Derivation of EOF

Yuchen Li

June 2023

Notation: In this document I use $\mathbf{a}^T\mathbf{a}$ instead of the dot product $\mathbf{a} \cdot \mathbf{a}$. They are equivalent.

1 Calculating anomaly matrix

Suppose we have a m by n matrix X that represents some spatiotemporal data with m time measurements and n spatial measurements. Then $(X_{k1}, X_{k2}, \cdots, X_{kn})$ represents the data at the k'th time step. If we want to study the variability in this data over time and space, we might care only about the anomalies, or the difference from the time-mean, rather than the actual values.

The time-mean at each spatial location can be computed by averaging over the m time measurements:

$$\mu_j = \frac{1}{m} \sum_{k=1}^{m} X_{kj} \tag{1}$$

where μ_j is the time-mean at spatial location j. Let $\bar{\mathbf{x}} = (\mu_1, \cdots, \mu_n)$ be the time-mean at all spatial locations. We then compute the anomaly matrix X':

$$X' = X - \mathbf{1}\bar{\mathbf{x}}^T \tag{2}$$

where $\mathbf{1}$ is the ones vector and $\mathbf{1}\bar{\mathbf{x}}^T$ is a matrix with the time mean for each spatial location in every row. X' has the nice property that its time-mean at each spatial location is zero; this will simplify calculations later.

2 Maximizing variance

The high level goal of EOF/PCA is to find a new set of coordinates (or basis) that maximizes the variance (a measure of the spread) of the data. If we project the data into this new coordinate system, then we will see distinct "modes" of variability that dominate the overall spatiotemporal variability of the data.

The *variance* of a vector $\mathbf{x}$ is defined as the average of the squared values of the vector minus the square of the average value of the vector. In vector language, it can be written as

$$\text{var}(\mathbf{x}) = \frac{1}{n} (\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{x} - \bar{\mathbf{x}}) \tag{3}$$

if n is the number of elements in $\mathbf{x}$. If you haven't encountered variance before, the square might be a confusing choice; the point is that since squares are always positive, a negative deviation from the mean would also increase the variance. This variance formula gives a measure of how spread out the values of $\mathbf{x}$ are around its mean.

Going back to EOF, we seek a n-dimensional vector $\mathbf{a}$ that represents "weights" for each spatial location such that the linear combination of the spatial locations according to weights given by $\mathbf{a}$ yields the greatest variance over time. That is, we seek $\mathbf{a}$ for

$$\max(\text{var}(X'\mathbf{a})) \tag{4}$$

You might notice that if the entries of $\mathbf{a}$ become arbitrarily large, we end up inflating the variance without bound. In the end, we only seek a direction for $\mathbf{a}$ (i.e., all we need are *relative* weights), so we constrain the length of $\mathbf{a}$ to be 1: $\mathbf{a}^T\mathbf{a} = 1$.

Now we show that the mean of $X'\mathbf{a}$ is zero:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}(X'\mathbf{a})_i &= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}X'_{ij}a_j \\ &= \sum_{j=1}^{n}a_j\left(\frac{1}{m}\sum_{i=1}^{m}X'_{ij}\right) \\ &= \sum_{j=1}^{n}a_j(0) = 0\end{aligned}$$

where we have used the fact that the time mean of the anomaly matrix X' at any location is zero, by construction. Hence, the variance of $X'\mathbf{a}$ simplifies to

$$\begin{aligned}\mathrm{var}(X'\mathbf{a}) &= \frac{1}{m}(X'\mathbf{a})^T(X'\mathbf{a}) \\ &= \frac{1}{m}\mathbf{a}^T X'^T X'\mathbf{a}\end{aligned}$$

How do we find a direction for $\mathbf{a}$ that maximizes this expression?

3 The covariance matrix

We define the *covariance* between two variables $\mathbf{x}$ and $\mathbf{y}$ to be the average of the product of their deviations from their respective means. Mathematically, this is given by:

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{m}(\mathbf{x}-\bar{\mathbf{x}})^T(\mathbf{y}-\bar{\mathbf{y}}) \tag{5}$$

where $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}x_i$ and $\bar{\mathbf{y}} = \frac{1}{m}\sum_{i=1}^{m}y_i$ are the means of the variables $\mathbf{x}$ and $\mathbf{y}$, respectively. Note the similarity to the definition of variance; in fact, variance is simply the covariance of a variable with itself.

The *covariance matrix*, denoted as Σ, for a set of m-dimensional variables $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is a $n \times n$ matrix, where the element in the ith row and jth column is the covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$. If we have a data matrix X with each column representing one variable, the covariance matrix can be computed as:

$$\Sigma = \frac{1}{m}(X - \mathbf{1}\bar{X}^T)^T(X - \mathbf{1}\bar{X}^T) \tag{6}$$

where $\bar{X}$ is a row vector containing the means of each column of X, and $\mathbf{1}$ is a ones vector of length m. This matrix Σ represents all pairwise covariances for n variables and provides a measure of the degree to which pairs of the variables $\mathbf{x}_i$ and $\mathbf{x}_j$ change together.

