4. Matrix Methods for Analysis of Structure in Data Sets:

Empirical Orthogonal Functions, Principal Component Analysis, Singular Value Decomposition, Maximum Covariance Analysis, Canonical Correlation Analysis, Etc.

In this chapter we discuss the use of matrix methods from linear algebra, primarily as a means of searching for structure in data sets.

Empirical Orthogonal Function (EOF) analysis seeks structures that explain the maximum amount of variance in a two dimensional data set. One dimension in the data set represents the dimension in which we are seeking to find structure, and the other dimension represents the dimension in which realizations of this structure are sampled. In seeking characteristic spatial structures that vary with time, for example, we would use space as the structure dimension and time as the sampling dimension. The analysis produces a set of structures in the first dimension, which we call the EOF's, and which we can think of as being the structures in the spatial dimension. The complementary set of structures in the sampling dimension (e.g. time) we can call the Principal Components (PC's), and they are related one-to-one to the EOF's . Both sets of structures are orthogonal in their own dimension. Sometimes it is helpful to sacrifice one or both of these orthogonalities to produce more compact or physically appealing structures, a process called rotation of EOF's.

Singular Value Decomposition (SVD) is a general decomposition of a matrix. It can be used on data matrices to find both the EOF's and PC's simultaneously. In SVD analysis we often speak of the left singular vectors and the right singular vectors, which are analogous in most ways to the empirical orthogonal functions and the corresponding principal components. If SVD is applied to the covariance matrix between two data sets, then it picks out structures in each data set that are best correlated with structures in the other data set. They are structures that 'explain' the maximum amount of covariance between two data sets in a similar way that EOF's and PC's are the structures that explain the most variance in a data set. It is reasonable to call this Maximum Covariance Analysis (MCA).

Canonical Correlation Analysis (CCA) is a combination of EOF and MCA analysis. The two input fields are first expressed in terms of EOF's, the time series of PC's of these structures are then normalized, a subset of the EOF/PC pairs that explain the most variance is selected, and then the covariance(or correlation) matrix of the PC's is subjected to SVD analysis. So CCA is MCA of a covariance matrix of a truncated set of PC's. The idea here is that the noise is first reduced by doing the EOF analysis and so including only the coherent structures in two or more data sets. Then the time series of the amplitudes of these EOFs are normalized to unit variance, so that all count the same, regardless of amplitude explained or the units in which they are expressed. These time series of normalized PCs are then subjected to MCA analysis to see which fields are related.

4.1 Data Sets as Two-Dimensional Matrices

Imagine that you have a data set that is two-dimensional. The easiest example to imagine is a data set that consists of observations of several variables at one instant of time, but includes many realizations of these variable values taken at different times. The variables might be temperature and salinity at one point in the ocean taken every day for a year. Then you would have a data matrix that is 2 by 365; 2 variables measured 365 times. So one dimension is the variable and the other dimension is time. Another example might be measurements of the concentrations of 12 chemical species at 10 locations in the atmosphere. Then you would have a data matrix that is 12x10 (or 10x12). One can imagine several possible generic types of data matrices.

a) A space-time array: Measurements of a single variable at *M* locations taken at *N* different times, where *M* and *N* are integers.

b) A parameter-time array: Measurements of *M* variables (e.g. temperature, pressure, relative humidity, rainfall, . . .) taken at one location at N times.

c) A parameter-space array: Measurements of *M* variables taken at *N* different locations at a single time.

You might imagine still other possibilities. If your data set is inherently three dimensional, then you can string two variables along one axis and reduce the data set to two dimensions. For example: if you have observations at *L* longitudes and *K* latitudes and *N* times, you can make the spatial structure into a big vector *L*x*K*=*M* long, and then analyze the resulting (*L*x*K*)xN=*M*x*N* data matrix. (A vector is a matrix where one dimension is of length 1, e.g. an 1x*N* matrix is a vector).

So we can visualize a two-dimensional data matrix **X** as follows:

$$\mathbf{X} = \begin{array}{c} \\ M \end{array} \overset{N}{\left[\quad \right]} = X_{i,j} \ \textit{where}\ i = 1, M;\ j = 1, N$$

Where *M* and *N* are the dimensions of the data matrix enclosed by the square brackets, and we have included the symbolic bold **X** to indicate a matrix, the graphical box that is *MxN* to indicate the same matrix, and finally the subscript notation $X_{i,j}$ to indicate the same matrix. We define the transpose of the matrix by reversing the order of the indices to make it an NxM matrix.

$$\mathbf{X}^T = \overset{M}{N\left[\quad\right]} = X_{j,i} \text{ where } i = 1, M;\ j = 1, N$$

In multiplying a matrix times itself we generally need to transpose it once to form an inner product, which results in two possible "dispersion" matrices.

$$\mathbf{X}\mathbf{X}^T = \overset{N}{M\left[\quad\right]}\overset{M}{\left[\quad\right]}N = \overset{M}{\left[\quad\right]}M$$

Of course, in this multiplication, each element of the first row of **X** is multiplied times the corresponding element of the first column of $\mathbf{X}^T$, and the sum of these products becomes the first (first row, first column) element of $\mathbf{X}\mathbf{X}^T$. And so it goes on down the line for the other elements. I am just explaining matrix multiplication for those who may be rusty on this. So the dimension that you sum over, in this case *N*, disappears and we get an *MxM* product matrix. In this projection of a matrix onto itself, one of the dimensions gets removed and we are left with a measure of the dispersion of the structure with itself across the removed dimension (or the sampling dimension). If the sampling dimension is time, then the resulting dispersion matrix is the matrix of the covariance of the spatial locations with each other, as determined by their variations in time. One can also compute the other dispersion matrix in which the roles of the structure and sampling variables are reversed.

$$\mathbf{X}^T\mathbf{X} = \overset{M}{N\left[\quad\right]}\overset{N}{\left[\quad\right]}M = \overset{N}{\left[\quad\right]}N$$

Both of the dispersion matrices obtained by taking inner products of a data matrix with itself are symmetric matrices. They are in fact covariance matrices. In the second case the covariance at different times is obtained by projecting on the sample of different spatial points. Either of these dispersion matrices may be scientifically meaningful, depending on the problem under consideration.

EOF (or PCA) analysis consists of an eigenvalue analysis of these dispersion matrices. Any symmetric matrix **C** can be decomposed in the following way through a diagonalization, or eigenanalysis.

Or,

$$\mathbf{C}\mathbf{e}_i = \lambda_i \mathbf{e}_i \tag{4.1}$$

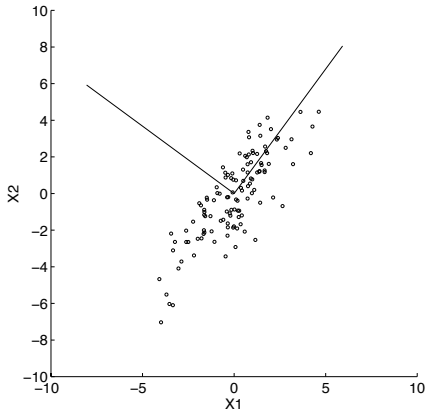
$$\mathbf{C}\mathbf{E} = \mathbf{E}\Lambda \tag{4.2}$$

Where **E** is the matrix with the eigenvectors $\mathbf{e}_i$ as its columns, and $\mathbf{\Lambda}$ is the matrix with the eigenvalues λ_i, along its diagonal and zeros elsewhere.

The set of eigenvectors, $\mathbf{e}_i$, and associated eigenvalues, λ_i, represent a coordinate transformation into a coordinate space where the matrix **C** becomes diagonal. Because the covariance matrix is diagonal in this new coordinate space, the variations in these new directions are uncorrelated with each other, at least for the sample that has been used to construct the original covariance matrix. The eigenvectors define directions in the initial coordinate space along which the maximum possible variance can be explained, and in which variance in one direction is orthogonal to the variance explained by other directions defined by the other eigenvectors. The eigenvalues indicate how much variance is explained by each eigenvector. If you arrange the eigenvector/eigenvalue pairs with the biggest eigenvalues first, then you may be able to explain a large amount of the variance in the original data set with relative few coordinate directions, or characteristic structures in the structure space. A derivation showing how a desire to explain lots of variance with few structures leads to this eigenvalue problem is given in the Section 4.3.

Two-Dimensional Example:

It is simplest to visualize EOFs in two-dimensions as a coordinate rotation that maximizes the efficiency with which variance is explained. Consider the following scatter plot of paired data (x_1, x_2). The eigenvectors are shown as lines in this plot. The first one points down the axis of the most variability, and the second is orthogonal to it.



4.2 EOF/Principal Component Analysis - Introduction

In this section we will talk about what is called Empirical Orthogonal Function (EOF), Principle Component Analysis (PCA), or factor analysis, depending on the tradition in the discipline of interest. EOF analysis follows naturally from the preceding discussion of regression analysis and linear modeling, where we found that correlations between the predictors causes them to be redundant with each other and causes the regression equations involving them to perform poorly on independent data. EOF analysis allows a set of predictors to be rearranged into a new set of predictors that are orthogonal with each other and which maximizes the amount of variance in the dependent sample that can be explained with a the smallest number of EOF predictors. It was in this context that Lorenz (1956) introduced EOF's into the meteorological literature. The same mathematical tools are used in many other disciplines, under a variety of different names.

In addition to providing better predictors for statistical forecasting, EOF analysis can be used to explore the structure of the variability within a data set in an objective way, and to analyze relationships within a set of variables. Examples include searching for characteristic spatial structures of disturbances and for characteristic relations between parameters. The relationships between parameters may be of scientific interest in themselves, quite apart from their effect on statistical forecasting. The physical interpretation of EOFs is tricky, however. They are constructed from mathematical constraints, and may not have any particular physical significance. No clear-cut rules are available for determining when EOFs correspond to physical entities, and their interpretation always requires judgment based on physical facts or intuition.

EOF's and PC's in Forecasting

Suppose we wish to predict y given x_1, x_2, x_3, ...x_M, but we know that the x_i's are probably correlated with each other. It is possible and desirable to first determine some new predictors z_i, which are linear combinations of the x_i.

$$\begin{aligned} z_1 &= e_{11}x_1 + e_{12}x_2 + e_{13}x_3 + ... + e_{1M}x_M \\ z_2 &= e_{21}x_1 + e_{22}x_2 + e_{23}x_3 + ... + e_{2M}x_M \\ .. &= ..... \quad ..... \quad ..... \quad ... \quad ..... \\ z_M &= e_{M1}x_1 + e_{M2}x_2 + eM3x_3 + ... + e_{MM}x_M \end{aligned} \tag{4.3}$$

that is, $z_i = e_{ij}\, x_j$. The matrix of coefficients e_{ij} rotates the original set of variables into a second set.

It is possible to determine the e_{ij} in such a way that:

1) z_1 explains the maximum possible amount of the variance of the x's; z_2 explains the maximum possible amount of the remaining variance of the x's; and so forth for the remaining z's. The e_{ij} are the set of empirical orthogonal functions, and when we project the data x_j onto the EOF's we obtain the principal components, z_i, which are the expression of the original data in the new coordinate system. The EOF's are spatially orthonormal, that is

$$\overline{e_{ki}e_{ij}} = \begin{cases} 1, \; for\, i = j \\ 0, \; for\, i \neq j \end{cases} \quad \text{or} \quad \mathbf{E}^T\mathbf{E} = \mathbf{I} \tag{4.4}$$

where **I** is the identity matrix.

