Chapter 5
Seeking Structure in Data

5.1 Introduction

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.

5.2 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 space-time array: Measurements of a single variable at M locations taken at N different times, where M and N are integers.
- A parameter-time array: Measurements of M variables (e.g. temperature, pressure, relative humidity, rainfall, . . .) taken at one location at N times.
- 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 LxK=M long, and then analyze the resulting (LxK)xN=MxN data matrix. (A vector is a matrix where one dimension is of length 1, e.g. an 1xN matrix is a vector).

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

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & . . & x_{1N} \\ x_{21} & x_{22} & . . & x_{2N} \\ x_{31} & x_{32} & . . & x_{3N} \\ . & . & . . & . \\ x_{M1} & x_{M2} & . . & x_{MN} \end{bmatrix} = X_{ij}, \quad i = 1, M; \quad j = 1, N \tag{5.1}$$

So the state space, which might represent spatial grid points or a set of biological or chemical measurements has a dimension of M, which is the first index. This state vector is observed N times, which is the second index. So a column of this matrix represents the state of the system, and the N different columns represent different samples or times of measurement. So M is the state space and N is the sampling space.

Here we have included the symbolic bold $\mathbf{X}$ to indicate a matrix and the subscript notation 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.

$$\mathbf{X}^T = N \overbrace{\begin{bmatrix} . . \\ . . \\ . . \end{bmatrix}}^{M} = X_{ji}, \quad j = 1, N; \quad i = 1, M \tag{5.2}$$

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.

$$\mathbf{X}\mathbf{X}^T = M \overbrace{\begin{bmatrix} . . . \\ . . . \end{bmatrix}}^{N} \overbrace{\begin{bmatrix} . . \\ . . \\ . . \end{bmatrix}}^{M} N = D_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.3}$$

Of course, in this multiplication, each element of the first row of $\mathbf{X}$ 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. This explains 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 M x M 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, and we have removed the time mean of the data, 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.

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = N \overbrace{\begin{bmatrix} . & . \\ . & . \\ . & . \end{bmatrix}}^{M} \overbrace{\begin{bmatrix} . & . & . \\ . & . & . \end{bmatrix}}^{N} M, \quad i = 1, N; \quad j = 1, N \tag{5.4}$$

Both of the dispersion matrices obtained by taking inner products of a data matrix with itself are symmetric matrices. They become covariance matrices, if we divide by the sample size.

$$\mathbf{X}\mathbf{X}^{\mathsf{T}}/N = \mathbf{C} = C_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.5}$$

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 used to study the temporal/spatial structure of data sets.

5.3 Empirical Orthogonal Functions

Suppose we wish to define a vectors $\mathbf{e}$ that has maximum similarity to a data set $\mathbf{X}$. To measure the similarity we can project the vector $\mathbf{e}$ onto the data set $\mathbf{X}$ as follows,

$$\mathbf{e}^{\mathsf{T}}\mathbf{X} = [e_1, e_2, ., ., e_M] \begin{bmatrix} x_{11} & x_{12} & . & . & x_{1N} \\ x_{21} & x_{22} & . & . & x_{2N} \\ x_{31} & x_{32} & . & . & x_{3N} \\ . & . & . & . & . \\ x_{M1} & x_{M2} & . & . & x_{MN} \end{bmatrix} = [q_1, q_2, ., ., q_N] = \mathbf{q} \tag{5.6}$$

We can make a scalar measure of the similarity by taking the inner product of this projection time series $\mathbf{q}$ with itself and dividing by the sample size, N, to make the measure of similarity independent of the sample size.

$$\mathbf{q}\mathbf{q}^{\mathsf{T}}/N = \mathbf{e}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{e}/N = \mathbf{e}^{\mathsf{T}}\mathbf{C}\mathbf{e} = \lambda \tag{5.7}$$

where λ is the similarity measure that we want to maximize, which has units of the square of $\mathbf{X}$. One final constraint we need to apply is to limit the length of $\mathbf{e}$, which we can do by requiring that it have a length of one, $\mathbf{e}\mathbf{e}^{\mathsf{T}} = 1$. The problem

$$\mathbf{e}^{\mathsf{T}}\mathbf{C}\mathbf{e} = \lambda, \quad \text{subject to,} \quad \mathbf{e}\mathbf{e}^{\mathsf{T}} = 1 \tag{5.8}$$

is the standard eigenanalysis problem, which can be applied to any symmetric matrix. Up to M such eigenvectors and eigenvalues can be found.

