ATM/OCN/ESS 589
Principle Component Analysis
May 25, 2009

May 25, 2009

Principle Component Analysis (PCA) can do many things for you (and to you, if you are not careful!). In this class, we will demonstrate how and when PCA can be used to find order in a data set, mainly by reducing the dimensionality of the data. PCA is a way to represent your data in a very compact form by identifying the most frequently recurring (energetic) spatial structures in the data, and projecting the data onto these structures. PCA is also known as Factor Analysis, Empirical Orthogonal Function (EOF) Analysis, and a host of other names – depending on the discipline you were raised in.

1 Background, Notation, and Matrix Concepts

1.1 Review of Vector Notation

Define two vectors $\mathbf{a}$ and $\mathbf{b}$ made of elements a_i and b_i

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_N \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_N \end{pmatrix} . \tag{1}$$

where we use lower case bold letters to denote vectors. The inner (dot) product of the vectors $\mathbf{a}$ and $\mathbf{b}$ is written:

$$< \mathbf{a}, \mathbf{b} > \; = \sum_{i=1}^{N} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_N b_N \; . \tag{2}$$

In vector notation, this is written

$$< \mathbf{a}, \mathbf{b} > \; \equiv \; \mathbf{a}^T \mathbf{b} \; = \; (a_1, a_2, \cdots a_N) \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_N \end{pmatrix} \; . \tag{3}$$

1.2 Review of Matrix Notation

Now lets say we have a data set $\mathbf{Z}$ with information at M locations in space and N realizations in time. Hence, $\mathbf{Z}$ is a matrix (written in bold capital letters) that is written

$$\mathbf{Z} \quad = \quad \overbrace{\left(\begin{array}{cccc} z_{11} & z_{12} & \cdot & z_{1N} \\ z_{21} & z_{22} & \cdot & z_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ z_{M1} & z_{M2} & \cdot & z_{MN} \end{array} \right)}^{\text{N columns (realizations)}} \Bigg\} \text{M rows (locations)} \; . \tag{4}$$

It is also convenient to express the columns in a matrix as vectors with a subscript to denote it is a vector in a matrix. For the example $\mathbf{Z}$ above, we would write: $\mathbf{z}_j$, where $\mathbf{z}_j$ is the j^{th} vector in the matrix $\mathbf{Z}$.

Each row "m" of $\mathbf{Z}$ is the time series at a unique place in space, z_{mt}. Each column "n" is a value of the data at all places in space at a single time $t = n$, $z_{\mathbf{x}n}$. The $\mathbf{z}_j$ are called the state vectors, and we have N of them.

Lets look at a concrete example. Lets say we have maps of sea surface air temperature over the northern hemisphere for each month, over the period 1948 to 2006. Each map has data averaged over a 10 deg by 10 deg area, so there are $M = 324$ locations (90/10 x 360/10) in a single map, and $N = 708$ maps (12 * 59 years). Thus our temperature matrix $\mathbf{Z}$ has $324x708$ pieces of information. In this case we have N realizations of our "state vector" of air temperature $\mathbf{z}_j$, where each state vector has M measurements at time j.

1.3 Review of Matrix Multiplication

Figure 1 shows how you do the multiplication of matrices.
Here

$$AB_{32} \quad = \quad a_{31}\, b_{12} + a_{32}\, b_{22} + a_{33}\, b_{32} + a_{43}\, b_{42} \; . \tag{5}$$

2 The Covariance Matrix C

Now lets go back and revisit our data matrix and calculate the covariance $\mathbf{C}$ matrix of our data $\mathbf{Z}$:

$$\mathbf{C} \quad \equiv \quad \mathbf{Z}\mathbf{Z}^T /N \quad =$$

$$= \quad \left(\begin{array}{cccc} z_{11} & z_{12} & \cdot & z_{1N} \\ z_{21} & z_{22} & \cdot & z_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ z_{M1} & z_{M2} & \cdot & z_{MN} \end{array} \right) \left(\begin{array}{cccc} z_{11} & z_{21} & \cdot & z_{M1} \\ z_{12} & z_{22} & \cdot & z_{M2} \\ \cdot & \cdot & \cdot & \cdot \\ z_{1N} & z_{2N} & \cdot & z_{MN} \end{array} \right)$$

$$= \quad \left(\begin{array}{cccc} c_{11} & c_{12} & \cdot & c_{1M} \\ c_{21} & c_{22} & \cdot & c_{2M} \\ \cdot & \cdot & \cdot & \cdot \\ c_{M1} & c_{M2} & \cdot & c_{MM} \end{array} \right) \; , \tag{6}$$

where

$$c_{jk} \quad = \quad \frac{1}{N}\sum_{l=1}^{N} z_{jl}\, z_{kl} \quad = \quad c_{kj} \ . \tag{7}$$

So long as we have subtracted the time mean from each record, Eq 7 expresses the covariance between the temperature at $x = j$ and $x = l$. Hence, $\mathbf{C}$ is the covariance matrix and a diagonal element c_{kk} of $\mathbf{C}$ is the variance at each point k. So the total variance in the data is

$$\sum_{m=1}^{M} C_{mm} \quad = \quad \text{trace}(\mathbf{C}) \ . \tag{8}$$

[The "trace" is the sum of the diagonal elements.]