We note that for our anomaly matrix X', the time-means of each spatial variable are zero, so $\bar{X}'$ vanishes! Thus,

$$\Sigma = \frac{1}{m}X'^T X' \tag{7}$$

for our anomaly matrix X'. We note that Σ is symmetric, since cov($\mathbf{x}$, $\mathbf{y}$) = cov($\mathbf{y}$, $\mathbf{x}$).

4 Eigenvalue problem

Returning to our maximization problem, we now see that var($X'\mathbf{a}$) can be written as $\mathbf{a}^T\Sigma\mathbf{a}$ where Σ is the covariance matrix for variables $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ representing all spatial locations. Maximizing $\mathbf{a}^T\Sigma\mathbf{a}$ turns out to be an *eigenvalue* problem, i.e., finding $\mathbf{a}$ such that $\Sigma\mathbf{a} = \lambda\mathbf{a}$ for a scalar eigenvalue λ. We explain this below.

To find the vector $\mathbf{a}$ that maximizes the expression $\mathbf{a}^T\Sigma\mathbf{a}$ subject to the constraint $\mathbf{a}^T\mathbf{a} = 1$, we use the method of Lagrange multipliers. This involves setting up a Lagrangian function $\mathcal{L}$ that combines the original function and the constraint, like so:

$$\mathcal{L}(\mathbf{a}, \lambda) = \mathbf{a}^T\Sigma\mathbf{a} - \lambda(\mathbf{a}^T\mathbf{a} - 1) \tag{8}$$

Here, λ is the Lagrange multiplier that we introduce to handle the constraint. The next step is to find the vector $\mathbf{a}$ and the scalar λ that make the derivative of this function equal to zero. This is equivalent to solving the following system of equations:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{a}} = 2\Sigma\mathbf{a} - 2\lambda\mathbf{a} = 0 \tag{9}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = \mathbf{a}^T\mathbf{a} - 1 = 0 \tag{10}$$

The first equation simplifies to $\Sigma\mathbf{a} = \lambda\mathbf{a}$, which is an eigenvalue equation. This equation has non-trivial solutions (i.e., solutions where $\mathbf{a}$ is not the zero vector) precisely when values of λ are eigenvalues of Σ. The corresponding solutions for $\mathbf{a}$ are the eigenvectors of Σ. The second equation is just the original constraint that $\mathbf{a}$ be a unit vector.

5 Punchline

We have shown that $\mathrm{var}(X'\mathbf{a})$ is maximized precisely when $\Sigma\mathbf{a} = \lambda\mathbf{a}$. Going back to our expression for $\mathrm{var}(X'\mathbf{a})$, we have then that

$$\mathrm{var}(X'\mathbf{a}) = \mathbf{a}^T\Sigma\mathbf{a} = \mathbf{a}^T\lambda\mathbf{a} = \lambda \tag{11}$$

where λ is an eigenvalue of Σ. In words,

> The variance is maximized when $\mathbf{a}$ is the eigenvector corresponding to the biggest eigenvalue $\lambda_{\max}$. The variance is $\lambda_{\max}$.

Since Σ is symmetric, it is diagonizable and has n such eigenvalues λ. We call $\mathbf{a}_k$ the k'th EOF of X' and it is associated with explaining a variance of λ_k. The beauty of EOF is that since λ_k's generally fall off quickly with increasing k, we can represent most of the variability in a spatiotemporal dataset by taking only the first few EOFs! This is also why EOF/PCA is a common dimensionality-reduction algorithm.

Since an EOF is an eigenvector in the spatial phase space (the space spanned by all of the spatial variables), it has the same dimension as the number of spatial variables, and corresponds to a spatial pattern. The *principal component* (PC) is a time series and corresponds to how much an EOF pattern "shows up" at every given time. We define the k'th principal component $\mathbf{p}_k = X'\mathbf{a}_k$ where $\mathbf{a}_k$ is the k'th EOF, such that

$$p_k(t) = \sum_{s=1}^{n} X'_{t,s} a_s^k \tag{12}$$

where a_s^k is the s'th component of $\mathbf{a}_k$. Since Σ is symmetric, the Spectral Theorem tells us that the eigenvectors form an orthogonal basis for $\mathbf{R}^n$. By construction, all EOFs and their principal components are orthogonal, so the principal components are pairwise uncorrelated. The original spatiotemporal anomaly field can thus be reconstructed from only the EOFs $\mathbf{a}_k$ and the principal components $\mathbf{p}_k$:

$$X'(t, s) = \sum_{k=1}^{n} p_k(t) a_k(s) \tag{13}$$

> The k'th EOF $\mathbf{a}_k$ corresponds to a spatial pattern that explains λ_k amount of variance in the dataset. The PC $\mathbf{p}_k$ for that EOF is a time series representing how much the EOF shows up at each time step.