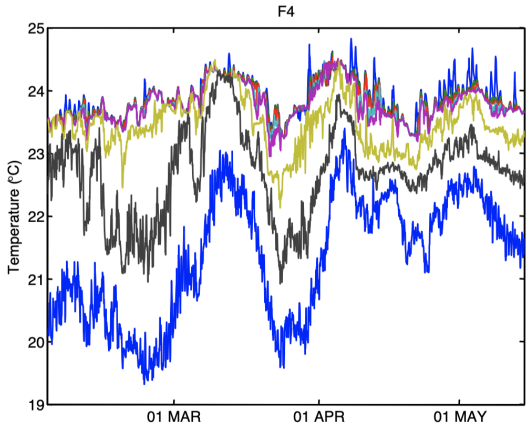

Figure 2: Temperature records from FASINEX mooring F4 at depths 1, 10, 20, 30, 40, 80, 120, and 160 m.

EOFs and amplitudes of the temperature data are found by doing a singular value decomposition as described in (11-15). The first step in the analysis is to examine the fraction of variance accounted for by each EOF (Figure 3). The first EOF accounts for 57% of the variance, the second 18%, and so on. The first 5 EOFs together account for 90% of the variance, so we get a reasonably complete description of the variance with only 5 spatial functions (EOFs) and time series (amplitudes). If the goal is a description of the largest variance, this is certainly a simpler than having to look at all 38 temperature time series.

The first EOF of temperature describes most of the variance observed in all 38 time series (Figure 4). Temperature at all depths and moorings is in phase, with the largest signal at depth. The EOF is nearly uniform in the upper 40 m, consistent with the presence of a mixed layer. The structure of the amplitude is similar to the large fluctuations in Figure 2. A first mode like this one, with all locations in phase is common in EOF analysis.

The second EOF describes variability that is of opposite phase on either side of the array (Figure 4). The combination of the first two EOFs are able to describe the passage of mesoscale features past the array. The second EOF is generally larger at depth, consistent with Figure 2. Playing this game of describing EOFs is fun, but not necessarily useful as there is no guarantee that the EOFS have a one-to-one correspondence to physical modes. However, no other functions can describe more variance than the EOFs, even if the physical modes were known perfectly.

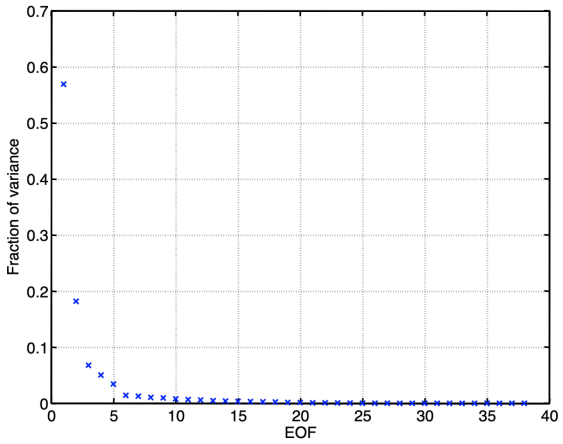

Figure 3: The fraction of variance accounted for by the EOFs of temperature on the FASINEX moorings.

Testing EOFs on white noise

Finally we can carry out an exercise to ask what happens if we compute EOFs on a matrix of random white noise. We can define a data set (in this case a 100×10 matrix:

```
A=randn(100,10);
```

Then we can decompose the data, either by computing the svd of the full matrix, or by converting the matrix $\mathbf{A}$ into a covariance matrix and finding its eigenvalues or svd. All of these produce the same results, although the eigenvalue solver orders the eigenvalues from smallest to largest:

```
[u,s,v]=svd(A);
[eu,es]=eig(A'*A);
[uu,ss,vv]=svd(A'*A);
```