2) The z's are orthogonal, linearly independent or uncorrelated, over the sample. That is

$$\overline{z_{ik}z_{kj}} = 0 \quad \text{for} \quad i \neq j\,. \tag{4.5}$$

where the overbar indicates an average over the sample or summing over the k index, which in many applications will be an average over time. This property of orthogonality or lack of correlation in time makes the principal components very efficient for statistical forecasting, since no variance is shared between the predictors and the minimum useful correlation is any nonzero one.

4.3 EOFs Derived as Efficient Representations of Data Sets

Following Kutzbach (1967), we can show how the mathematical development of empirical orthogonal functions (EOFs) follows from the desire to find a basis set that explains as much as possible of the variance of a data set with the fewest possible

coordinate directions (unit vectors). This will illustrate that the structure functions we want are the eigenvectors of the covariance matrix. Suppose we have an observation vector $\mathbf{x}_m$, of length M, so that it requires M components to describe the state of the system in question. It can also be termed the state vector. If we have N observations of the state vector $\mathbf{x}_m$ then we can think of an observation matrix $\mathbf{X}=x_{nm}$, whose n columns represent the n observation times and whose m rows represent the m components of the state vector. We want to determine a vector **e**, which has the highest possible resemblance to the ensemble of state vectors. The projection of this unknown vector onto the ensemble of data is measured by the inner product of the vector **e** with the observation matrix **X**. To produce an estimate of the resemblance of **e** to the data that is unbiased by the size of the data set, we must divide by N. To make the measure of resemblance independent of the length of the vector and dependent only on its direction, we should divide by the length of the vector **e**. The measure of the resemblance of **e** to the ensemble of data derived by this line of reasoning is,

$$\left(\mathbf{e}^T\mathbf{X}\right)^2 N^{-1}\left(\mathbf{e}^T\mathbf{e}\right)^{-1} = \mathbf{e}^T\mathbf{X}\mathbf{X}^T\mathbf{e}\, N^{-1}\left(\mathbf{e}^T\mathbf{e}\right)^{-1} \tag{4.6}$$

This is equivalent to maximizing,

$$\mathbf{e}^T\mathbf{C}\,\mathbf{e}\ ;\ \text{subject to}\ \mathbf{e}^T\mathbf{e}=1 \tag{4.7}$$

and where,

$$\mathbf{C}=\mathbf{X}\mathbf{X}^T\ N^{-1} \tag{4.8}$$

is the covariance matrix of the observations (**C** is really only equal to the usual definition of the covariance matrix if we have removed the mean value from our data). The length of the unknown vector **e** is constrained to be equal to one so that only the direction of it can affect its projection on the data set. Otherwise the projection could be made arbitrarily large simply by increasing the length of **e**.

To facilitate a solution to (4.7), suppose that we assume the maximum value of the squared projection we are looking for is λ. It is equal to the variance explained by the vector **e**.

$$\mathbf{e}^T\mathbf{C}\,\mathbf{e}=\lambda \tag{4.9}$$

(4.9) corresponds to the classic eigenvalue problem,

$$\mathbf{C}\,\mathbf{e}=\mathbf{e}\,\lambda \quad \text{or} \quad \left\{\mathbf{C}-\lambda\,\mathbf{I}\right\}\mathbf{e}\ =\ 0 \tag{4.10}$$

This can only be true if,

$$\left|\mathbf{C}-\lambda\mathbf{I}\right|=0 \tag{4.11}$$

Thus λ is an eigenvalue of the covariance matrix **C**. The eigenvalues of a real symmetric matrix are positive, as we expected our λ's to be when we defined them with (4.9). In general there will be *M* of these eigenvalues, as many as there are elements of the state vector **x** unless the matrix is degenerate (in which case there are only *r* nonzero eigenvalues, where *r* is the rank of the matrix). From (4.9) we can see that each of these λ_j is equal to the variance explained by the corresponding e_j. To find the principal component, or empirical orthogonal function, e_j, we can use any number of standard techniques to solve the system,

$$\left(\mathbf{C} - \lambda_j \mathbf{I}\right) e_j = 0 \tag{4.12}$$

which is equivalent to the standard linear algebra problem of diagonalizing the matrix **R**.

It is easy to show that the eigenvectors, e_j, are orthogonal. Suppose we have two eigenvectors e_j and e_k. From (4.3) we know that,

$$\mathbf{C}\, e_j = \lambda_j e_j \qquad\qquad \mathbf{C}\, e_k = \lambda_k e_k \tag{4.13}$$

(a) (b)

Multiply (4.13a) by e_k^T and transpose the equation. Multiply (4.13b) by e_j^T. Subtracting the resulting equations from each other yields,

$$e_j^T \mathbf{C}^T e_k - e_j^T \mathbf{C}\, e_k = \left(\lambda_j - \lambda_k\right) e_j^T e_k \tag{4.14}$$

Since the covariance matrix **C** is symmetric, the left-hand side of (4.14) is zero and we have,

$$e_j^T e_k = 0 \quad \text{unless} \quad \lambda_j = \lambda_k \tag{4.15}$$

Therefore the eigenvectors are orthogonal if the eigenvalues are distinct.

We have defined a new basis in the orthonormal eigenvectors e_i that can be used to describe the data set. These eigenvectors are called the *empirical orthogonal functions*; empirical because they are derived from data; orthogonal because they are so.

4.4 Manipulation of EOFs and PCs

Original Space to EOF Space and back again.

It is convenient to order the eigenvalues and eigenvectors in order of decreasing magnitude of the eigenvalue. The first eigenvector thus has the largest λ and explains the largest amount of variance in the data set used to construct the covariance matrix. We need to know how to find the loading vectors that will allow us to express a particular

state in this new basis. Begin by expanding the equation for the eigenvalues to include all the possible eigenvalues. The equation (4.9) becomes,

$$\mathbf{E}^T \mathbf{C} \mathbf{E} = \Lambda \quad \text{or} \quad \mathbf{E}^T \mathbf{X}\mathbf{X}^T \mathbf{E} = \Lambda N \tag{4.16}$$

Where **E** is the matrix whose columns are the eigenvectors e_i and **L** is a square matrix with the M eigenvalues down the diagonal and all other elements zero. If **C** is MxM and has M linearly independent eigenvectors, then the standard diagonalization (4.16) is always possible.

If we define

$$\mathbf{Z} = \mathbf{E}^T \mathbf{X} \tag{4.17}$$

Then it follows that,

$$\mathbf{X} = \mathbf{E}\mathbf{Z}, \quad \text{since} \quad \mathbf{E}\mathbf{E}^T = \mathbf{I} \tag{4.18}$$

---

where **I** is the identity matrix. Equation (4.18) shows how to express the original data in terms of the eigenvectors, when the coefficient matrix **Z** is defined by (4.17). **Z** contains the principal component vectors, the amplitudes by which you multiply the EOFs to get the original data back. One can go back and forth from the original state vector in the original components to the new representation in the new coordinate system by using the eigenvector matrix, as indicated by the transform pair in (4.17) and 4.18). An individual observation vector $\mathbf{x}_n = X_{in}$ can thus be expressed as

$$\mathbf{x}_n = X_{in} = \sum_{j=1}^{M} E_{ij} Z_{jn} \tag{4.19}$$

Projections 101:

Suppose we have a set of orthogonal and normalized eigenvectors. The first one might look like the following:

$$\begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \\ \cdots \\ e_{M1} \end{bmatrix}$$

Putting all the eigenvectors into the columns of a square matrix, gives me **E**, which

has the following orthonormality property.

$$\mathbf{E}^T\mathbf{E} = \mathbf{I}$$

where **I** is the identity matrix.

If I want to PROJECT a single eigenvector onto the data and get an amplitude of this eigenvector at each time, I do the following, $\mathbf{e}^{\mathrm{T}}\mathbf{X}$

$$\begin{bmatrix} e_{11} & e_{21} & e_{31} & \cdots & e_{M1} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2N} \\ x_{31} & x_{32} & x_{33} & \ldots & x_{3N} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{M1} & x_{M2} & x_{M3} & \ldots & x_{MN} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \ldots & z_{1N} \end{bmatrix}$$

where, for example, $z_{11} = e_{11}x_{11} + e_{21}x_{21} + e_{31}x_{31} + \ldots + e_{M1}x_{M1}$. If we do the same for all the other eigenvectors, we will get time series of length n for each EOF, which we call the principle component (PC) time series for each EOF, **Z**, which is an *N*x*M* matrix.

$$\mathbf{E}^T\mathbf{X} = \mathbf{Z}$$

Orthogonality of the Principal Component Time Series.

The matrix **Z** is the coefficient matrix of the expansion of the data in terms of the eigenvectors, and these numbers are called the principal components. The column

vectors of the matrix are the coefficient vectors of length M for the N observation times(or cases). Substituting (4.18) into (4.16) we obtain,

$$\mathbf{E}^T \mathbf{X}\mathbf{X}^T \mathbf{E} = \mathbf{E}^T \mathbf{E}\mathbf{Z}\mathbf{Z}^T \mathbf{E}^T \mathbf{E} = \mathbf{Z}\mathbf{Z}^T = \Lambda N$$

$$\mathbf{Z}\mathbf{Z}^T = \Lambda N \qquad or \qquad \frac{1}{N}\mathbf{Z}\mathbf{Z}^T = \Lambda \tag{4.20}$$

Thus not only the eigenvectors, but also the PCs are orthogonal. If you like, the N realizations expressed in principal component space are uncorrelated in time. The basis set in principal component space, the e_i's, is an orthogonal basis set, both in the 'spatial' dimension M and the 'temporal' dimension N.

Note that the fraction of the total variance explained by a particular eigenvector is equal to the ratio of that eigenvalue to the trace of the eigenvalue matrix, which is equal to the trace of the covariance matrix. Therefore the fraction of the variance explained by the first k of the M eigenvectors is,

$$V_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \tag{4.21}$$

No other linear combination of k predictors can explain a larger fraction of the variance than the first k principal components. In most applications where principal component analysis is useful, a large fraction of the total variance is accounted for with a relatively small number of eigenvectors. This relatively efficient representation of the variance and the fact that the eigenfunctions are orthogonal, makes principal component analysis an important part of statistical forecasting.

To this point we have assumed that the matrix $\mathbf{C}$ is a covariance matrix and that the λ's corresponded to partial variances. In many applications, however, it is desirable to normalize the variables before beginning so that $\mathbf{C}$ is in fact a correlation matrix and the λ's are squared correlation coefficients, or fractions of explained variance. The decision to standardize variables and work with the correlation matrix or, alternatively, to use the covariance matrix depends upon the circumstances. If the components of the state vector are measured in different units (e.g., weight, height, and GPA) then it is mandatory to use standardized variables. If you are working with the same variable at different points (e.g., a geopotential map), then it may be desirable to retain a variance weighting by using unstandardized variables. The results obtained will be different. In the case of the covariance matrix formulation, the elements of the state vector with larger variances will be weighted more heavily. With the correlation matrix, all elements receive the same weight and only the structure and not the amplitude will influence the principal components.

