Empirical Orthogonal Functions

Juan José Gómez Navarro

23 de mayo de 2008

MAR

Index

Data

- In climate research, we have to deal with a huge amount of data.
- This data is full of noise (weather is just noise).

"Climate is what we hope, but weather is what we get"

Mark Twain

- The information is also very redundant, because of correlation.
- We would like to get the main information with as few numbers as possible.

Empirical Orthogonal Functions

Empirical Orthogonal Functions (EOF) is a statistical tool which allow us to reduce the large dimensionality of weather data, and also help us to get physical interpretations of our result.

To understand correctly this tool, we need basic knowledge of statistics, but also, and mainly, of matrix algebra.

We will describe the most importants theorems which make this tool possible. We will also try to get some intuition of why does this tool work, and will show some examples at the end.

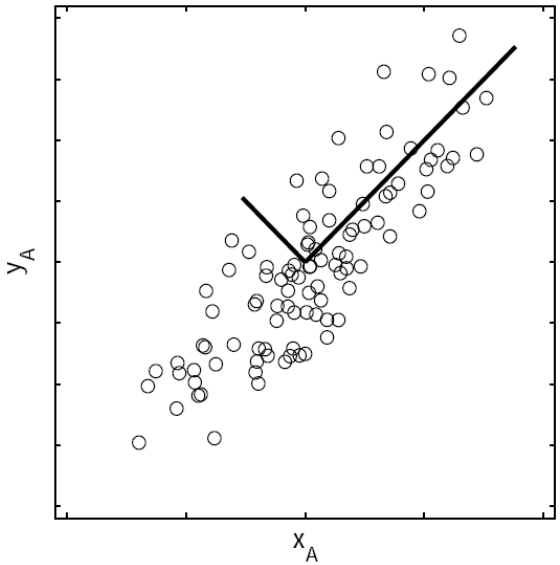
Index

Map's vectorial space

- We have n maps.
- Each one is a p-dimensional vector (usually, we have $n \ll p$).
- So these vectors belong to a p-dimensional vectorial space, and are expressed in the canonical p-dimensional basis.
- Is not there other more convenient basis to express our data?
- The aim is to get profit of correlation to reduce dimensionality of the problem.

Getting some intuition

The key is the correlation. Let's see an example:



- We have a two dimensional data, x and y.
- In the picture, we can see a more convenient choice of axis to express the data.
- These axis maximize the variance.
- This only have sense with highly correlated data.

Getting some intuition

- Correlation is implicit in the covariance matrix (wich depends on the basis!).
- The diagonal elements are just the variance of each dot in the maps.
- The non-diagonal elements are the correlation between dots in the maps.
- We would like to minimize the correlation,because this means redundancy in the data set.
- To take advantage of correlation, we would like to select a basis wich minimize the non-diagonal elements.
- All this sounds like we need to diagonalize something... let's formalize this idea!

Basis

If $\{\hat{u}_i\}_{i=1,2,\ldots,p}$ is an orthonormal basis of $\mathbb{R}^p$, then we can write any vector $\vec{X}_i$ in our data as the sum

Linear combination

$$\vec{X}_i = \sum_{j=1}^{p} \left(\vec{X}_i \cdot \hat{u}_j \right) \hat{u}_j.$$

Going a bit ahead, what we would like is to get the approximation

What we would like to get

$$\vec{X}_i = \sum_{j=1}^{p} \left(\vec{X}_i \cdot \hat{u}_j \right) \hat{u}_j \simeq \sum_{j=1}^{q} \left(\vec{X}_i \cdot \hat{u}_j \right) \hat{u}_j.$$

MAR

Data matrix

But we can also set the information in matrix form. For instance in a matrix with n rows (temporal dimension) and p columns (spatial dimension):

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Change of basis like a matrix multiplication

If we set U as a matrix which columns are the vectors of any orthonormal basis $\{\hat{u}\}$,

$$U = \begin{pmatrix} | & | & & | \\ \hat{u}_1 & \hat{u}_2 & \dots & \hat{u}_p \\ | & | & & | \end{pmatrix}$$

then $UU^t = I$. This is, U is an orthogonal matrix.