$$\mathbf{e_i}^{\mathsf{T}}\mathbf{C}\mathbf{e_i} = \lambda_i, \quad = 1, M, \tag{5.9}$$

The problem can be written in matrix form as follows.

$$\mathbf{E}^{\mathsf{T}}\mathbf{C}\mathbf{E} = \tag{5.10}$$

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

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & . & . & e_{1M} \\ e_{21} & e_{22} & . & . & e_{2M} \\ e_{31} & e_{32} & . & . & e_{3M} \\ . & . & . & . & . \\ e_{M1} & e_{M2} & . & . & e_{MM} \end{bmatrix} = E_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.11}$$

Each column in 5.11 is a distinct eigenvector, with the order chosen so that the first eigenvalue is the largest, so that the first eigenvector explains the most variance.

$$= \begin{bmatrix} \lambda_1 & 0 & . & . & 0 \\ 0 & \lambda_2 & . & . & 0 \\ 0 & 0 & \lambda_3 & . & 0 \\ . & . & . & . & . \\ 0 & 0 & . & . & \lambda_M \end{bmatrix} = \Lambda_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.12}$$

The set of eigenvectors, e_{ij}, and associated eigenvalues, λ_j, represent a coordinate transformation into a coordinate space where the covariance matrix $\mathbf{C}$ 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 relatively few coordinate directions, which correspond to characteristic structures in the original structure space.

The sum of all the eigenvalues is the variance, so that the eigenvalue/divided by the sum of the eigenvalues gives the fraction of variance explained by each EOF. Since the EOFs are ordered from largest eigenvalue to smallest, plotting the eigenvalues as a function of their index provides a measure of how effective the eigenanalysis has been in explaining the variance with a small number of structures.

5.3.1 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.

5.3.2 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 that maximizes the amount of variance in the dependent sample that can be explained with 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.

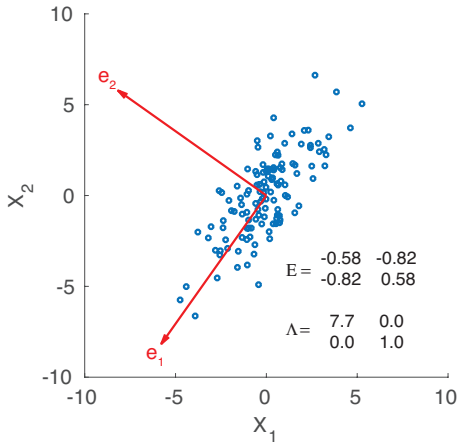

Figure 5.1 Scatter plot of two variables x_2 vs x_1. Eigenvectors are shown as red arrows and are elongated by a factor of ten for better viewing. Eigenvector matrix $\mathbf{E}$ and eigenvalue matrix Λ are shown as arrays of numbers.

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.

One area where EOF analysis is useful is in fitting a line to data. In ordinary least squares the line that is obtained depends on which variable is chosen to be the independent variable and which is chosen to be the dependent variable, so long as the data are not perfectly colinear. EOF analysis minimizes the perpendicular distance from the line and is not dependent on which variable is viewed as independent. Generally, both variables are noisy so that EOF line fitting is more robust.