3 Principle Component Analysis

3.1 Eigen Analaysis of the Covariance Matrix C

If there is a lot of structure (covariance) in the data such that much of the information at different places in space is linearly related, then it is useful to find a new, smaller set of state vectors that contains most of the variance and covariance information in the data. This is done by doing an eigen analysis of the covariance matrix $\mathbf{C}$. First, we decompose $\mathbf{C}$ into its M eigenvectors $\mathbf{e}_j$ and eigenvalues λ_j:

$$\mathbf{C}\,\mathbf{e}_j \quad = \quad \lambda_j \mathbf{I}\,\mathbf{e}_j \ , \tag{9}$$

or

$$(\mathbf{C} \quad - \quad \lambda_j \mathbf{I})\,\mathbf{e}_j \quad = \quad 0 \ , \tag{10}$$

where $\mathbf{I}$ is the identity matrix of one on diagonal and zero everywhere else. Importantly, since $\mathbf{C}$ is symmetric, the eigenvectors $\mathbf{e}_j$of $\mathbf{C}$ are orthogonal to each other:

$$\mathbf{e}_k{}^T\,\mathbf{e}_j \quad = \quad \delta_{jk} \ , \tag{11}$$

where δ_{jk} is the Kronecker delta (who was Leopold Kronecker anyway?):

$$\delta_{jk} \quad = \quad \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases} \ , \tag{12}$$

3.2 The Empirical Orthogonal Functions

Since the eigenvectors of the covariance matrix are orthogonal, we can use the eigenvectors as a set of basis functions to re-express our data set in terms of each eigenfunction. This is completely analogous to expressing your data in terms of Fourier coefficients, only now the orthogonal basis functions are informed by the covariance in the data, rather than by sines and cosines.

The eigenvectors $\mathbf{e}_j$ of the covariance matrix $\mathbf{C}$ are known as the *Empirical Orthogonal Functions*, or EOFs, because they are a basis set that is determined

from data rather from the physics (the eigenvectors of the physical system are are called often called "normal modes" of the system. The EOFs and normal modes are almost never the same thing; we will return to this in section 5).

In addition to being orthogonal, the eignenvectors have the wonderful property:

$$\sum_{j=1}^{M} \lambda_j \quad = \quad \sum_{j=1}^{M} \mathbf{C}_{mm} \quad = \quad \text{trace}(\mathbf{C}) \quad = \quad \text{total variance in } \mathbf{Z} \; . \tag{13}$$

Since the eigenvectors are orthogonal, the fraction of variance in the entire data set that is explained by eigenvector (EOF) j is

$$\lambda_j \Big/ \sum_{j=1}^{M} \lambda_j \; . \tag{14}$$

So, if we order the eigenvectors in such a way that the first eigenvector has the largest eignvalue, then the first eigenvector explains the largest fraction of total variance. If there is a lot of structure (covariance) in the data, it will take only a few eigenvectors to explain most of the data (imagine if we could reduce our 324 grid boxes to, say, two maps that explain most of the variance in all 324 of the grid boxes!)

3.3 The Principle Components, PCs

The principle components $\mathbf{P}$ are the time series of each of the eigenvectors that, when added together, reconstitute the original data. Hence,

$$z_{mk} \quad = \quad \sum_{j=1}^{M} \mathbf{e}_{mj} \, \mathbf{P}_{jk} \qquad k = 1, 2, 3, \cdots N \; , \tag{15}$$

Hence, the PCs are found by projecting the eigenvectors (EOFs) onto the data:

$$\mathbf{P}_{jk} \quad = \quad \mathbf{e}_j{}^T \, \mathbf{z}_k \qquad \begin{array}{l} j = 1, 2, 3, \cdots M \\ k = 1, 2, 3, \cdots N \end{array} \; . \tag{16}$$