In this way, the last change of basis is just a matrix multiplication:

Matrix point of view of a change of basis

$$X = (XU)U^t.$$

- $(XU)_{np}$ will be the Principals Components matrix (PCs).
- $(U^t)_{np}$ will be the EOFs matrix.
- But for being, we have to choose carefully the elements of the basis $\{\hat{u}\}$.

Approximation

- Let's suppose we retain only one map (that is, the first EOF):

$$\vec{X}_i = \sum_{j=1}^{p} (\vec{X}_i \cdot \hat{u}_j)\hat{u}_j \simeq (\vec{X}_i \cdot \hat{u}_1)\hat{u}_1, \quad i = 1, 2, ..., n.$$

- $(\vec{X}_i \cdot \hat{u}_1)$ is a series of numbers, the weights of first EOF in every moment. It is what we call the first Principal Component.

- We are making a quadratic error

$$\epsilon_1 = \sum_{i=1}^{n} \frac{1}{n} ||\vec{X}_i - (\vec{X}_i \cdot \hat{u}_1)\hat{u}_1||^2.$$

- The point is, of course, to chose $\hat{u}_1$ in order to set the error as small as possible.

Developing the error

Working a bit...

$$\begin{aligned}
\epsilon_1 &= \sum_{i=1}^{n} \frac{1}{n} ||\vec{X}_i - (\vec{X}_i \cdot \hat{u}_1)\hat{u}_1||^2 \\
&= \sum_{i=1}^{n} \frac{1}{n} \left[||\vec{X}_i||^2 + ||\hat{u}_1||^2 |\vec{X}_i \cdot \hat{u}_1|^2 - 2|\vec{X}_i \cdot \hat{u}_1|^2 \right] \\
&= \underbrace{\sum_{i=1}^{n} \frac{1}{n} ||\vec{X}_i||^2}_{(a)} - \underbrace{\sum_{i=1}^{n} \frac{1}{n} |\vec{X}_i \cdot \hat{u}_1|^2}_{(b)}
\end{aligned}$$

Let's see each term separately.

Developing of (a)

If we suppose that data is zero-averaged:

$$\sum_{i=1}^{n} \frac{1}{n} \vec{X}_i = \vec{0} \quad \left(\text{or better} \quad \sum_{i=1}^{n} \frac{1}{n} x_{ij} = 0 \quad j = 1, 2, ..., p \right)$$

In this case, (a) is the total variance of data

$$(a) = \sum_{i=1}^{n} \frac{1}{n} ||\vec{X}_i||^2 - \left|\left| \sum_{i=1}^{n} \frac{1}{n} \vec{X}_i \right|\right|^2 = \text{Var}(\vec{X}).$$

Developing of (b)

$$
\begin{aligned}
(b) &= \sum_{i=1}^{n} \frac{1}{n} |\vec{X}_i \cdot \hat{u}_1|^2 \\
&= \sum_{i=1}^{n} \frac{1}{n} \left(\sum_{k=1}^{p} x_{ik} u_{k1} \right) \left(\sum_{l=1}^{p} x_{il} u_{l1} \right) \\
&= \sum_{k=1}^{p} u_{k1} \underbrace{\left(\sum_{i=1}^{n} \frac{1}{n} x_{ik} x_{il} \right)}_{\Sigma_{kl}} \sum_{l=1}^{p} u_{l1}
\end{aligned}
$$

If we think in $\hat{u}_1$ as a column matrix, the last is a matrix multiplication $(b) = \hat{u}_1^t \Sigma \hat{u}_1$.

What is Σ?

Let's see the covariance matrix (between the p dots in the map):

$$\begin{aligned}\Sigma_{kl} &= \sum_{i=1}^{n} \frac{1}{n} \left(x_{ik} - \frac{1}{n} \sum_{j=1}^{n} x_{jk} \right) \left(x_{il} - \frac{1}{n} \sum_{j=1}^{n} x_{jl} \right) \\ &= \sum_{i=1}^{n} \frac{1}{n} x_{ik} x_{il}\end{aligned}$$

This is, working with zero-averaged data (and only in this case), allow us to set $\Sigma = \frac{1}{n} X^t X = \mathrm{Cov}(\vec{X})$ as the covariance matrix of our data.