5.4 Principal Components and EOFs

The EOFs form an orthogonal coordinate system that is a rotation of the original coordinate system. It is convenient to order the eigenvalues and eigenvectors in order of decreasing magnitude of the eigenvalue. The first eigenvector thus has the largest and explains the largest amount of variance in the data set used to construct the covariance matrix. In this new coordinate system, each data point is defined by a set of distances in the new coordinate system. To get these distances we project the eigenvectors onto the original data. We will call these new distances the Principal Components (PCs). To obtain the PCs we project the eigenvectors onto the original data, since the eigenvectors are directions defined in the original coordinate space. If we define the PC matrix as $\mathbf{Z}$, then the projection to obtain $\mathbf{Z}$ is a simple matrix multiplication.

$$\mathbf{Z} = \mathbf{E}^{\mathsf{T}}\mathbf{X} \tag{5.13}$$

Because the eigenvectors are orthogonal,

$$\mathbf{E}^{\mathsf{T}}\mathbf{E} = \mathbf{I} \tag{5.14}$$

where $\mathbf{I}$ is the identity matrix, with ones down the diagonal and zeros off the diagonal. Because of the orthogonality of the EOFs we can multiply (5.13) on the left by $\mathbf{E}$ and obtain an equation for the original data in terms of the PC matrix $\mathbf{Z}$.

$$\mathbf{X} = \mathbf{E}\mathbf{Z} \tag{5.15}$$

So it is easy (EZ) to get the original data back from the PCs since they are both expressions for the same data set, but in different reference frames or coordinate systems. The PC matrix $\mathbf{Z}$ has the same shape and information as the original data matrix $\mathbf{X}$ (5.1), but the information is expressed in a different coordinate system, which is defined by the EOF directions.

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & . . & z_{1N} \\ z_{21} & z_{22} & . . & z_{2N} \\ z_{31} & z_{32} & . . & z_{3N} \\ . & . & . . & . \\ z_{M1} & z_{M2} & . . & z_{MN} \end{bmatrix} = Z_{ij}, \quad i = 1, M; \quad j = 1, N \tag{5.16}$$

Now the columns represent the state of the system at different times, but in the new, more efficient, coordinate system.

5.4.1 Orthogonality of the Principle Components

Now we can easily show that the principle component time series, the time series of the amplitudes of each eigenvector, are uncorrelated in the sample space (*e.g.* time). The covariance matrix of the PCs is written as,

$$\mathbf{C}_Z = \mathbf{Z}\mathbf{Z}^T/N \tag{5.17}$$

Substituting in the expression for $\mathbf{Z}$ from (5.13), and using (5.10) we get,

$$\mathbf{C}_Z = \mathbf{Z}\mathbf{Z}^T/N = \mathbf{E}^T\mathbf{X}\mathbf{X}^T\mathbf{E}/N = \mathbf{E}^T\mathbf{C}\mathbf{E} = \tag{5.18}$$

From which we see that the covariance matrix of the PCs is also diagonal and equal to the eigenvalue matrix of the eigenanalysis.

The PCs are useful for prediction, since the PC time series are uncorrelated with each other, thus they share no variance between them. Truncating the EOF and PC representation of the data by removing EOFs that explain a small amount of variance may allow one to construct a set of predictors for which noise is reduced and the predictors are uncorrelated in the sample space. Noise is reduced if one assumes that the signal is correlated in space, so that eliminating EOFs that explain a small amount of variance is removing noise and not signal.

5.4.2 EOF Analysis via Singular Vector Decomposition

EOF/PC analysis can also be done by direct singular value decomposition of the data matrix $\mathbf{X}$ instead of doing an eigenanalysis of the covariance 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 $\mathbf{X}$ can be factored into

$$\mathbf{X} = \mathbf{U}\ \mathbf{V}^T \tag{5.19}$$

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

So we suppose that the data matrix $\mathbf{X}$ 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 and brevity let's call them space and time. Now $\mathbf{X}\mathbf{X}^\mathsf{T}$ is the dispersion matrix obtained by taking an inner product over time, leaving the covariance between spatial points. Thus the eigenvectors of $\mathbf{X}\mathbf{X}^\mathsf{T}$ are the spatial eigenvectors, and appear as the columns of $\mathbf{U}$ in the SVD. Conversely, $\mathbf{X}^\mathsf{T}\mathbf{X}$ 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 $\mathbf{V}$ 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 . 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 $\mathbf{X}$ 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

$$\mathbf{C} = \mathbf{U}\,\mathbf{V}^\mathsf{T} \tag{5.20}$$

and the eigenvectors and eigenvectors are defined by the diagonalization of $\mathbf{C}$.