As a specific example, the amplitude of the second eigenvector (EOF #2) at time $k = 4$ is

$$p_{24} \quad = \quad e_{12} \, z_{14} \, + \, e_{22} \, z_{24} \, + \, \cdots + \, e_{M2} \, z_{M4} \; . \tag{17}$$

We can write this more compactly as

$$\mathbf{Z} \quad = \quad \mathbf{E} \, \mathbf{P} \; , \tag{18}$$

or

$$\mathbf{P} \quad = \quad \mathbf{E}^T \, \mathbf{Z} \; , \tag{19}$$

where

$$\mathbf{E} \quad = \quad \begin{pmatrix} \text{eigenvector 1} & \text{eigenvector 2} & \cdots & \text{eigenvector M} \end{pmatrix} \tag{20}$$

and $\mathbf{P}$ is the matrix of PCs that express the time series of the amplitude of each eigenvector (EOF), one row per eigenvector:

$$\mathbf{P} \quad = \quad \overbrace{\begin{pmatrix} \leftarrow & \mathrm{PC}_1 & \rightarrow \\ \leftarrow & \mathrm{PC}_2 & \rightarrow \\ & \cdots\cdots\cdots & \\ \leftarrow & \mathrm{PC}_M & \rightarrow \end{pmatrix}}^{\text{N times}} . \tag{21}$$

[Note: one can also show that the PCs are orthogonal (try it!).]

3.4 Summary of the EOF/PC Machinery

The eigenanalysis of the covariance matrix gives new functions that allow us to re-express our data.

- The eigenvectors, or EOFs, are orthogonal in space.
- The time series of the EOFs, called the PCs, are also orthogonal to one another, and they tell us the time evolution of the EOFs.
- The eigenvectors and eigenvalues contain all of the variance and covariance in the original data.
- When there is a lot of shared variance between spatial points (i.e., a lot of covariance), most of the variance and structure in the data can be expressed in terms of only a few eigenvectors (EOFs) and their time series. Guidelines for how many modes you retain are discussed in section 4.1.

It is very important to understand the strengths and weaknesses of the EOF analysis, and these are strongly dependent on the physical system that you are analyzing. The next section should help illustrate some of the issues that the analysis must take into consideration.

4 How many modes should we retain, and how should we interpret them?

So now you have a compact representation of your data set in terms of a fewer number of EOFs and their associated PCs. How many should you keep, and when do they have physical meaning? The answer to these questions are not easy, and it helps to have some experience and an *a priori* understanding of the underlying dynamics of the system you are analyzing.

4.1 How many modes should we retain?

This depends on whether you are using EOFs as a filter on your data set, or as a way to identify or isolate special modes of variability. In either case, it is an art form.

4.1.1 Using EOF Analysis as a filter of the data

A typical eigenmode spectrum is shown in Fig 2, ordered from most variance explained to least variance explained. The first handful of modes explain most of the variance (and covariance) in the entire data set, so if you keep these you retain most of the information. The common *assumption* is that all of the remaining modes (which together explain only a small fraction of the total information in your data) is not interesting, or perhaps is even unwanted instrumental noise that you don't want anyway. If your guess right, the stuff you throw away is just uninteresting signal (noise) or instrument error. In this case, using the smaller set of EOFs and PCs to reconstitute your data essentially a filter to remove uninteresting signals or observational error. So where do you draw the line?

Some have advocated keeping enough EOFs/PCs so that you retain 90% of the information in the data set: that is, ordering the eigenvalues from largest to smallest, keep enough to explain 90% of the variance in the data set (determined in Eq. 14). Others have argued that you should look for a break in the spectrum in Fig. 2 and keep only those modes that lay to the left of the break. The logic here is that, if you default model for the system is an AR(1) model (a temporal spectrum that is "red"), then a plot $log\lambda_i$ vs λ_i will have a negative linear slope. If the goal is filtering the data, the overwhelmingly popular choice is the former: keeping enough eigenvectors to explain the overwhelming bulk of the covariance in the data.

4.1.2 Using EOFs to ascertain or isolate special patterns of variability

You might also want to analyze the space and time structure of the leading modes (patterns) *if* you think they have some special physical meaning. Determining whether the modes have special physical meaning can be a bit tricky, however, and requires some experience (and an *a priori* knowledge of the underlying physics).