4.5 EOF Analysis via Singular Vector Decomposition of the Data Matrix

If we take the two-dimensional data matrix of structure (e.g. space) versus sampling (e.g. time) dimension, and do direct singular value decomposition of this matrix, we recover the EOFs, eigenvalues, and normalized PC's directly in one step. If the data set is relatively small, this may be easier than computing the dispersion matrices and doing the eigenanalysis of them. If the sample size is large, it may be computationally more efficient to use the eigenvalue method. Remember first our definition of SVD of a matrix:

Singular Value Decomposition: Any *m* by *n* matrix **X** can be factored into

$$\mathbf{X} = \mathbf{U}\,\Sigma\,\mathbf{V}^T \tag{4.22}$$

where **U** and **V** are orthogonal and **Σ** is diagonal. The columns of **U** (*m* by *m*) are the eigenvectors of $\mathbf{X}\mathbf{X}^T$, and the columns of **V** (*n* by *n*) are the eigenvectors of $\mathbf{X}^\mathbf{T}\mathbf{X}$. The *r* singular values on the diagonal of **Σ** (*m* by *n*) are the square roots of the nonzero eigenvalues of both $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^\mathbf{T}\mathbf{X}$. .

So we suppose that the data matrix **X** is MxN, where M is the space or structure dimension and N is the time or sampling dimension. More generally, we could think of the dimensions as the structure dimension M and the sampling dimension N, but for concreteness let's call them space and time. Now $\mathbf{X}\mathbf{X}^T$ is the dispersion matrix obtained by taking an inner product over time leaving the covariance between spatial points. Thus the eigenvectors of $\mathbf{X}\mathbf{X}^T$ are the spatial eigenvectors, and appear as the columns of **U** in the SVD. Conversely, $\mathbf{X}^T\mathbf{X}$ is the dispersion matrix where the inner product is taken over space and it represents the covariance in time obtained by using space as the sampling dimension. So the columns of **V** are the normalized principal components, that are associated uniquely with each EOF. The columns of **U** and **V** are linked by the singular values, which are down the diagonal of **Σ**. These eigenvalues represent the amplitude explained, however, and not the variance explained, and so are the proportional to the square roots of the eigenvalues that would be obtained by eigenanalysis of the dispersion matrices. The eigenvectors and PC's will have the same structure, regardless of which method is used, however, so long as both are normalized to unit length.

To illustrate the relationship between the singular values of SVD of the data matrix and the eigenvalues of the covariance matrix, consider the following manipulations. Let's assume that we have modified the data matrix X to remove the sample mean from every element of the state vector, so that $\mathbf{X} = \mathbf{X} - \overline{\mathbf{X}}$. The covariance matrix is given by

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T / n \tag{4.23}$$

and the eigenvectors and eigenvectors are defined by the diagonalization of **C**.

$$\mathbf{C} = \mathbf{E}\Lambda\mathbf{E}^\mathbf{T} \tag{4.24}$$

Now if we take the SVD of the data matrix, $X = \mathbf{U}\Sigma\mathbf{V}^\mathbf{T}$, and use it to compute the covariance matrix, we get:

$$\begin{aligned} \mathbf{C} &= \mathbf{U}\Sigma\mathbf{V}^\mathbf{T}(\mathbf{U}\Sigma\mathbf{V}^\mathbf{T})^T / n \\ &= \mathbf{U}\Sigma\mathbf{V}^\mathbf{T}\mathbf{V}\Sigma^T\mathbf{U}^\mathbf{T} / n \\ &= \mathbf{U}\Sigma\Sigma^T\mathbf{U}^\mathbf{T} / n \end{aligned} \tag{4.25}$$

From this we infer that: $\mathbf{U} = \mathbf{E}$, $\Lambda = \Sigma\Sigma^T / n$ *or* $\lambda_i = \sigma_i^{\,2} / n$, so there is a pesky factor of n, the sample size, between the eigenvalues of the covariance matrix, and the singular values of the original data matrix.

Also, from EOF/PC analysis we noted that the principal component time series are obtained from $\mathbf{Z} = \mathbf{E}^T\mathbf{X}$, if we apply this to the singular value decomposition of **X**, we get (see 4.17-4.18),

$$\begin{aligned} \mathbf{Z} = \mathbf{E}^T\mathbf{X} = \mathbf{E}^T\mathbf{U}\Sigma\mathbf{V}^T = \mathbf{E}^T\mathbf{E}\Sigma\mathbf{V}^T = \Sigma\mathbf{V}^T \\ \mathbf{Z} = \Sigma\mathbf{V}^T \end{aligned} \tag{4.26}$$

Notice that as far as the mathematics is concerned, both dimensions of the data set are equivalent. You must choose which dimension of the data matrix contains interesting structure, and which contains sampling variability. In practice, sometimes only one dimension has meaningful structure, and the other is noise. At other times both can have meaningful structure, as with wavelike phenomena, and sometimes there is no meaningful structure in either dimension.

Note that in the eigenanalysis,

$$\mathbf{Z}\mathbf{Z}^T = \mathbf{E}^\mathbf{T}\mathbf{X}\mathbf{X}^T\mathbf{E} = \mathbf{E}^\mathbf{T}\mathbf{C}\mathbf{E}N = \Lambda N, \tag{4.27}$$

whereas in the SVD representation,

$$\mathbf{Z}\mathbf{Z}^T = \Sigma\mathbf{V}^T\mathbf{V}\Sigma^\mathbf{T} = \Sigma\Sigma^\mathbf{T}, \tag{4.28}$$

so we must have that

$$\Sigma\Sigma^\mathbf{T} = \Lambda N, \text{ or } \sigma_k^2 = \lambda_k N. \tag{4.29}$$

4.6 Presentation of EOF Analysis Results

After completing EOF analysis of a data set, we have a set of eigenvectors, or structure functions, which are ordered according to the amount of variance of the original data set that they explain. In addition, we have the principal components, which are the amplitudes of these structure functions at each sampling time. Normally, we only concern ourselves with the first few EOFs, since they are the ones that explain the most variance and are most likely to be scientifically meaningful. The manner in which these are displayed depends on the application at hand. If the EOFs represent spatial structure, then it is logical to map them in the spatial domain as line plots or contour plots, possibly in a map projection that shows their relation to geographical features.

One can plot the EOFs directly in their normalized form, but it is often desirable to present them in a way that indicates how much real amplitude they represent. One way to represent their amplitude is to take the time series of principal components for the spatial structure (EOF) of interest, normalize this time series to unit variance, and then regress it against the original data set. This produces a map with the sign and dimensional amplitude of the field of interest that is explained by the EOF in question. The map has the shape of the EOF, but the amplitude actually corresponds to the amplitude in the real data with which this structure is associated. Thus we get structure and amplitude information in a single plot. If we have other variables, we can regress them all on the PC of one EOF and show the structure of several variables with the correct amplitude relationship, for example, SST and surface vector wind fields can both be regressed on PCs of SST.

How to scale and plot EOF's and PC's:

Let's suppose we have done EOF/PC analysis using either the SVD of the data (described in Section 4.5), or the eigenanalysis of the covariance matrix. We next want to plot the EOF's to show the spatial structure in the data. We would like to combine the spatial structure and some amplitude information in a single plot. One way to do this is to plot the eigenvectors, which are unit vectors, but to scale them to the amplitude in the data set that they represent.

Let's Review:

SVD of Data Matrix Approach:

Before we look at the mathematics of how to do these regressions, let's first review the SVD method of computing the EOFs and PCs.

We start with a data matrix X, with n columns and m rows (nxm), and SVD it, giving the following representation.

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T \tag{4.30}$$

the columns of U are the column space of the matrix, and these correspond to the eigenvectors of EOF analysis. The columns of V are the unit vectors pointing in the same direction and the PC's of EOF analysis. They are the normalized time variability of the amplitudes of the EOFs, the normalized PCs. The diagonal elements of Σ, are the amplitudes corresponding to each EOF/PC pair.

Eigenanalysis of Covariance Matrix:

If we take the product across the sampling dimension of the original data matrix X, we get the covariance matrix. Let's assume that the means in the sampling dimension have been removed, but not necessarily the amplitudes (we have not necessarily divided by the standard deviation in the sampling dimension). The covariance matrix is,

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T / N \tag{4.31}$$

Where we have to divide by the sample size, N=n. We use eigenanalysis of C to get,

$$\mathbf{E}^T \mathbf{X}\mathbf{X}^T \mathbf{E} / N = \mathbf{E}^T \mathbf{C}\mathbf{E} = \Lambda \tag{4.32}$$

where the columns of the **E** matrix are the eigenvectors and the elements of the diagnonal matrix **Λ** are the eigenvalues. The eigenvalues are the squares of the singular values of the SVD analysis of the data matrix.

$$\mathbf{E} = \mathbf{U} \qquad \Lambda = \Sigma^2 / N \tag{4.33}$$

You need to remember that Σ has at most m nonzero elements if m<n.

Now suppose we want to plot the eigenvectors, but with the amplitude corresponding to the amplitude in the data set that they represent. In EOF analysis of the covariance matrix we just multiply the square root of the eigenvalue matrix times the eigenvector matrix, and in the SVD world, the same is accomplished by multiplying the left singular vector (the eigenvalue) matrix by the corresponding matrix.

$$\mathbf{D}^{SVD} = N^{-1/2}\mathbf{U}\Sigma \quad equals \quad \mathbf{D}^{EOF} = \mathbf{E}\Lambda^{1/2} \tag{4.34}$$

In each case you can show that $\mathbf{D}\mathbf{D}^{\mathrm{T}}=\mathbf{C}$, so if you put in the amplitudes and take the inner product you get back the covariance matrix of the input data. The columns of the matrix **D** are the eigenvectors, scaled by the amplitude that they represent in the original data. It might be more interesting to plot **D** than **E**, because then you can see how much amplitude in an RMS sense is associated with each EOF and you can plot the patterns in millibars, ˚C, kg, or whatever units in which the data are given. This really only works if the data that you subject to SVD are dimensional and have not already been

nondimensionalized. Rather we should use the procedure below, which works even if the data have been normalized prior to calculating the eigenvectors by SVD or eigenanalysis of the covariance matrix. The time structures in SVD analysis, **V**, are normalized in a different way from the normalized time series of principle components, as we will note below. Nonetheless, we can do the regression in a variety of ways.

Regression maps for EOF Analysis Based on Normalized Data

Sometimes it is advisable to take the amplitudes out of the data before conducting the EOF analysis. Reasons for this might be 1) the state vector is a combination of things with different units or 2) the variance of the state vector varies from point to point so much that this distorts the patterns in the data. If you are looking for persistent connections between the data, you may want to look at correlation rather than covariance. The data are normalized such that the variance of the time series of each element of the state vector is 1. We define this new normalized data set as $\tilde{\mathbf{X}}$.