$$\mathbf{C} = \mathbf{E}\,\mathbf{E}^\mathsf{T} \tag{5.21}$$

Now if we take the SVD of the data matrix, , and use it to compute the covariance matrix, we get,

$$\mathbf{C} = \mathbf{U}\,\mathbf{V}^\mathsf{T}(\mathbf{U}\,\mathbf{V}^\mathsf{T})^\mathsf{T})/\mathrm{N} = \mathbf{U}\,\mathbf{V}^\mathsf{T}\mathbf{V}\,^\mathsf{T}\mathbf{U}^\mathsf{T}/\mathrm{N} = \mathbf{U}\,\,^\mathsf{T}\mathbf{U}^\mathsf{T}/\mathrm{N} \tag{5.22}$$

Comparing 5.21 and 5.22 one can infer that $\mathbf{U} = \mathbf{E}$ and $= \,^\mathsf{T}/\mathrm{N}$ or $\lambda_\mathrm{i} = \sigma_\mathrm{i}^2/\mathrm{n}$. The singular values represent amplitudes across $\mathbf{X}$, and the eigenvalues represent variance.

We can also see that $\mathbf{V}$ represents normalized PC time series in the following way,

$$\mathbf{Z} = \mathbf{E}^\mathsf{T}\mathbf{X} = \mathbf{E}^\mathsf{T}\mathbf{U}\,\mathbf{V}^\mathsf{T} = \mathbf{E}^\mathsf{T}\mathbf{E}\,\mathbf{V}^\mathsf{T} = \,\mathbf{V}^\mathsf{T} \tag{5.23}$$

Here we have used 5.13, 5.14 and 5.19.

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,

$$\mathbf{Z}\mathbf{Z}^\mathsf{T} = \mathbf{E}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{E} = \mathbf{E}^\mathsf{T}\mathbf{C}\mathbf{E}\mathrm{N} = \,\mathrm{N} \tag{5.24}$$

while in the SVD computation,

$$\mathbf{Z}\mathbf{Z}^\mathsf{T} = \,\mathbf{V}^\mathsf{T}\mathbf{V}\,^\mathsf{T} = \,\,^\mathsf{T} \tag{5.25}$$

so we must have, as show before, that,

$$^\mathsf{T} = \,\mathrm{N} \quad \text{or} \quad \sigma_\mathrm{i}^2 = \lambda_\mathrm{i}\mathrm{N} \tag{5.26}$$

From **(5.24)** and **(5.25)** we see that the covariance matrix of the principle components is diagnonal, and the principle components are orthogonal (uncorrelated) in the structure dimension.

5.4.3 A very simple example

Consider the following simple 4x2 data set. Imagine that the structure dimension is 2 and the sampling dimension is 4.

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & -6 & 8 \\ 1 & 2 & -3 & 4 \end{bmatrix}$$

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

$$\mathbf{U}\,\mathbf{V}^{\mathsf{T}} = \mathbf{X} \tag{5.27}$$

The singular value matrix contains only one non-zero 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 data points in mathbfX all fall on the same line. The singular value contains all of the amplitude information. The spatial and temporal singular vectors are both of unit length.

$$= \begin{bmatrix} 12.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Next U, which contains the spatial singular vector in the first column.

$$\mathbf{U} = \begin{bmatrix} 0.894 & -0.447 \\ 0.447 & 0.894 \end{bmatrix}$$

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

$$\mathbf{V} = \begin{bmatrix} 0.183 & -0.119 & -0.976 & 0.000 \\ 0.365 & -0.239 & 0.098 & -0.894 \\ -0.548 & -0.837 & 0.000 & 0.000 \\ 0.730 & -0.478 & 0.195 & 0.447 \end{bmatrix}$$

We can reconstruct the data matrix by first multiplying the singular value matrix times the transpose of the temporal variation matrix to form a traditional PC matrix.