MAR

Minimization problem

Finally, we have got the next result

$$\epsilon_1(\hat{u}_1) = \mathrm{Var}(\vec{X}) - \hat{u}_1^t \mathrm{Cov}(\vec{X})\hat{u}_1.$$

So the problem now is

Optimization problem

Maximize $\hat{u}_1^t \mathrm{Cov}(\vec{X})\hat{u}_1$ with $\hat{u}_1^t\hat{u}_1 = 1$.

To solve this problem is equivalent to the eigenvalue problem

Eigenvalue problem

Find $\hat{u}_1$ such as $\mathrm{Cov}(\vec{X})\hat{u}_1 = \lambda_1\hat{u}_1$.

Error got

After solving the eigenvalue problem, the error we got in the approximation is

$$\begin{aligned} \epsilon_1 &= \mathrm{Var}(\vec{X}) - \hat{u}_1^t \mathrm{Cov}(\vec{X}) \hat{u}_1 \\ &= \mathrm{Var}(\vec{X}) - \hat{u}_1^t \lambda_1 \hat{u}_1 \\ &= \mathrm{Var}(\vec{X}) - \lambda_1 \end{aligned}$$

So as bigger the eigenvalue, smaller the error, and more amount of variance explained by the eigenvector (EOF) associated.

Others EOF

Let's retain now two EOFs

$$\vec{X}_i \simeq (\vec{X}_i \cdot \hat{u}_1)\hat{u}_1 + (\vec{X}_i \cdot \hat{u}_2)\hat{u}_2, \quad i = 1, 2, ..., n$$

in this case, it is possible to prove that the error now becomes

$$\epsilon_2 = \mathrm{Var}(\vec{X}) - \lambda_1 - \lambda_2,$$

and so on.
Finally, if we consider all eigenvalues

Total variance as sum of variances of every EOF

$$\mathrm{Var}(\vec{X}) = \sum_{i=1}^{p} \lambda_i.$$

Strategy

1. Remove the (temporal) average in each dot in the map.
2. Diagonalice the covariance matrix.
3. Sort the eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$.
4. Retain only the first m eigenvalues, which are the EOFs.
5. The explained percentage of variance is $\frac{\lambda_1+\lambda_2+...+\lambda_m}{\lambda_1+\lambda_2+...+\lambda_p}$ % with $m \ll p$.

The point here is to retain only some EOFs (as few as possible), in order to explain a big *enough* part of variance.

MAR

Spacial and temporal covariance matrix

Up to now, we have seen that the key here is to solve the eigenvalue problem with the spatial covariance matrix,

$$\Sigma^s_{kl} = \sum_{i=1}^{n} \frac{1}{n} x_{ik} x_{il} \quad k, l = 1, 2, .., p.$$

In general, this is a quite big matrix.

It would be better to work with the temporal covariance matrix (because this is quite smaller)

$$\Sigma^t_{ij} = \sum_{k=1}^{p} \frac{1}{p} x_{ik} x_{jk} \quad i, j = 1, 2, .., n.$$

Spacial vs. temporal covariance matrix

Or in the matrix flavour

s covariance matrix

$$\Sigma^s = \frac{1}{n} X^t X$$

t covariance matrix

$$\Sigma^t = \frac{1}{p} X X^t$$

But, what is the relation (if any) between Σ^s and Σ^t?: eigenvalues and eigenvectors are related.

Spacial vs. temporal covariance matrix

It can be proved that if two matrix are proportional, then

1. their eigenvalues are as well
2. their eigenvectors are the same

So from now on, we will forget about the constants $1/n$ and $1/p$, and we will work with X^tX and XX^t for s and t modes, respectively.

But, again, what is the relation between X^tX and XX^t?