In this case we can take the EOFs and project them onto the normalized data to get the normalized principal component time series. We first construct the PCs from the data by projecting the EOFs from the correlation matrix on the normalized data used to calculate them.

$$\mathbf{Z} = \mathbf{E}^T \tilde{\mathbf{X}} \tag{4.35}$$

Even though we have normalized the input data so that each component of the input data has the same variance, this variance gets unevenly distributed over the EOFs, and we have to renormalize the PC time series as below. We can normalize the principal components by dividing through by the standard deviation of the PC time series. In matrix notation, this is especially compact.

$$\tilde{\mathbf{Z}} = \Lambda^{-1/2}\mathbf{Z} \tag{4.36}$$

We can then calculate

$$\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T = \Lambda^{-1/2}\mathbf{Z}\mathbf{Z}^T\Lambda^{-1/2T} = \Lambda^{-1/2}\mathbf{E}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{E}\Lambda^{-1/2T} = \Lambda^{-1/2}\mathbf{E}^T\mathbf{C}\mathbf{E}\Lambda^{-1/2T}N = \Lambda^{-1/2}\Lambda\Lambda^{-1/2T}N = \mathbf{I}N$$

Thus we normalize the principle components by requiring that :

$$\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T = \mathbf{I}N \tag{4.37}$$

Whereas the eigenvectors from the SVD analysis are normalized in the following way,

$$\mathbf{V}\mathbf{V}^T = \mathbf{I} \tag{4.38}$$

The rows of the matrix $\tilde{\mathbf{Z}}$ are now the normalized time series of the individual PCs corresponding to each eigenvector. Now we project the original, unnormalized data, **X**, onto the normalized PCs to construct a set of functions that represent the amplitude of the patterns.

$$\mathbf{D} = \mathbf{X}\tilde{\mathbf{Z}}^T / N \tag{4.39}$$

$$\underset{m}{\underset{k}{\begin{bmatrix} D \end{bmatrix}}} = \underset{m}{\underset{n}{\begin{bmatrix} X \end{bmatrix}}} \underset{n}{\underset{k}{\begin{bmatrix} \tilde{Z}^T \end{bmatrix}}} n / N$$

Here k is the number of nonzero eigenvalues, or the truncation limit at which you wish to stop the expansion, k cannot be greater than the structure dimension m.

The matrix product divided by the length of the sample space is equivalent to a covariance, which is like a regression coefficient (See (3.6) or (3.21), if the predictors are orthonormal, then the regression coefficients are just the covariance of the individual predictors with the predictand). The columns of the matrix **D** are the EOFs with their dimensional units. The **D** calculated this way (4.39) is the eigenvector matrix, except with amplitudes that are equal to the amplitude in the original **X** field that is associated with a one standard deviation variation of the PC time series. In other words, the amplitudes that you explain with this structure. This works whether or not the data are normalized prior to EOF analysis, as long as you use the unnormalized data in (4.39).

Let's note one more interesting thing about **D**.

$$\mathbf{D}\mathbf{D}^T = \mathbf{X}\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}\mathbf{X}^T / N^2 = \mathbf{X}\mathbf{I}N\mathbf{X}^T / N^2 = \mathbf{X}\mathbf{X}^T / N = \mathbf{C} \tag{4.40}$$

So if we take the inner product of this regression matrix with itself, we get back the covariance matrix, **C**, provided that we retain all the nonzero eigenvalues or singular values and their associated EOFs and PCs, along the way.

The structures in **D** give the amplitudes in the original **X** field that are associated with a one standard deviation variation in the principle component time series. This is a reasonable number to look at, since it is the amplitude you might see on a typical day.

Why is this like a regression?

Suppose we want to regress y onto x, $y = a x + \varepsilon$. a has units of y/x, but if we normalized x such that $\overline{x'^2} = \sigma_x = 1$, then the units of a are the units of y, and a is the amplitude of y that can be associated with a one standard deviation variation of x. In linear regression, a involves the covariance between x and y (see (3.6)). In plotting the EOF's we might choose to regress the original **X** data onto normalized time series of the amplitudes, or the normalized PC time series. Using the SVD of the data matrix, we can write

$\mathbf{Z} = \Sigma \mathbf{V}^T$, where $\mathbf{V}\mathbf{V}^T = \mathbf{I}$,

so a formula for the normalized time series of the PC's is

$\tilde{\mathbf{Z}} = \mathbf{V}^T n$,

but our regression coefficient is

$a = \mathbf{X}\tilde{\mathbf{Z}}^T / n$,

so, using $\tilde{\mathbf{Z}} = \mathbf{V}^T n$,

we get finally,

$a = \mathbf{X}\mathbf{V}$, where $\mathbf{X} = U\Sigma\mathbf{V}^T$.

Here only the first r columns of **V** would be used, and probably far fewer than that, because only the first few principal components would be meaningful.

Examples:

We will now do a couple of examples of EOF analysis using Matlab on a very simple data set. We will do EOF/PC analysis first using SVD of the data matrix. Then we will calculate the covariance matrix of the data set and compute the EOFs as the eigenvectors of the covariance matrix.

Here is an example from Matlab, where we input a data matrix a, which has a spatial dimension of 2 and a sampling dimension of 4. Do EOF analysis by SVD method.

```
>> clear
>> a=[2 4 -6 8; 1 2 -3 4]
a =
    2    4   -6    8
    1    2   -3    4
```

Do SVD of that data matrix to find its component parts.

```
>> [u, s, v]=svd(a)
```

First U, which contains the spatial singular vectors as columns.

```
u =
   0.8944   -0.4472
   0.4472    0.8944
```

Then the singular value matrix, which only contains one value. This means the data matrix is singular and one structure function and one temporal function can explain all of the data, so only the first column of the spatial eigenvector matrix is significant. The singular value contains all of the amplitude information. The spatial and temporal singular vectors are both of unit length.

```
s =
  12.2474         0         0         0
       0         0         0         0
```

Finally, the temporal structure matrix. Only the first column is meaningful in this context and it gives the normalized temporal variation of the amplitude of first spatial structure function.

```
v =
   0.1826   -0.1195   -0.9759         0
   0.3651   -0.2390    0.0976   -0.8944
  -0.5477   -0.8367    0.0000    0.0000
   0.7303   -0.4781    0.1952    0.4472
```

We can reconstruct the data matrix by first multiplying the singular value matrix times the transpose of the temporal variation matrix.

```
>> sv=s*v'
sv =
   2.2361    4.4721   -6.7082    8.9443
        0         0         0         0
```

Only the first row of this matrix has nonzero values, because the amplitude of the second structure function is zero. The second spatial structure is the left null space of the data matrix. If you multiply it on the left of the data matrix, it returns zero. The first row of sv is the principal component vector, including the dimensional amplitude. Finally we can recover the data matrix by multiplying the spatial eigenvector matrix times the previous product of the singular value and the temporal structure matrices. This is equivalent to multiplying the eigenvector matrix times the PC matrix, and gives us the original data back.

```
>> A=u*sv
A =
   2.0000    4.0000   -6.0000    8.0000
   1.0000    2.0000   -3.0000    4.0000
```

That was fun! This time let's take the same data matrix and use Matlab to find the EOFs and PCs by using an eigenanalysis of the covariance matrix.

First, enter the data matrix a, as before.

```
>>  clear
>> a=[2 4 -6 8; 1 2 -3 4]
a =
    2    4   -6    8
    1    2   -3    4
```

Now compute the covariance matrix AA^T. We won't even bother to remove the mean or divide by N.

```
>> c=a*a'      (normally, you divide by N here, to make the covariance independent of the sample size.)
c =
  120    60
   60    30
```

Next do the eigenanalysis of the square covariance matrix c.

```
>> [v,d]=eig(c)
v =
  -0.8944    0.4472
  -0.4472   -0.8944
d =
  150     0
    0     0
```

v contains the eigenvectors as columns. We get the same normalized eigenvector in the first column as before, except that its sign is reversed. The sign is arbitrary in linear analysis, so let's just leave it reversed. The one eigenvalue (d) is 150, the total variance of the data set. It's all explained by one function. Well, two actually, the spatial structure and the temporal structure functions.

We can get the PCs by projecting the eigenvectors on the original data. The result will be exactly as we got in the SVD method, except for that pesky change of sign. Take the transpose of the eigenvector matrix and multiply it on the left into the original data.

```
>> p=v'*a
p =
  -2.2361   -4.4721    6.7082   -8.9443
        0         0         0         0
```

Finally recover the original data again by multiplying the eigenvector matrix(v) times the PC matrix(p).

```
>> A=v*p
A =
   2.0000    4.0000   -6.0000    8.0000
   1.0000    2.0000   -3.0000    4.0000
```

Wow, isn't math wonderful? And computers and software are amazing, too!

4.8 Applications of EOF/PC Analysis

4.8.1 Data Compression

EOF/PC analysis is a kind of functional representation, where a set of spatial/structure functions are presumed (or derived from the data in the case of EOFs), and the data is represented in terms of the amplitude coefficients of these structures for a set of observation times/samples. In numerical weather prediction the horizontal structure is often represented by smooth analytical/mathematical functions on the sphere, called spherical harmonics, that form an orthogonal and complete basis set. The data are then stored and manipulated as the coefficients of these structure functions, rather than as values at specific locations. Normally, the representation in terms of functions takes fewer numbers (fewer numerical degrees of freedom) than a similarly accurate representation in terms of the values at specified latitude/longitude grid points. That's because of the auxiliary information contained in the functions themselves.

EOF/PC analysis is different in that the structure functions are not pre-specified mathematical functions, but rather are determined from the data itself. They represent the correlated structures in the data. Often a large amount of the variance of a data set can be represented with a relatively small number of EOFs, so that when the data are stored as the PCs, the volume required is small. For example, the human fingerprint can be represented in great detail with about 10 EOFs. It can be shown that Fourier analysis is optimal in a least squares sense. But EOF analysis will often beat Fourier analysis in terms of efficiency of representation, when the data contain structures that are not easily represented by a standard set of analytical functions used in Fourier analysis (e.g. sines and cosines). Fingerprints are better represented by EOFs than by Fourier series because the fingerprint patterns are simpler than they appear, being composed of a set of whorls that occupy a rather small range of wavenumbers in the x-y space of fingerprint area. Fourier analysis has to carry along all of the wavenumbers needed to span a print of a certain size, whereas EOF analysis can concentrate on the scales and shapes that contain the real information, and thus require far fewer stored numbers to reproduce a given individual's print.

In general EOF analysis performs well when most of the variability is contained in a small number of localized or irregular structures. Even in situations where EOFs produce a more efficient representation of the data Fourier representation may be preferred, because of its uniqueness, familiarity, and precise mathematical definition. In addition, because of Fast Fourier Transform techniques, Fourier analysis is generally much more efficient computationally than EOF analysis, which requires matrix decompositions and naive transform methods.

4.8.2 Statistical Prediction

Earlier we supposed that a transformation existed for converting a data set expressed in x space in a new z space, according to the following transformation.

$$\begin{aligned} z_1 &= e_{11}x_1 + e_{12}x_2 + e_{13}x_3 + ... + e_{1M}x_M \\ z_2 &= e_{21}x_1 + e_{22}x_2 + e_{23}x_3 + ... + e_{2M}x_M \\ .. &= ..... \quad ..... \quad ..... \quad ... \quad ..... \\ z_M &= e_{M1}x_1 + e_{M2}x_2 + eM3x_3 + ... + e_{MM}x_M \end{aligned} \tag{4.41}$$

The equation above is equivalent to (4.17), so that the z's are equivalent to the columns in the coefficient matrix **Z**. These are the loadings or distances needed to express an observation in the new orthonormal basis set defined by e_{ij} = **E**. These loadings are often called the principal components The e_{ij} express the unit vectors of the new space in terms of the old x coordinates. The statement (4.20) that the coefficients necessary to describe the observations in the new coordinate space are orthogonal over the data set is equivalent to the statement that the covariance matrix of the z's is diagonal. We have in particular that,

$$\begin{bmatrix} \overline{z_1^2} & \overline{z_2 z_1} & ... & \overline{z_N z_1} \\ \overline{z_1 z_2} & \overline{z_2^2} & ... & \overline{z_N z_2} \\ ... & ... & ... & ... \\ \overline{z_1 z_N} & \overline{z_2 z_N} & ... & \overline{z_N^2} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_N \end{bmatrix} \tag{4.42}$$

which is, again, another way of writing (4.13), the result that the coefficients are orthogonal when summed over all of the realizations. Following (4.13) we can write this in matrix notation as,

$$\frac{1}{N}\mathbf{Z}^T\mathbf{Z} = \Lambda \tag{4.43}$$

where $\mathbf{\Lambda}$ is again a matrix with the eigenvalues down the diagonal. If we think of the realizations as observations of the state vector at different times, then the coefficient vectors in the new basis set are orthogonal in time. The diagonal elements of the covariance matrix of the coefficients are equal to the eigenvalues of the covariance matrix of the x's. The trace of the covariance matrix of the x's must equal the trace of the covariance matrix of the z's, which is equal to the sum of the eigenvectors, which is equal to the total variance.

$$\sum_{i=1}^{M} \overline{x_i^2} = \sum_{i=1}^{M} \overline{z_i^2} = \sum_{i=1}^{M} \lambda_i \tag{4.44}$$

If the variables have been standardized, then the trace of the covariance (in this case the correlation) matrix is just M, the number of components in the state vector.