$$\mathbf{Z} = \mathbf{V}^{\mathsf{T}} \tag{5.28}$$

$$\mathbf{Z} = \begin{bmatrix} 2.236 & 4.472 & -6.7082 & 8.944 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

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 a row of zeros. The first row of **Z** is the principal component vector for the first EOF, 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 matrix. This is equivalent to multiplying the eigenvector matrix times the PC matrix, and gives us the original data back.

$$\mathbf{X} = \mathbf{U}\mathbf{Z} \tag{5.29}$$

$$\mathbf{X} = \begin{bmatrix} 2.00 & 4.00 & -6.00 & 8.00 \\ 1.00 & 2.00 & -3.00 & 4.00 \end{bmatrix}$$

5.5 Presenting the Results of EOF and PC Analysis

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, standardize 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.

5.5.1 How to scale and plot EOFs and PCs

Let's suppose we have done EOF/PC analysis using either the SVD of the data, or the eigenanalysis of the covariance matrix. We next want to plot the EOF's to show the structure in the state space of the data. The EOFs are normalized to unit length, but we would like to combine this structure with some amplitude information in a single plot. One way to do this is to scale the eigenvectors according to the amplitude in the data set that they represent. A simple way to do this is to multiply the eigenvectors by the square root of the eigenvalue. We learned in the previous section that,

$$\mathbf{E} = \mathbf{U} \quad \text{and} \quad = \Sigma^2/\mathsf{N} \tag{5.30}$$

We define the EOF with amplitude as $\mathbf{D}$.

$$\mathbf{D}^{\mathsf{EOF}} = \mathbf{E}\,^{1/2} = \mathbf{D}^{\mathsf{SVD}} = \mathsf{N}^{-1/2}\mathbf{U} \tag{5.31}$$

In each case you can show that $\mathbf{D}\mathbf{D}^{\mathsf{T}} = \mathbf{C}$, so if you put in the amplitudes and take the inner product you get back the covariance matrix of the input data. Note also that $\mathbf{D}^{\mathsf{T}}\mathbf{D} =$. The columns of the matrix $\mathbf{D}$ are the eigenvectors, scaled by the amplitude that they represent in the original data. It might be more interesting to plot $\mathbf{D}$ than $\mathbf{E}$, because then you can see how much amplitude in an RMS sense is associated with each EOF, and you can plot the patterns in hectoPascals, °C, kg, or whatever units in which the data are given. These methods work if the data analyzed is dimensional, if the data have first been standardized, then the principle components can be regressed onto the un-standardized data to get structures with dimensional amplitudes.

In most cases the sample mean of each state variable would be removed before doing EOF or PC analysis. Sometimes it is also advisable to take the amplitudes out of the data before conducting the EOF analysis by dividing each observation by its standard deviation over the sample. Subtracting the mean and dividing by the standard deviation can be called standardizing the data, and we can denote the standardized data set as $\widetilde{\mathbf{X}}$. Reasons for standardizing 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, rather than just an efficient expression of variance, you may want to look at correlation rather than covariance. In such cases the data

can be standardized such that the variance of the time series of each element of the state vector is 1. The covariance matrix of this standardized data set is a correlation matrix.

If the data have been standarized, we can still get the amplitude into the structure by regressing the principle components of the EOF analysis onto the original dimensional data. Regression can also be used to see how the EOF of the state variable is related to other variables not in the state vector used to do the EOF analysis.

To do the regression analysis, we first want to get the PC time series normalized so that they have unit variance in time. Regardless of whether the original data were standarized or not, we can obtain a standardized principle component time series by dividing by the square root of the eigenvalue.

$$\widetilde{\mathbf{Z}} = {}^{-1/2}\mathbf{Z} \tag{5.32}$$

As an exercise, show that

$$\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\mathsf{T}}/\mathsf{N} = \mathbf{I} \tag{5.33}$$

The regression to determine $\mathbf{D}$ is then,

$$\mathbf{D} = \mathbf{X}\widetilde{\mathbf{Z}}^{\mathsf{T}}/\mathsf{N} \tag{5.34}$$

The columns of the matrix $\mathbf{D}$ are the EOFs, 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. This is a reasonable number to look at, since it is the amplitude that you might typically see associated with this EOF.