Rules of thumb to determine which modes *might* be physically meaningful have been developed by several investigators. A very popular one is due to North et al. (*Mon. Wea. Rev.*, **110**, p699-706, 1982), which says that the eigenvectors are distinct from one another when the eigenvalues are well-separated from each other. North et al. derived the following equation that describes the uncertainty in each eigenvalue:

$$\Delta\lambda \quad = \quad \lambda\sqrt{2/N^*}\,, \tag{22}$$

where N^* are the effective (temporal) degrees of freedom in the data set (see

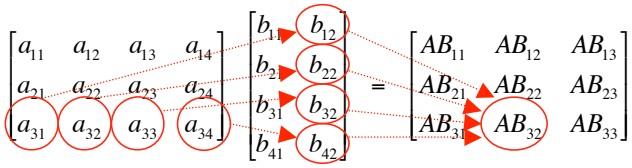

Figure 1: An example of how to do matrix multiplication. The product of the 3 x 4 matrix **A** and 4 x 2 **B** is a 3 x 4 matrix **AB**. Element $AB_{32} = a_{31}\, b_{12} + a_{32}\, b_{22} + a_{33}\, b_{32} + a_{43}\, b_{42}$.

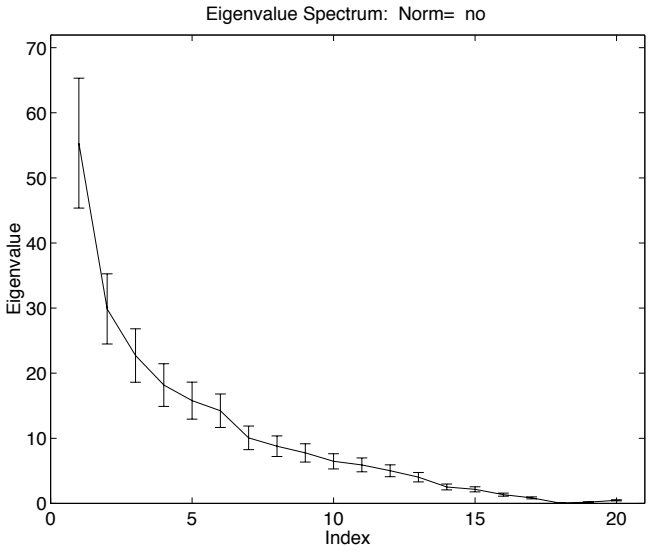

Figure 2: An example eigenspectrum (aka a scree diagram). Plotted along the abcissa is the eigenvalue number, ordered from largest to smallest. Plotted along the ordinate is the eigenvalue. The confidence limits on the eigenvalues are denoted by the whiskers. In this case, only the first eigenvalue/vector is distinct. From D.L. Hartmann.

the notes from the lecture on spectral analysis for a discussion of N^*). North et al. recommend that one not focus on a particular eigenvector/value if there is overlap with another eigenvalue.

In addition, if modes are identified *a posteriori* as being special, they must also exceed in amplitude the default model. If the default model is an AR(1) process, then the constraint on the eigenspectrum discussed in section 4.1.1 also holds.

5 Caveats and Limitations of EOF Analysis

Here is a partial list of caveats and concerns of the EOF Analysis.

- The method tends to favor places where variance is large. So, for example, if you think winds (currents) *u*are the dynamically interesting quantities (or the state variables that are simply the most interest), you would get biased results if you did the EOF analysis of the geopotential (dynamic height) Φ because the geopotential is weighted by the Coriolis parameter f:

$$\mathbf{u} \quad = \quad \mathbf{k} \, \mathbf{x} \, \frac{\nabla \Phi}{f} \; . \tag{23}$$

  Note that spatial weighting between winds and geopotential or vorticity is also problematic because of the spatial weighting by wavenumber (ditto for currents and dynamic height).

- The technique works best when data across space is linearly related (because the eigenvalue decomposition is a linear decomposition of the covariance matrix). When there are nonlinear relationships in space (which is almost always the case), you have to be very careful when you assign physical meaning to the eigenvectors.

  In a paleo context, analogous troubles arise if the proxy index is not linearly related to the climate variable that you are reconstructing.

- Since all of the variance and covariance is contained within the eigenvectors, the EOFs tend to have large spatial structures. Since, in the atmosphere and ocean, large spatial structures tend to also be lower frequency phenomenon, the EOFs will tend to emphasize large scale, lower frequency phenomenon.

- **IMPORTANT.** When the eigenvalues are not well separated, the eigenanalysis often will scramble information between the modes, and one should be very cautious about interpreting these modes as physically. In fact, in general, don't try to interpret them physically.

  An example of such a problem can be seen using the supplied Matlab program.