Suppose that we now try to express a variable y at some time n in terms of our set of predictors expressed in the new coordinate system, the z's.

$$\hat{y}_n = a_1 z_{1n} + a_2 z_{2n} + ... + a_M z_{Mn} \tag{4.45}$$

Since the z's are uncorrelated in time, the regression can be done as an independent linear fit for each z. With these predictors there is no problem with automatic correlation (it is zero), and the multiple correlation coefficient increases monotonically as we add additional predictors.

$$R_m^2 = r_m^2(\hat{y}, y) = r^2(z_1, y) + r^2(z_2, y) + ... + r^2(z_m, y) \tag{4.46}$$

where $m<M$.

4.8.3 Prefiltering

In addition to restructuring a data set for more efficient and stable statistical prediction as discussed above, it may be desirable to prefilter data sets by decomposition into EOFs for other purposes. EOF analysis can make the analysis variables uncorrelated in time, and can reduce the number of variables necessary to explain a given fraction of the variance. These features are important in regression analysis as described above, and also in the Canonical Correlation Analysis method, where such prefiltering may be necessary in order to insure stable results.

4.8.4 Exploratory Data Analysis

Given a data set consisting of observations of a single variable at many spatial points, or several variables at the same point, both observed at many different times, EOF analysis can be used as a tool to search for characteristics structures in the spatial, parameter, or time dimensions of the data set. The spatial structure and associated temporal structure of a data field may be helpful in identifying mechanisms that produce variability in the data set.

Wavelike phenomena are easily picked up by EOF analysis. For example, if the data set consists of a spatial pattern that oscillates in time,

$$\phi(x,t) = f(x)\cos\omega t$$

then EOF analysis will pick out the spatial pattern $f(x)$ as the first EOF, and its corresponding principal component will oscillate in time with the frequency ω. Such a pattern is often called a standing wave, because it has a fixed spatial pattern that oscillates in time. For example, consider a data set consisting of the time series of monthly mean surface air temperature at a regular grid of points in the Northern Hemisphere for several years. The first several EOFs would be associated with the annual cycle of temperature,

which would have several spatial patterns and several periods corresponding to harmonics of the annual cycle (e.g. periods of 12, 6, 4, … months).

A traveling wave is one whose phase is not fixed in space. A traveling wave can be represented by two patterns in space with time signatures that are in quadrature.

$$\phi(x,t) = f(x)\cos\omega t + g(x)\sin\omega t$$

In this case EOF analysis would produce two spatial structures, with approximately equal eigenvalues, whose associated principal components are in quadrature (out of phase by 90°, or one quarter of a wavelength).

4.8.4 Factor Analysis[1]

According to Dillon and Goldstein(1984) factor analysis is the search for that part of the variance of a particular variable that it shares with other variables in the set. There are many ways to calculate this. Here we mention only the method whereby we start from the principle component analysis and approach the factors by truncating and rotating the principle components or EOFs. In this way we get a few structures that are simple in the sense that some variables in the set are "on" (e.g. have amplitudes of one) while the rest are "off" (have amplitudes close to zero). The basic orthogonal EOF or PC analysis is unique, whereas one can construct many possible sets of factors by varying the methods used to construct the factors or the parameters used with these methods. Seems to me that it gets murky quickly.

Suppose we wish to express our original variables in terms of a new set of variables, as we did in the case of principal component analysis, except that we limit the number of new variables to be relatively small. We therefore cannot require that our new "factors" explain all of the variance of the original variables. We would like to express a large fraction of the variance with a fairly small number of factors. We want a parsimonious description of the essential behavior. From (4.18) we have

$$\mathbf{X} = \mathbf{EZ} \qquad or \qquad X_{im} = E_{ij} Z_{jm} \tag{4.46}$$

where i is the index of the original vector space ($1<i<M$), j is the index of the eigenvector space ($1<j<M$), and n is the index of realizations or time ($1<n<N$). We can construct an approximate expression for the state matrix $\mathbf{X}$ by making the summation over less than the total number of eigenvectors by rewriting (4.19a) as,

$$\hat{X}_{in} = \sum_{j=1}^{m} E_{ij} Z_{jn} \qquad where\ m<M \tag{4.47}$$

[1]Note that the use of the descriptors EOF's, principal components and factors varies from author to author and among fields (e.g. among sociology, psychology, and meteorology).

Here, as before, N is the sample size (number of realizations) and M is the length of the state vector (the number of original variables). The new index limit $m < M$ is the number of factors retained, or the number of eigenvectors used to represent the data.

How Many Factors Should Be Retained?

A logical first proposal would be to simply choose the first few principal components as the factors to be retained in the approximation (4.47). How would we determine the number of principal components to retain? The usual procedure is to look at the eigenvalues. If we are working with standardized variables, so that the eigenvalues are those of a correlation matrix, then the sum of the eigenvalues must be M. The average value is therefore 1.0. If our principal components are drawn from a completely random distribution with no real structure, then we would expect each eigenvalue to be close to 1. We might therefore begin by saying that no factors associated with eigenvalues smaller than 1 will be retained. Are the remaining principal components necessarily meaningful? The answer is, not necessarily and probably not.

We know that when we sample correlation coefficients, they are distributed about their true mean value. Therefore, so long as we have a finite sample, we will always get some eigenvalues greater than l and some less than 1, even from a population with zero cross correlation. That is, the true correlation matrix has diagonal elements equal to l and all other elements zero, but we never get that because of sampling errors. To address the question of the statistical significance of principal components we need to consider the sample size, N, the number of variables and hence eigenvalues, M, and use a t-statistic to estimate the distribution of eigenvalues we are likely to get by chance. Often in geophysical applications the data are correlated in space and time. In this case we should expect smooth functions in space and time to explain a disproportionate amount of the variance.

These considerations are illustrated in the diagram below. The ordinate is the fraction of variance of the x's, that can be explained by the fraction of the total number of principal components m/M, which is the abscissa. The null hypothesis might be that the x's have no structure, so that all the eigenvalues should be l.0. In practice not all the sample eigenvalues will be l.0, and because we order the eigenvalues from largest to smallest *a posteriori*, the actual curve we get will start out steeply and then level off like the dashed line. The question is, will the curve we get start out more steeply than we would expect to get by chance from a sample of unstructured data? This question is addressed in the references by North et al. (1982), and Overland and Preisendorfer (1982). The answer depends on a lot of things, including the spacing of the data compared to the natural scale of variation of the variable in question.

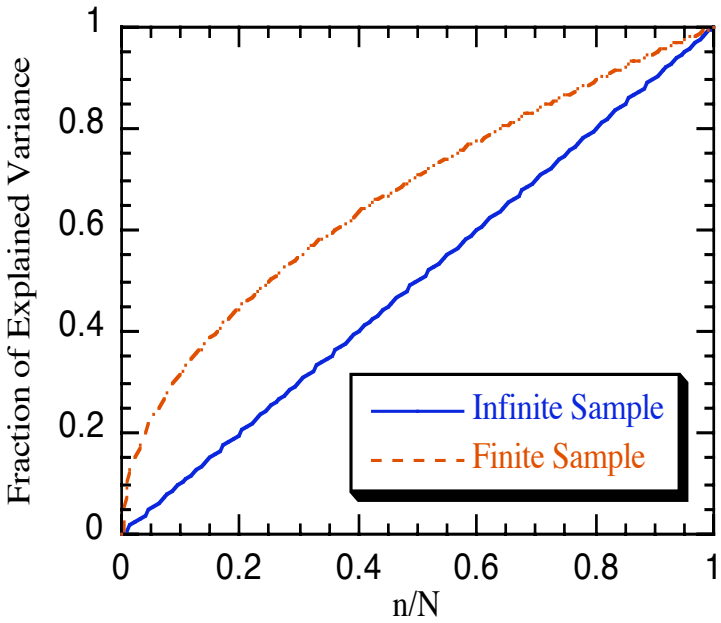


Figure. 4.1 Illustration of explained variance as a function of the number of principal components retained, *n*, out of the total possible, *N*, for a data set with no real structure (N is the structure dimension here, not the sample size.). Because of sampling errors, some EOFs will always appear to explain more of the variance than others do.

We may also consider the derivative of the curve above, which is just proportional to the eigenvalue spectrum. North, et al.(1982) suggested using the point in the eigenvalue spectrum where the spectrum appears to break from a rapid decline from the first few significant eigenvalues, to the more gradual decline of the remaining eigenvalues, which we expect to be associated with noise. Noise in the atmosphere tends to be autocorrelated in space and time, which gives it the character of what we call "red noise". This means that the largest scales and lowest frequencies will tend to explain more of the variance than the smaller scales and higher frequencies. The first few EOF's thus tend to have a smooth, appealing large-scale structure, or even look wavelike, even when the data have no real coherent structures within them, other than this autocorrelation.

North et al.(1982) showed that the 95% confidence error in the estimation of the eigenvalues is approximately,

$$\Delta\lambda = \lambda\sqrt{2/N^*} \qquad (4.48)$$

where l is the eigenvalue and N^* is the number of degrees of freedom in the data set. We will return to discussing how to estimate the number of degrees of freedom later. It involves estimating the autocorrelation of the data. If the eigenvalues of adjacent EOF's are closer together than this standard error, then it is unlikely that their particular structures are significant, since any linear combination of the two eigenvectors is equally

significant. Usually we would look for a place in the eigenvalue spectrum where it levels off so that successive eigenvalues are indistinguishable. We would not consider any eigenvectors beyond this point as being special.

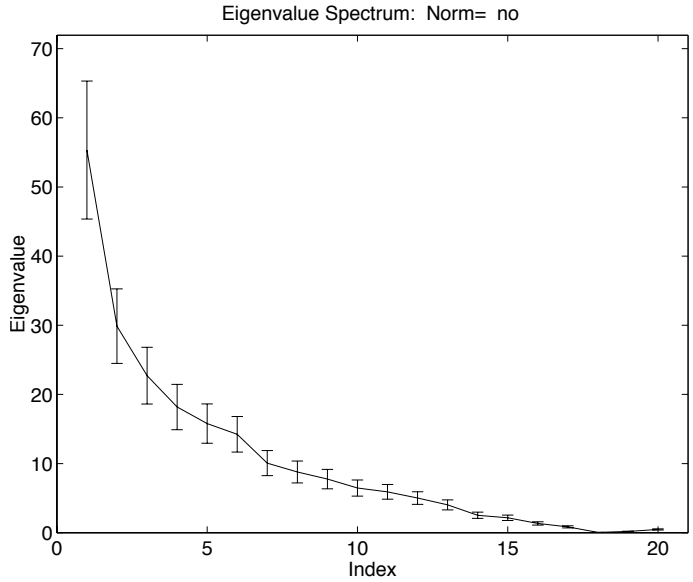


Figure 4.2 Example eigenvalue spectrum, with confidence limits. Index number of the eigenvalue, ordered from largest to smallest. It looks like only the first one is distinct.