A useful feature

- XX^t is a $n \times n$ matrix. It has a characteristic polynomial ($|XX^t - \lambda I| = 0$) of degree n which we will call $f(\lambda) = 0$.
- In the other hand, X^tX is a $p \times p$ matrix. It has a characteristic polynomial ($|XX^t - \lambda I| = 0$) of degree p, $g(\lambda) = 0$.

The great thing is that it can be proved than then we have the next equation

Theorem

$$g(\lambda) = f(\lambda)\lambda^{p-n}.$$

A useful feature

This have many meanings, but in particular that the **first** n eigenvalues are equal, and the next $p - n$ are zero.

Being zero, they explain no variance. In this way, we can retain only the first n in the development (and now is exact, not an approximation):

$$\vec{X}_i = \sum_{j=1}^{p} \left(\vec{X}_i \cdot \hat{u}_j \right) \hat{u}_j = \sum_{j=1}^{n} \left(\vec{X}_i \cdot \hat{u}_j \right) \hat{u}_j .$$

And moreover, there is a simple connection between eigenvectors.

A useful feature

Let be $\hat{u}$ a eigenvector of XX^t with non-zero eigenvalue λ. This is,

$$XX^t\hat{u} = \lambda\hat{u}.$$

If we multiply by X^t by the left, we get

$$X^tX\left(X^t\hat{u}\right) = \lambda\left(X^t\hat{u}\right)$$

This is, $X^t\hat{u}$ is eigenvector of X^tX.

A useful feature

So we have proved the next theorem, which relate the eigenvalues and eigenvectors of the s and t modes:

Theorem

Let be $\Sigma^s = \frac{1}{n} X X^t$ and $\Sigma^t = \frac{1}{p} X^t X$ the s and t mode covariance matrix, respectively (and we will also suppose $p > n$). Then, the number of non-zero eigenvalues of Σ^s is equal or smaller to n. Moreover, if $\hat{u}$ is eigenvector of Σ^t, then $X^t \hat{u}$ is also eigenvector of Σ^s.

Strategy

So if we are worried (and we are) about to work with small matrix, we have to follow these steps:

1. If the number of time steps is bigger than the dimension of maps, we will diagonalize Σ^s, after removing the time average in each dot.
2. Otherwise, which is the most regular in climatology, we will diagonalize Σ^t, after removing the spatial average in each dot.
3. The relative weight of each eigenvalue will inform us about the percentage of variance explained by the associated eigenvector.
4. We get the EOFs by just the matrix multiplication $X^t \hat{u}$.