5.6 Significance of EOF Analysis

EOF/PC analysis makes sense when the data contain a lot of correlation between the elements of the state space, so that the variance in the data can be explained with a number of EOFs that is smaller than the number of elements in the state vector $\mathbf{X}$. The fraction of variance explained is measured by the eigenvalue λ_i divided by the sum of the eigenvalues, since the sum of the M eigenvalues is the total variance.

$$\mathsf{FoV}_i = \frac{\lambda_i}{\sum_{m=1}^{M} \lambda_i} \tag{5.35}$$

For a random state vector sampled a finite number of times, however, it is likely that EOF analysis will by chance find some EOFs that explain more variance than others, even if the state vectors are uncorrelated with each other. This happens due to random chance, but the EOF analysis orders the eigenvalues from largest to smallest, and so it appears that the data is structured, when if fact it is not. When the analysis is performed on another sample from the same uncorrelated data set, a different set of EOFs are found that appear, by chance, to explain more variance than expected for an uncorrelated data set.

This can be illustrated with a simple example. Suppose we consider a state vector of Gaussian white noise that has no correlation in the state space (M=20) or the sample space N. We consider samples of size N=20, 50, 200, 5,000. Since each eigenvalue spectrum will be a little different due to sampling variations, we do this 1000 times and average the eigenvalue spectra to give the most likely spectrum (**Fig. 5.2**). The true eigenvalue spectrum is uniform with a each eigenvalue equal to one. A large sample of N=5,000, nearly produces a uniform eigenvalue spectrum, but for smaller samples the eigenvalue spectrum is strongly sloped, indicating that the first eigenvector can explain much more variance than the last one. This happens because for any small sample some structure will explain a lot of the variance by chance. This structure will be assigned the first position. Each time this is done the EOF that explains the most variance is different, however, so the large explained variance by the first EOF is not robust or meaningful.

5.6.1 The North Test

North et al. (1982) suggested a "rule-of-thumb" for assessing the statistical significance of the eigenvalue spectrum. For a Gaussian distribution the eigenvalues should fall within a range given by a standard error that depends on the eigenvalue, λ and the number of independent samples, N.

$$\Delta\lambda = \lambda\left(\frac{2}{N}\right)^{1/2} \tag{5.36}$$

They argued that 68 percent of sample eigenvalues should fall within these limits, and that if the uncertainties of two adjacent eigenvalues overlapped, then eigenvalues are not really distinct and the EOFs associated with those eigenvalues will show large inter-sample variability, would not be robust, and should not be taken seriously. **Fig. 5.2** includes the the North et al. (1982) eigenvalue uncertainty bars. In each case the uncertainties of adjacent eigenvalues overlap, so we conclude that none of the eigenvalues or eigenvectors are meaningful. We could illustrate the inter-sample variation in the eigenvectors by plotting the eigenvector that explains the most variance for each of the 1000 realizations associated with the cases in **Fig. 5.2** and this would show that they are random in appearance. The rule of thumb expressed by **(5.36)** correctly suggests that the eigenvalues are not distinct for all the cases shown in **Fig. 5.2**.

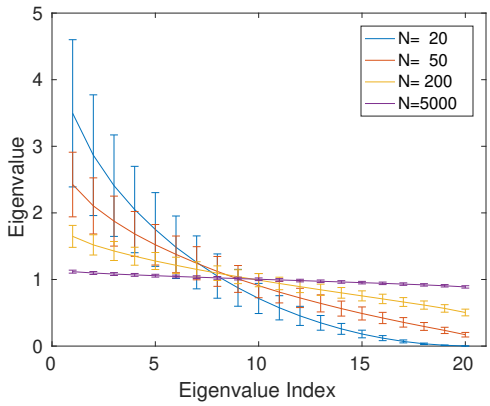

Figure 5.2 Eigenvalue spectra for white noise in a state space of M=20 with different sample sizes N=20, 50, 200, and 5,000. The North et al. (1982) uncertainty estimates are shown for each spectrum.