In the above example only the first eigenvalue is different from the others according to the North et al. criterion indicated by the error bars on the eigenvalue estimates. The second and third eigenvalues are indistinguishable, so you want to think carefully about how you interpret their structures, or just don't. Since their eigenvalues are the same to within the statistical uncertainty, any linear combination of the two eigenvectors is just as significant as the eigenvectors individually. Lag correlation analysis of the principal components will sometimes show that pairs of eigenvectors are in phase quadrature in time, which may indicate that they represent a propagating structure.

4.9 Interpretation of EOFs

In interpreting EOFs one must remember exactly what they are. They are mathematical constructs that are chosen to represent the variance over the domain of interest as efficiently as possible, and also be orthogonal with each other. Sometimes these mathematical constraints will select out scientifically interesting structures in a data set, but not always. EOF analysis will always pick out some structures that represent more of the variance than the others will, and they will always tend to look wavelike, because they are constrained to be orthogonal. If you put white or slightly red noise through EOF analysis, it will produce structures that resemble the Fourier modes for the

domain of interest, even when the data set is pure noise. If the data are autocorrelated in time or space, for which we can use the model of red noise, then the eigenvalue spectrum will be peaked, with some EOFs explaining much more of the variance than others. These will be smoothly varying large-scale structure, and it is tempting to interpret them physically, although they are telling you nothing but that the adjacent data points are correlated with each other. The particular structures obtained will depend on the particular spatial and temporal slice of data that was used to compute them. Just because EOF analysis produces large-scale wavelike structures does not mean the data contain coherent wave structures shaped like the sample EOFs. Below are some suggestions for steps to follow in attempting to ascertain whether EOFs produced in an analysis are a reflection of scientifically meaningful structures in the data.

1. Is the variance explained by the EOF more than you would expect if the data had no structure? What is your null hypothesis for the data field? Is it white noise? Is it red noise? Can the EOFs of interest support a rejection of this uninteresting null hypothesis?

2. Do you have any *a priori* reason for expecting the structures that you find? Are the structures explainable in terms of some theory? Do the spatial and temporal structures of the modes behave consistently with theory and *a priori* expectation?

3. How robust are the structures to the choice of structure domain? If you change the domain of the analysis, do the structures change significantly? If the structure is defined in geographical space, and you change the size of the region, do the structures change significantly? If the structures are defined in a parameter space, and you add or remove a parameter, do the results change in a sensible manner, or randomly?

4. How robust are the structures to the sample used? If you divide the sample into randomly chosen halves and do the analysis on each half, do you consistently get the same structures?

4.10 Rotation of EOFs and Factors

Sometimes the orthogonality constraint will cause structures to have significant amplitude all over the domain (e.g. spatial domain) of the analysis, when we expect the structures to be much more localized. To reduce the effect of the orthogonality constraint and allow more localized structures to emerge, we can consider rotation of the eigenvectors (Horel 1984, Richman 1986). After we have chosen a subset of the principal components as our factors, we may continue to improve the factors by rotating the eigenvectors to form new factors. If we do this we must sacrifice one or both of the orthogonalities. Often these orthogonalities are artificial constraints, anyway, and we can get a much more compact and desirable factorization by removing these constraints. Rotation of EOFs is most desirable when we are using EOF analysis to look for physical structure in fields, rather than using the analysis for predictive purposes. Often the orthogonality constraints combined with the requirement that the EOF explain as much as possible the variance over the whole domain yield artificial global-looking structures. If

we expect the structures to be more local in nature, then orthogonal or oblique rotations of EOFs may be advantageous. Orthogonal rotations relax the constraint on the orthogonality of the factors, but retain the constraint that the data be uncorrelated in time when described with the new factors. If we are looking at geopotential maps, then the factors will be correlated in space but uncorrelated in time. We can thus use orthogonally rotated principal components (factors) in statistical prediction, since orthogonality in time is retained, so that the EOFs remain uncorrelated in time.

The criteria used to determine the optimal rotation seek to maximize the "simplicity" of the structure of the factors, where the structure is defined as the matrix eigenvectors, e_{jp}. To do this, of course, we must have some reason to expect that the real structures are 'simple' and produce a mathematical definition of simplicity. Recall that the principal components were chosen to maximize the variance of all of the variables (x's) that can be explained with a single vector, e_j. The sum of the variance-explained criterion often produces eigenvectors that are complex and difficult to interpret. Simplicity of structure of the eigenvectors or parallel loading vectors can be measured by several criteria. Basically, simplicity of structure is supposed to occur when most of the elements of the eigenvector are either of order one (absolute value) or zero, but not in between.

The Quartimax Criterion seeks an orthogonal rotation of the factor-loading matrix $e_{jp} = \mathbf{E}$ into a new factor matrix $b_{jp} = \mathbf{B}$ for which the variance of squared elements of the eigenvector is a maximum. The quantity to be maximized is,

$$s_{b^2}^2 = \frac{1}{mM} \sum_{j=1}^{M} \sum_{p=1}^{m} \left(b_{jp}^2 - \overline{b^2} \right)^2 \tag{4.49}$$

where

$$\overline{b^2} = \frac{1}{mM} \sum_{j=1}^{M} \sum_{p=1}^{m} b_{jp}^2$$

The quantity (4.49) to be maximized can be simplified to,

$$\mathrm{Q} = \frac{1}{mM} \sum_{j=1}^{M} \sum_{p=1}^{m} b_{jp}^4 - \overline{\left(b^2 \right)} \tag{4.50}$$

Since the mean-squared loading remains constant under orthogonal rotations, the criterion is simply equivalent to maximizing the sum of the fourth power of the loadings, hence the name Quartimax.

The Varimax Method more nearly approximates simple structure of the individual factors. The simplicity of an individual factor is defined as the variance of its squared loadings.

$$s_p^2 = \frac{1}{M}\sum_{j=1}^{M}\left\{b_{jp}^2\right\}^2 - \frac{1}{M^2}\left\{\sum_{j=1}^{M} b_{jp}^2\right\}^2 \quad , p = 1,2,...,m \tag{4.51}$$

When the variance s_p^2 is at a maximum the factor has the greatest simplicity in the sense that its factor loadings tend toward unity or zero. The criterion of simplicity of the complete factor matrix is defined as the maximization of the sum of the simplicities of the individual factors,

$$s^2 = \sum_{p=1}^{m} s_p^2 \tag{4.52}$$

Equation (4.52) is called the *raw varimax criterion*. In this form the factors with larger eigenvalues contribute more to the criterion than the others because of their larger factor loadings. This bias toward the first eigenvector, which tends to have a simple structure anyway, can be removed by taking out the weighting of the eigenvectors by their explained variance. Different weights are applied to the variables according to their communalities. The communality is related to the sum of the squares of the factor loadings for a particular variable,

$$h_j^2 \propto \sum_{p=1}^{m} a_{jp}^2 \tag{4.53}$$

The final normalized varimax criterion is,

$$V = m\sum_{p=1}^{m}\sum_{j=1}^{M}\left\{\frac{b_{jp}}{h_j}\right\}^4 - \sum_{p=1}^{m}\left\{\sum_{j=1}^{M}\frac{b_{jp}^2}{h_j^2}\right\}^2 \tag{4.54}$$

The Varimax method is often preferred over the Quartimax method because the sensitivity to changes in the number (or choice) of variables is less. The difference between the results obtained with the two methods in practice is usually small.

4.12 Eight Physical Variables Example

A classic example of the use of factor analysis is the 'Eight Physical Variables Example' which looks at the correlations between eight measures of human anatomy. These eight variables are highly redundant, since length of forearm, arm span, and height are all highly correlated. In cases such as this factor analysis can describe the anatomy with fewer variables. The correlation matrix for the eight physical variables is shown below.

Correlation matrix

	height	arm span	forearm...	lower le...	weight	bitrocha...	chest gi...	chest wi...
height	1							
arm span	.846	1						
forearm l...	.805	.881	1					
lower leg...	.859	.826	.801	1				
weight	.473	.376	.38	.436	1			
bitrochan...	.398	.326	.319	.329	.762	1		
chest girth	.301	.277	.237	.327	.73	.583	1	
chest wid...	.382	.415	.345	.365	.629	.577	.539	1

The correlation matrix is of course closely related to multiple regression and explained variance. The matrix below shows the fraction of the variance of each variable that can be explained by all of the other variables. Since this is relatively large, it suggests that the variables are closely related and that the data set is therefore a good candidate for factor analysis. The off diagonal terms are the fractions of the variance of each variable that can only be explained by the variable indicated. Since the diagonal terms are large, and most of the off diagonal terms are fairly small, this again means that the data set is a good candidate for factor analysis, since the variables are highly redundant, none of them explain a lot of the variance all by themselves.

Partials in off-diagonals and Squared Multiple R in diagonal

	height	arm span	forearm...	lower le...	weight	bitrocha...	chest gi...	chest wi...
height	.816							
arm span	.346	.849						
forearm l...	.072	.584	.801					
lower leg...	.479	.179	.188	.788				
weight	.183	-.196	.1	.056	.749			
bitrochan...	.103	-.005	.027	-.122	.492	.604		
chest girth	-.146	.091	-.116	.131	.491	.054	.562	
chest wid...	-.086	.248	-.087	-.025	.238	.177	.12	.478

The eigenvalue spectrum for the first four eigenvectors is shown in the panel below. The first two exceed the expected value of one, and together explain 80.7 percent of the variance. Notice that the third and fourth eigenvalues are almost the same and small. Pretty clearly we have only two significant eigenvectors here.

Eigenvalues and Proportion of Original Variance

	Magnitude	Variance Prop.
Value 1	4.673	.584
Value 2	1.771	.221
Value 3	.481	.06
Value 4	.421	.053

The eigenvectors are shown below as the columns of the E matrix, although only the first four are shown. The first is a constant; all variables go up and down together. The second is a square wave, the first four go up when the second four go down. The third and fourth have more wiggles.

Eigenvectors

	Vector 1	Vector 2	Vector 3	Vector 4
height	-.398	.28	-.101	-.107
arm span	-.389	.331	.113	.068
forearm leng...	-.376	.345	.015	-.047
lower leg len...	-.388	.297	-.145	.124
weight	-.351	-.394	-.213	-.114
bitrochanter...	-.312	-.401	-.073	-.713
chest girth	-.286	-.436	-.421	.63
chest width	-.31	-.314	.853	.221

We can also show the eigenvector structure as the correlation of the principal component of the eigenvector with the original data. This correlation matrix may be calculated according to the formula,

$$\mathbf{S}_{ij} = \frac{X_{in} Z_{nj}}{\left(Z_{jn} Z_{nj} \bullet X_{in} X_{ni} \right)^{1/2}} \tag{4.55}$$

where summation only over the inner repeated index is implied in the denominator. This gives a matrix where the columns are the correlation of the principal components with the individual original physical variables, as shown below. The shape is similar to the EOF's, but the magnitudes are now correlations. The correlations for the first two unrotated factors are shown below.