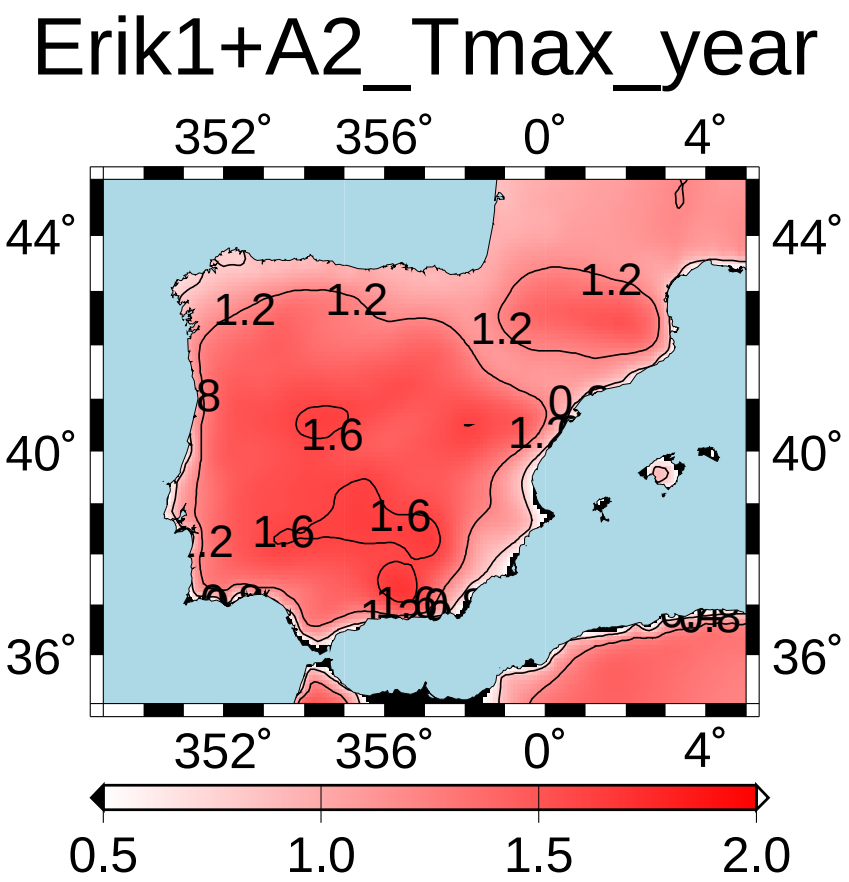
Index

Stating the problem

- We have a series of max daily temperatures averaged over one month (v.g. January) for 199 years.
- Each map has a resolution of 1512 points.
- So in this case $n = 199$, $p = 1512$.
- The problem is, *how can we understand this huge amount of data?*
- We performed an EOF analysis.

Spatial pattern

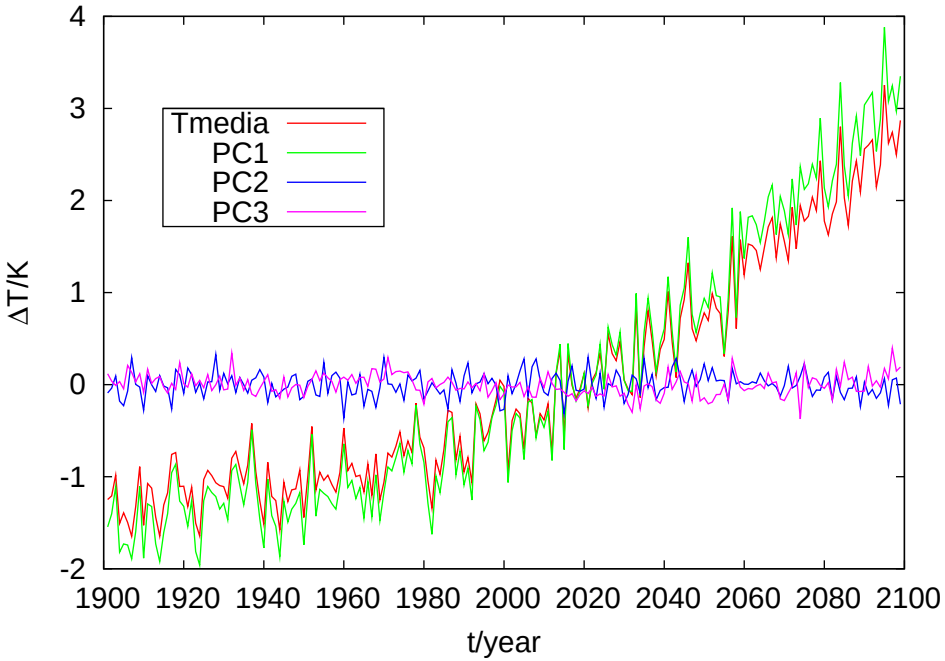
Here we can see the spatial pattern



- It explains more than 80 % of the total variance.
- It tell us the main spatial pattern of warming.

Trend

In the other hand, the temporal behaviour is in the associated to PCs



- First PC follows almost exactly the mean temperature.
- Other PCs are not related (at least directly) with warming.

MAR

Index

Advantages of this tool

- We have reduced strongly the amount of data (just one spatial pattern and one time series).
- We have divided the information into spatial and temporal behaviour.
- We are able to get an easy physical interpretation of the results.

Disadvantages of this tool

- This only works propertly with highly correlated data (very redundant).
- Sometimes is not so easy to get a physical interpretation.
- The EOFs are, by construction, orthogonals, but the physical system may not be $\longrightarrow$ Rotated EOFs.
- But hopefully for you, I'm not going to talk about them...

Thank you

and...

that's all folks