When structure is present in the data, the first few eigenvalues will set themselves apart from the rest, and the error bars suggested by North et al. (1982) will not overlap with the others as shown in **Fig. 5.3**.

5.6.2 Assessing Physical Significance

If the North criterion is satisfied for a particular analysis, one must still be careful in interpreting the 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 and sample of interest as efficiently as possible, and also be orthogonal with each other. Sometimes these mathematical constraints will select 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 often tend to look wavelike, because they are constrained to be orthogonal. If you put autocorrelated 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

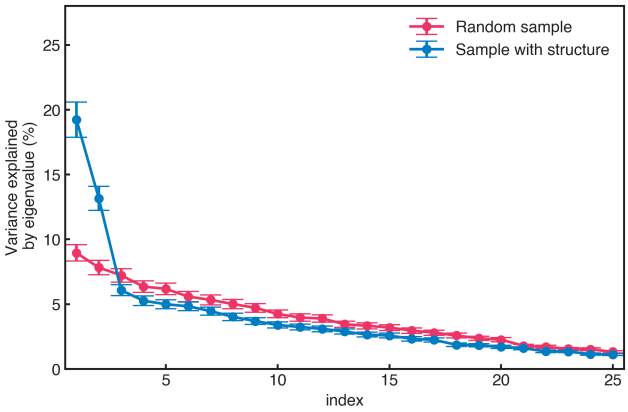

Figure 5.3 Eigenvalue spectrum of white noise (purple) and data with structure (blue) for a state space of $M = 25$ and sample of size $N = 100$.

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. Sometimes, in an effort to explain the variance of a finite sample, EOF analysis will combine distinct physical modes into a single EOF, which may be called mode mixing. 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? Do you expect adjacent data points to be correlated? Can the EOFs of interest support a rejection of the uninteresting null hypothesis that adjacent points are correlated?
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?

5.7 Applications of EOF/PC Analysis

Rearrangement of data into Empirical Orthogonal Functions (EOFs) and their Principal Components (PCs) is useful in a variety of contexts.

5.7.1 Data Compression

EOF/PC analysis is a kind of functional representation, where a set of spatial/structure functions is derived that explains the largest amount of variance with the smallest number of functions. The functions are

arranged according to their rank in explaining variance. EOFs 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, if the PCs of EOFs that explain a small amount of variance are discarded.

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 represented well by a small range of Fourier wavelengths. 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 structures. This is indicated by the eigenvalue spectrum. If the first few eigenvalues are large and most of the rest are small, then the number of degrees of freedom of the data set is much less than the number of data points, and use of EOFs to compress the data can provide benefits.

5.7.2 Determining Degrees of Freedom

If a spatiotemporal data set has large correlations between the state-space variables, then the data set may have fewer independent degrees of freedom than the number of state-space variables (*e.g.* spatial grid points or parameters). For many purposes it is important to evaluate how many independent degrees of freedom a data set has. This can be assessed using EOF analysis, as has been reviewed by (Bretherton et al., 1999).

Consider a spatial data set of dimension m that is stationary on the time interval for which it is sampled. Define a quadratic functional of some vector variable $\mathbf{X}(t)$, where the vector is of length m.

$$E(t) = \sum_{i=1}^{m} \mathbf{X}_i^2(t) \tag{5.37}$$

The number of spatial degrees of freedom m^* is defined to be the number of uncorrelated random normal variables a_k, each having zero mean and the same population variance $\langle a^2 \rangle$, for which the χ^2 distribution for the specified functional most closely matches the PDF of $E(t)$. In order to approximate this one can require that the χ^2 distribution match the observed distributions ensemble mean value $\langle E \rangle$ and the temporal variance about this mean,

$$var(E) = \langle E'^2 \rangle = \left\langle \left(E - \langle E \rangle \right)^2 \right\rangle \tag{5.38}$$