- When are the EOFs true physical modes? Lets define a true physical mode as the solution to the linear equation:

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\,\mathbf{x}\,, \tag{24}$$

where $\mathbf{x}$ is the state vector $\mathbf{M}$ is a matrix that contains the physics and thermodynamics. The eigenvectors $\mathbf{f}_j$ and eigenvalues of $\mathbf{M}$ are then the solution to this equation. If $\mathbf{M}$ is not Hermitian, however, then the eigenvectors $\mathbf{f}_j$ are not orthogonal and hence there can not be a one-to-one relationship between the true modes and the EOF modes of the output from this system.

What would make the matrix $\mathbf{M}$ non-Hermitian, and thus destroy a perfect relationship between the dynamical and empirically determined modes of the system? Anything that makes $\mathbf{M}$ not symmetric. For example, sheared mean flow or coupled between the atmosphere and ocean (because they have different Rossby numbers). That is, the EOFs are almost never true modes of the dynamical system. They can be close, however, and so there are times when it is useful and appropriate to think of the two as being nearly synonymous.

6 Presentation of EOF Analyses, and some notes on EOF Analysis using Matlab

Note that Eq 11 uses the convention that is customary in physics and in linear algebra packages (such as Matlab): the eigenvectors are chosen to have unit length. In this case, all of the amplitude information is contained in the PCs. In the environmental science literature (including atmospheric sciences and oceanography), it is customary to have the PCs have unit length and instead place the amplitude information in the eigenvector (the EOF). This convection allows you to look at a map of the first EOF and say, "That is what I would see if there was a typical (1 σ) perturbation in this "mode." A 2 σ event would have the same pattern of anomaly, but twice the amplitude.

6.1 Calculating EOFs using Matlab

If your data are arranged in a matrix $\mathbf{Z}$ that is $N\,x\,M$ matrix, where the N rows are the spatial locations of your data and the M columns are the data at each time step, you are ready to go (just remember to remove the temporal mean from each location).

- *Get the eigenvectors and eigenvalues* To get the eigenvectors and eigenvalues, simply write:

  $[v, sig] = eig(Z)$

The eigenvectors are contained in the matrix v; the column j ($v(:,j)$) contains the j^{th} eigenvector. The eigenvalues are stored along the diagonal in the matrix sig and are ordered in order of increasing amplitude. If, you have M locations, then you will have M eigenvectors and values, and the vector (value) explaining the most variance is found in $v(:,M)$ ($sig(M,M)$).

- *Getting the PCs* Its easy, use Eq. 19. So type:

  $P \quad = \quad v'Z; \;\;,$

  and your PCs will be in the matrix P in the format outlined in Eq 21; hence, the PC for the leading (most variance eigenmode) will be found in $P(M,:)$.

- *To reconstruct the data using only a few modes* If you want to reconstruct the data using only a few modes, then you can do so as follows. Lets say we want to construct the data using the first two modes and the fourth mode. If we have a total of 12 modes, then we want only the modes in positions 9, 11 and 12 (the fourth, second and first modes, respectively). Hence, we can write

  $modes \quad = \quad [9,\ 11,\ 12];$
  $zt \quad = \quad v(:,modes) * P(modes,:);$

  And zt is now the sum of the product of $\mathbf{e}_j\ PC_j$ for each of these modes, $j = 12, 11, 9$.

- *Normalizing PC #1* You can normalize the PC by calculating the variance in the PC (which is the eigenvalue associated with that eigenvector), and then dividing the PC by the square root of the variance. For example, to normalize the leading PC, we would take:

  $p1 \quad = \quad P(M,:);$
  $p1 = p1/sqrt(sig(M,M))$

  which is equivalent to

  $p1 \quad = \quad P(M,:);$
  $p1 = p1/std(p1);$

- *Plotting the dimensional Eigenvector #1* This can be done by writing

  $e1 \quad = \quad v(:,M);$
  $e1 \quad = \quad e1 * sqrt(sig(M,M);$

- *Is another field related to the one you have just analyzed?* Lets say we have a separate data set $\mathbf{S}(x_j,t)$. We can find out if this data set is related to the leading eigenvector $\mathbf{e}_M$ of our data $\mathbf{Z}$ by correlating each time series in this new data $\mathbf{S}$ with our $PC\#1 \quad = \quad p1$:

  $for\ i = 1 : M$

F(i)= corrcoef(S(i,:), p1);
end

where the correlation coefficients$F(x_j)$ can be mapped to see how and where the variable S is related to the leading pattern of variability.

If you have normalize the PC to each have unit variance, then you can regress any data onto it and interpret the resulting spatial map as the "map that accompanies a typical (1σ) excursion of the phenomenon captured in EOF/PC#1."

$for\ i = 1 : M$
$R(i) \quad = \quad S(i,:) * p1'/N;$
end

where the $R(x_j)$ are the regression coefficients, which can also be mapped.