Unrotated Factor Matrix

	Factor 1	Factor 2
height	.859	-.372
arm span	.842	-.441
forearm leng...	.813	-.459
lower leg len...	.84	-.395
weight	.758	.525
bitrochanter...	.674	.533
chest girth	.617	.58
chest width	.671	.418

If we perform an orthogonal rotation we obtain the correlation structure below, in which we find that the first factor has strong correlations with the first four physical variables related to length (the bone factor), and weak correlations with the remaining four, and the second factor has strong correlations with the weight and width variables (the flesh factor) and weak correlations with the length variables.

Orthogonal Transformation Solution-Varimax

	Factor 1	Factor 2
height	.9	.26
arm span	.93	.195
forearm leng...	.919	.164
lower leg len...	.899	.229
weight	.251	.887
bitrochanter...	.181	.84
chest girth	.107	.84
chest width	.251	.75

If we perform an oblique rotation, in which both spatial and temporal orthogonality are relaxed, we can make the structures even more close to zero and one, but the basic structure remains the same.

Oblique Solution Primary Pattern Matrix-Orthotran/Varimax

	Factor 1	Factor 2
height	.919	.033
arm span	.973	-.047
forearm leng...	.971	-.08
lower leg len...	.928	-4.82E-4
weight	-.001	.922
bitrochanter...	-.064	.89
chest girth	-.146	.911
chest width	.043	.768

Which of the solutions, the unrotated eigenvectors or the rotated eigenvectors is more useful or reasonable? The orthogonal eigenvectors say that you can explain the variance of physical factors with two functions, one in which all physical variables go up and down together, and one in which the bone length and flesh volume factors are anti-correlated. You can see how this pair of functions would explain the variance very efficiently, but the physical interpretation of the second EOF is counterintuitive. It says that bone length and body mass are anti correlated, which is not reasonable. The rotated interpretation is that there are two factors, a bone factor and a flesh factor. These factors are correlated sometimes, but it is a more reasonable way to look at the data than blindly interpreting the EOF's.

References on EOF and SVD Analysis

Bretherton, C. S., C. Smith and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data sets. *J. Climate*, **5**, 541-560.

Dillon, W. R. and M. Goldstein, 1984: *Multivariate Analysis: Methods and Applications*. John Wiley & Sons, New York, 587.

Gorsuch, R., 1983: *Factor Analysis*, Lawerence-Erlbaum Pubs., BF 39 G58.

Harman, H.H., 1976: *Modern Factor Analysis*. U. Chicago Press, QA 276 H38.

Horel, J.D., 1981: A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field. *Mon Wea.Rev*., **109**, 2080-2092.

Horel, J.D., 1984: Complex principal component analysis: Theory and examples. *J. Appl. Meteorol*., **23**, 1660-1673.

Kutzbach, J.E., 1967: Empirical eigenvectors of sea-level pressure, surface temperature, and precipitation complexes over North America. *J. Appl. Meteorol*., **6**, 791-802.

Lorenz, E. N., 1956: *Empirical orthogonal functions and statistical weather prediction*. Science Rept. 1, Statistical Forecasting Project. Department of Meteorology, Massachusetts Institute of Technology, Cambridge,

Morrison, D.F., 1976: *Multivariate Statistical Methods*. McGraw-Hill.

Mulaik, S., 1972: *The Foundations of Factor Analysis*. McGraw-Hill.

North, G.R., T.L. Bell, R.F. Cahalan and F.J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev*., **110**, 699-706.

Overland, J.E., and R.W. Preisendorfer, 1982: A significance test for principal components applied to a cyclone climatology. *Mon. Wea. Rev*., **110**, 1-4.

Richman, M.B., 1986: Rotation of principal components. *J. Climatology*, **6**, 293-335. (Review Article)

Strang, G., 1988: Linear algebra and its Applications, 3rd. Edition, Harcourt Brace, 505pp.

Timm, N., 1975: Multivariate Analysis with Applications in Education and Psychology. Cole.

von Storch, H. and F. W. Zwiers, 1999: *Statistical analysis in climate research*. Cambridge University Press, 484 pp.

Wallace, J. M., C. Smith and C. S. Bretherton, 1992: Singular value decomposition of wintertime sea-surface-temperature and 500 mb height anomalies. *J. of Climate*, **5**, 561-576.

4.12 Maximum Covarance Analysis

In EOF analysis we looked for structures that explained the maximum amount of variance in some data matrix. We, somewhat arbitrarily, defined a structure dimension, which we thought of as space, and a sampling dimension, which we thought of as time. In Maximum Covarance Analysis (MCA) and Canonical Correlation Analysis (CCA) we take two data matrices that are examples of different structures, or state vectors, but which share a common sampling dimension. For example, you could consider the fields of sea surface temperature and surface chlorophyll content, measured at the same set of times. Or you could consider a case where the sampling dimension is a collection of hospital patients, and where one of the two state vectors is their "Eight Physical Variables", and the other state vector is their cholesterol data (say, 3 numbers). In the latter case suppose the sample of patients is N. You could make an augmented state vector by taking the 11 numbers that include their 8 physical variables and their 3 cholesterol values. You could then do EOF analysis of the 11 x N data matrix, and see if the structures chosen include both physical variables and cholesterol variables. You might expect that if the physical variables and the cholesterol variables vary together, then they should show up in the same structures that efficiently explain the variance of the augmented or combined state vector.

In MCA analysis you first compute the covariance matrix between the 8 x N and 3 x N data sets to make an 8x3 covariance matrix. Then you would do SVD analysis of this covariance matrix. The resulting singular vectors and singular values would tell you about structures in one data set that are correlated with structures in the other data set as you sample across the population of your hospital patients. For example, do the flesh variables correlate strongly with high levels of "bad" cholesterol??

Prohaska (1976) first perhaps used MCA in the meteorological literature, although it has long been used in the social sciences. Bretherton et al. (1992) and Wallace et al. (1992) popularized it for meteorological and oceanographic use.

4.12.2 MCA Mathematics

Let us suppose we have two data matrices **X** and **Y** of size *MxN* and *LxN*, where *M* and *L* are the structure dimensions and N is the shared sampling dimension. We begin by taking the inner product of these two matrices to obtain an *MxL* covariance matrix, as depicted graphically and formulaically below.

$$\frac{1}{N}\mathbf{X}\mathbf{Y}^T = \mathbf{C}_{XY} \tag{4.56}$$

Normally, we would remove the time mean (average over the sample *N*) from X and Y, so that $\mathbf{C}_{XY}$ is indeed a covariance matrix in the usual sense. The graphical representation of this matrix operation is:

$$\begin{matrix} & \text{N x M} & & \text{L x N} & & \text{L x M} \\ \frac{1}{N} \times M & \left[\quad X \quad \right] & & \left[\quad Y^T \quad \right] & N = & \left[\quad C_{XY} \quad \right] M \end{matrix}$$

Having formed the covariance matrix between the two data sets by projecting over their sampling dimension, SVD analysis can be done to decompose the covariance matrix into its column space and row space and associated singular values. The column space will be structures in the dimension M that are orthogonal and have a partner in the row space of dimension L. Together these pairs of vectors efficiently and orthogonally represent the structure of the covariance matrix. The hypothesis is that these pairs of functions represent scientifically meaningful structures that explain the covariance between the two data sets. Let's set the deeper issues aside for a moment and just look at some of the features of the mathematics. We consider the SVD of the MxL covariance matrix.

$$\mathbf{C}_{XY} = \mathbf{U}\Sigma\mathbf{V}^T \tag{4.57}$$

The columns of **U** (MxM)are the column space of $\mathbf{C}_{XY}$ and represent the structures in the covariance field of **X**. The columns of **V** are the row space of $\mathbf{C}_{XY}$ and are those structures in the **Y** space that explain the covariance matrix. The singular values are down the diagonal of the matrix **Σ**. The sum of the squares of the singular values σ_k is equal to the sum of the squared covariances between the original elements of **X** and **Y**.

$$\left\| C_{XY} \right\|^2 = \sum_{i=1}^{M} \sum_{j=1}^{L} \left(\overline{x_i y_j} \right)^2 = \sum_{k=1}^{L} \sigma_k^2 = \sum_{k=1}^{L} \left(\overline{x_i^* y_j^*} \right)^2 \tag{4.58}$$

Here x* and y* correspond to the principal components of EOF analysis. They are the projections of the singular vectors onto the original data.

$$\mathbf{X}^* = \mathbf{U}^T\mathbf{X} \quad ; \quad \mathbf{Y}^* = \mathbf{V}^T\mathbf{Y} \tag{4.59}$$

Since the input matrix is a covariance matrix, the singular values have units of covariance, or correlation if the original matrix is a correlation matrix. The singular value is also equal to the covariance between the expansion coefficients $\mathbf{X}^*$ and $\mathbf{Y}^*$ of the two fields.

$$\sigma_k = \overline{x_k^* y_k^*} \tag{4.60}$$

The sum of the squares of the singular values is equal to the square of the Frobenius Norm

(the sum of the squares of the elements) of the covariance matrix, which is the squared covariance. One can ask whether one mode stands out over the others by asking whether it explains a large fraction of the covariance, although it is also necessary that the total covariance between the two data sets be large, or the results are not meaningful.

4.12.3 Scaling and Display of Singular Vectors

The singular vectors are normalized and non-dimensional, whereas the expansion coefficients have the dimensions of the original data. Like EOFs, singular vectors can be scaled and displayed in a number of ways. The sign is arbitrary, but if you change the sign of one component, you must change the sign of everything, including either left or right singular vectors and their corresponding expansion coefficients. One must remember that the singular vectors, as defined here, are constructed to efficiently represent covariance, and they may not, in general, be very good at representing the variance structure.

In constructing the singular vector patterns by projecting the data onto the expansion coefficients, MCA analysis is a little different than EOF analysis, since to get the structure of the left field, you project the left field data onto the expansion coefficient of the right singular vector, and vice versa. These are heterogeneous regression maps.

$$\mathbf{u}_k = \frac{1}{N\sigma_k}\mathbf{X}\mathbf{y}_k^{*T} \quad or \quad u_{jk} = \frac{1}{N\sigma_k}\sum_{i=1}^{N} x_{ji}y_{ik}^{*} = \frac{1}{\sigma_k}\overline{x_j y_k^{*}} \tag{4.61}$$

$$\mathbf{v}_k = \frac{1}{N\sigma_k}\mathbf{Y}\mathbf{x}_k^{*T} \quad or \quad v_{jk} = \frac{1}{N\sigma_k}\sum_{i=1}^{N} y_{ji}x_{ik}^{*} = \frac{1}{\sigma_k}\overline{y_j x_k^{*}} \tag{4.62}$$

These identities can be used as consistency checks, and as a way of constructing dimensional singular vector patterns from the time series of expansion coefficients. The amplitude information associated with the singular vector expansion can be incorporated into the singular vector patterns for display purposes by mapping the covariance between the normalized expansion coefficient time series (set standard deviation to one) of the left singular vector with the right data field and vice versa – the covariance of the normalized expansion coefficient of the right field with the left data field. If the amplitudes are very small, this is a clue that the analysis may not be meaningful.