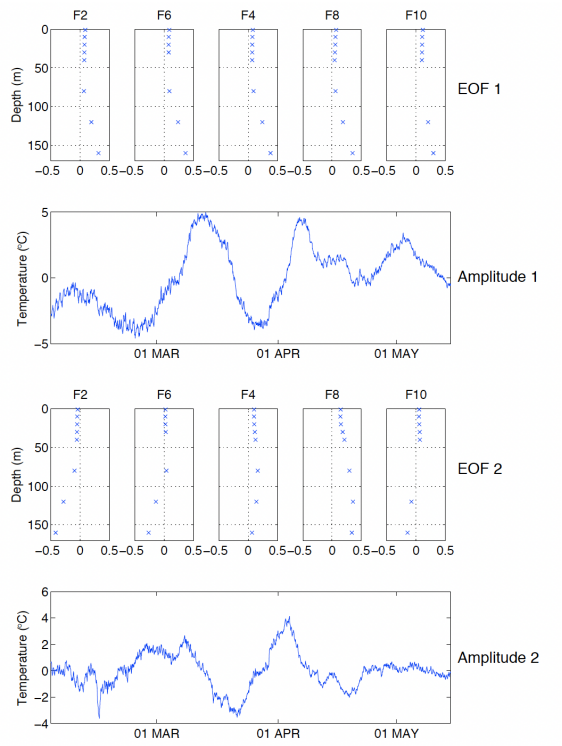

Figure 4: First two EOFs and amplitudes of temperature. The EOFs are shown by crosses as a function of depth on each mooring, and the amplitudes are shown as time series. In this case, the EOFs are dimensionless and the amplitudes carry the units of °C.

```
% what is the difference between the eigenvalue decomposition
% and the svd?
[diag(s) diag(es) diag(ss) diag(s).^2]
```

One question you can ask is how much of the variance is explained by each mode. You might be tempted to compute this using the singular values, but for variance you really need the squared singular values:

```
% how much of the variance is explained by each mode?
plot(diag(s).^2/sum(diag(s).^2),'LineWidth',2)
h=gca;
set(gca,'FontSize',14)
xlabel('Mode number','FontSize',14)
ylabel('Fraction of variance explained','FontSize',14)
```

You can also plot the structure of the modes by plotting the vectors $\mathbf{U}$ and $\mathbf{V}$, in this case, in Matlab, the columns of the matrices, where the first column is mode 1, the second column is mode 2, and so forth.

One thing we sometimes do with EOFs is to reconstruct the data using just the first few modes. The recoonstruct the first mode, you'd use

```
% reconstruct the data mode by mode?  Mode 1:
A1=u(:,1)*s(1,1)*v(:,1)';
```

Once you've done this, you might ask how the variance in mode 1 compares with the variance inferred from the singular values alone.

```
% variance in mode 1 reconstruction vs total variance
sum(A1(:).^2)/sum(A(:).^2)

% variance inferred from singular value 1:
s(1,1)^2/sum(diag(s).^2)
```

In this particular case, our data are white noise, so we expect no skill whatsoever from the EOFs. In this example, in the case of data that are white noise, the fraction of variance explained helps us address the null hypothesis: for a matrix with a given number of degrees of freedom, how much apparent skill would we see purely by chance? Formally, researchers sometimes use a rule-of-thumb text called the N-test, in which they compare EOFs for noise with true EOFs. To carry out the test, plot the fraction of variance explained for the true data, as a function of mode number, and plot the fraction of variance explained for noise with equivalent matrix dimensions (either white noise, or noise that has been filtered to match the effective smoothing in the data). The point N, where the lines cross provides a rough indication of the number of modes that provide more skill than we'd expect from pure noise. Beyond that point, it s hard to justify continuing to attempt to interpret EOFs. The N-test is a rough rule of thumb, but a useful starting point. For more details, see Preisendorfer (1988).

Bibliography

Preisendorfer, 1988. *Principal Component Analysis in Meteorology and Oceanography*, Elsevier, 425 pp.