In general, two types of covariance maps can be formed from the products of MCA analysis:

- Heterogeneous regression maps: regress (or correlate) the expansion coefficient time series of the left field with the input data for the right field, or do the same with the expansion coefficient time series for the right field and the input data for the left field. As computed above.

- Homogeneous regression maps: regress (or correlate) the expansion coefficient time series of the left field with the input data for the left field, or do the same with the right field and its expansion coefficients.

$$\mathbf{u}_k = \frac{1}{N\sigma_k}\mathbf{X}\mathbf{x}_k^{*T} \quad or \quad u_{jk} = \frac{1}{N\sigma_k}\sum_{i=1}^{N} x_{ji}x_{ik}^{*} = \frac{1}{\sigma_k}\overline{x_j x_k^{*}} \tag{4.63}$$

$$\mathbf{v}_k = \frac{1}{N\sigma_k}\mathbf{Y}\mathbf{y}_k^{*T} \quad or \quad v_{jk} = \frac{1}{N\sigma_k}\sum_{i=1}^{N} y_{ji}y_{ik}^{*} = \frac{1}{\sigma_k}\overline{y_j y_k^{*}} \tag{4.64}$$

The heterogeneous maps are characteristic of MCA analysis and are the same as the dimensional singular vectors. The homogenous fields show how the singular vectors do in explaining the variance of their own data set. If the patterns that explain covariance between two data sets are similar to the patterns that explain the variance in each data set, then the homogenous and the heterogeneous patterns should be similar. Another way to check this is to compare the singular vectors with the EOFs of each data set.

In contrast to the principal component time series of EOF analysis, the expansion coefficient time series of MCA are not mutually orthogonal. The correlation coefficient between the expansion coefficients for corresponding left and right singular vectors is a measure of the strength of the coupling between the two patterns in the two fields.

4.12.4 Normalized Root Mean Squared Covariance.

The total squared covariance, sum of the squares of all the elements of the covariance matrix is a useful measure of the strength of the simultaneous linear relationship between the fields. We can normalize this with the product of the variance of the left and right fields. If this statistic is very small, then the covariance between the two data sets is small, and it may not make sense to search for structure in this covariance.

$$RMSC = \left(\frac{\sum_{i=1}^{M}\sum_{j=1}^{L}\left(\overline{x_i y_j}\right)^2}{\left(\sum_{i=1}^{M}\overline{x_i^2}\right)\left(\sum_{j=1}^{L}\overline{y_j^2}\right)} \right)^{\frac{1}{2}} \tag{4.65}$$

The normalized root mean square covariance is on the order of 0.1 for well-correlated fields, but even smaller values may indicate “significant correlation” if the number of independent samples is large, so that small correlations can be distinguished from zero.

4.12.5 Statistical significance of MCA analysis:

I am unaware of any formal procedures for evaluating the statistical significance of MCA analysis, but MCA analysis is subject to the usual sampling fluctuations. Sampling errors can be significant if the number of degrees of freedom in the original data set is modest compared to the degrees of freedom in the structure (e.g. spatial). Some effort should be made to evaluate how many degrees of freedom the data set really has. The usual method of dividing the data set should be used, if possible, to test consistency. One should also try to evaluate how much variance and covariance are explained with a pair of patterns that may be of interest. Comparison against Monte Carlo experiments may also give insight into how probable it is that a given correlation pattern could have arisen by chance from essentially random data.

4.12.6 Criticisms of MCA Analysis:

Many caveats and criticisms have been offered for MCA analysis. Newman and Sardeshmukh(1995) asked what kind of relationship between x and y you might be trying to find with MCA analysis, and they mention three possibilities:

1. **x** and **y** are diagnostically related by a relationship like: $\mathbf{y} = \mathbf{L}\mathbf{x}$, where **L** is a linear operator, in the form of a matrix.
2. **x** and **y** are parts of a larger system.
3. **x** is a forcing for **y** in that the equation for **y** involves **x**, but not vice versa.

They asked whether MCA could distinguish these possibilities, but mostly investigated the first one. They showed that if $\mathbf{y} = \mathbf{L}\mathbf{x}$, then the singular vectors would bear this relationship to each other only under the very restrictive conditions that the operator **L** be orthogonal, or the covariance matrices of the two fields be diagonal. (An orthogonal matrix is one whose transpose is its inverse such that $\mathbf{L}^T = \mathbf{L}^{-1}$. They showed the example of the streamfunction and the vorticity field which are related by the Laplacian operator. The left and right singular vectors are not so related, so MCA analysis does not recover the underlying relationship between the variables.

Cherry(1996) compared MCA and CCA and recommended extreme caution in applying both techniques, since they tend to produce spurious spatial patterns. Cherry(1997) showed that singular vectors could be thought of as orthogonally rotated PC patterns, rotated so as to produce maximum correlation between pairs of rotated PCs. He recommends first carrying out separate PC analysis on the two data sets. If the two sets of patterns are significantly correlated and make sense scientifically for the data sets under consideration, then one has evidence of meaningful coupling. It can be argued that the correlation has not been forced on the two data sets by the analysis scheme. MCA will seek out coupled patterns in noise. It is less likely that patterns picked out from two data sets for the ability to explain variance in their own domain will be correlated with patterns in another domain, purely by

chance. Hu (1997) pointed out some lack of uniqueness problems with MCA analysis.

4.12 Canonical Correlation Analysis

Principal component analysis and MCA analysis can be performed in sequence, and we can call the result Canonical Correlation Analysis. The presentation here follows Barnett and Preisendorfer(1987), who recognized the need to reduce noise in statistical prediction by limiting the number of predictors. The raw predictors can be subjected to EOF/PC analysis, and new predictors formed from that subset the PC time series that explain most of the variance. The reduction of the dimension of the data set to the strongest PCs reduces the possibility that correlated patterns will emerge by chance from essentially random data. If desired, one can then normalize the time series of PCs so that they have unit variance (and this is what Barnett and Preisendorfer do). One then calculates the SVD from the covariance matrix of these normalized PC time series. This means that the correlation between patterns of EOFs is maximized by the SVD analysis, rather than the covariance. It is argued that CCA is more discriminating than MCA analysis, in that it is not overly influenced by patterns with high variance, but weak correlation, but it is also susceptible to sampling variability.

4.12.1 Treatment of input data

The first step is to perform EOF analysis of the original data for both the left and right fields and construct the time series of the PCs (principal components), which are the amplitudes of the EOFs at each sampling time for each data set. Of course this step only makes sense if the original data are highly correlated, so that EOF analysis makes sense. So the first step is an orthogonal rotation of the coordinate systems so that the first direction explains most of the variance, and so forth. The new data matrices, composed of the PCs for the two data sets are subject to two further procedures:

- Truncation to reduce the number of degrees of freedom in the input data sets from the original structure dimensions of the input fields x and y to some smaller dimension. Since the PCs are efficient in explaining the variance, a small number can explain a large fraction of the variance. In choosing the number of modes to be retained, one faces a tradeoff between statistical significance and explaining as much variance as possible. To have statistical significance argues for as few modes as possible so that the number of samples will be large compared to the number of degrees of freedom in the structure dimension. To include as much variance as possible, one would include more PCs in the analysis.
- Normalization to the make the variance over the sampling dimension unity for each PC. If the sampling dimension is time, this is just dividing each PC by its standard deviation in time. Hence all the PC time series are weighted equally, regardless of the amount of variance in the original data that they explain.

After these modifications, the modified data matrices no longer contain the

information necessary to reconstruct the original data sets.

The remainder of the analysis is very similar to MCA analysis. First construct the inner product across sampling dimension to form the covariance matrix between the two new data sets. Since these data set time series have been normalized, the covariance matrix is a correlation matrix between the retained PCs. The Frobenius norm of the correlation matrix may be interpreted as the total fraction of the variance of the left modified data set that is explained by the right modified data set, and vice versa.

4.12.2 The Canonical Correlations

Because the SVD is done on a correlation matrix, the singular values may be interpreted as correlation coefficients or "canonical correlations". SVD rearranges the PCs into combinations so that the first set in each modified input data series explains as much as possible of the correlation with the other modified data set. The structures in each field associated with these canonical correlations can be called the canonical correlation vectors, if you like.

4.12.3 How many PCs to Retain

Again this is more an art than a science, best determined by experimentation or *ad hoc* rules of thumb, such as retaining enough PCs to explain 70% of the variance in each data set. If too few PCs are retained, not enough variance may be retained to cover the important variability. If too many are retained, then statistical significance may be compromised. In any case, the number of degrees of freedom in the structure that are retained should be much less than the number of independent samples, or the results will have neither stability nor statistical significance. This is especially important when you have many more spatial grid points than independent samples, as is often the case in investigating interannual variability of global data fields.

If the coupling between the fields is large and the sample size is sufficiently large, the spatial patterns should be insensitive to the number of modes retained over a range of truncations.

References on MCA and CCA:

Barnett, T. P. and R. W. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825-1850.

Bretherton, C. S., C. Smith and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data sets. *J. Climate*, **5**, 541-560.

Cherry, S., 1996: Singular value decomposition analysis and canonical correlation analysis. *J. Climate*, **9**, 2003-9.

Cherry, S., 1997: Some comments on singular value decomposition analysis. *Journal of Climate*, **10**, 1759-61.

Fraedrich, K., C. Ziehmann and F. Sielmann, 1995: Estimates of spatial degrees of freedom. *J. Climate*, **8**, 361-369.

Hu, Q., 1997: On the uniqueness of the singular value decomposition in meteorological applications. *J. Climate*, **10**, 1762-6.

Jin, S.-X. and H. von Storch, 1990: Predicting the state of the Southern Oscillation using principal oscillation pattern analysis. *J. Climate*, **3**, 1316-29.

Newman, M. and P. D. Sardeshmukh, 1995: A Caveat Concerning Singular Value Decomposition. *J. Climate*, **8**, 352-360.

Prohaska, J., 1976: A technique for analyzing the linear relationships between two meteorological fields. *Mon. Wea. Rev.*, **104**, 1345-1353.

von Storch, H., U. Weese and J. S. Xu, 1990: Simultaneous analysis of space-time variability: principal oscillation patterns and principal interaction patterns with applications to the Southern Oscillation. *Zeitschrift fur Meteorologie*, **40**, 99-103.

Wallace, J. M., C. Smith and C. S. Bretherton, 1992: Singular value decomposition of wintertime sea-surface-temperature and 500 mb height anomalies. *J. of Climate*, **5**, 561-576.

Xinhua, C. and T. J. Dunkerton, 1995: Orthogonal rotation of spatial patterns derived from singular value decomposition analysis. *Journal of Climate*, **8**, 2631-43.

Yulaeva, E. and J. M. Wallace, 1994: The signature of ENSO in global temperature and precipitation fields derived from the Microwave Sounding Unit. *J. Climate*, **7**, 1719-36.

Zhang, Y., J. R. Norris and J. M. Wallace, 1998: Seasonality of large-scale atmosphere-ocean interaction over the North Pacific. *Journal Of Climate*, **11**, 2473-2481.

Zwiers, F. and H. von Storch, 1990: Regime-dependent autoregressive time series modeling of the Southern Oscillation. *J. Climate*, **3**, 1347-63.

Zwiers, F. W. and H. von Storch, 1989: Multivariate recurrence analysis. *J. Climate*, **2**, 1538-53.