For the χ^2 distribution $\langle E \rangle = m^* \langle a^2 \rangle$ and $var(E) = 2m^* \langle a^2 \rangle^2$. We can then solve for the spatial degrees of freedom that matches the first two moments of the normal distribution of variance. This is a "moment matching" estimate of the effective number of degrees of freedom.

$$m^*_{mm} = \frac{2\langle E \rangle^2}{var(E)} \qquad \langle a^2 \rangle^2_{mm} = \frac{var(E)}{\langle E \rangle} \tag{5.39}$$

These estimates can be obtained from the mxm covariance matrix of $\mathbf{X}$, $\mathbf{C}_{xx} = \mathbf{C}$, if $\mathbf{X}(t)$ is normally distributed and we know $\mathbf{C}$ well enough. Suppose we have the eigenvalues λ_k and the standardized principle components $z_k(t)$ of $\mathbf{C}$. We can now calculate m^* from the eigenvalues in the following way.

$$E(t) = \sum_{k=1}^{m} \lambda_k z_k^2(t) \qquad \langle E \rangle = \sum_{k=1}^{m} \lambda_k \tag{5.40}$$

and

$$\begin{aligned} var(E) = \sum_{k=1}^{m} \lambda_k^2 var(z_k^2(t)) &= \sum_{k=1}^{m} \lambda_k^2 \left\langle var\left(z_k^2 - \langle z_k^2 \rangle\right)^2 \right\rangle \\ &= \sum_{k=1}^{m} \lambda_k^2 \left\langle z_k^4 - \langle z_k^2 \rangle^2 \right\rangle \end{aligned} \tag{5.41}$$

Since we are assuming that the PCs are standardized Gaussian normal variables their variance is one and their kurtosis is 3, and we have that,

$$var(E) = \sum_{k=1}^{m} \lambda_k^2 \left\langle z_k^4 - \langle z_k^2 \rangle^2 \right\rangle = \sum_{k=1}^{m} \lambda_k^2 (3-1) = 2 \sum_{k=1}^{m} \lambda_k^2 \tag{5.42}$$

We can now write down an eigenvalue based estimate for the effective number of spatial degrees of freedom by substituting 5.40 and 5.42 into 5.39.

$$m_{eff}^* = \frac{\left(\sum_{k=1}^{m} \lambda_k\right)^2}{\sum_{k=1}^{m} \lambda_k^2} = \frac{\left(m\bar{\lambda}\right)^2}{m\overline{\lambda^2}} \tag{5.43}$$

This formula can also be written in terms of the covariance matrix from which the eigenvalues were derived.

$$m_{eff}^* = \frac{\left(\sum_{i=1}^{m} C_{ii}\right)^2}{\sum_{i=1}^{m} \sum_{j=1}^{m} C_{ij}^2} = \frac{\left(\operatorname{tr} \mathbf{C}\right)^2}{\operatorname{tr}\left(\mathbf{C}^2\right)} \tag{5.44}$$

Here tr indicates the trace of the matrix. Note that the denominator in **(5.43)** equals the square of the Frobenius norm of **C**.

5.7.3 Prefiltering

It might be reasonable to assume that EOFs with large eigenvalues are the structure in a data set and EOFs with small eigenvalues are noise. One could argue then that doing EOF analysis and removing the variance associated with small eigenvalues is a noise reduction procedure. For example, when reconstituting the original data from the EOF expansion, the PC time series associated with the smaller eigenvalues could be set to zero. This would produce a data set with the effects of those less correlated structures removed.

5.7.4 Statistical Prediction

EOFs are orthogonal in the structure dimension and in the sampling dimension. In statistical prediction, correlation between predictors is undesirable [reference earlier section]. PC series are uncorrelated in the sampling dimension, and the covariance of the PC matrices is diagnonal with the eigenvalues down the diagonal.

$$\mathbf{Z}\mathbf{Z}^T/N = \mathbf{C}_{ZZ} = \tag{5.45}$$

It is thus advantageous to first do EOF analysis and use the principal components of the EOFs as predictors rather than any other combination of the predictor variables. This makes the predictors uncorrelated and the inversion to obtain the regresson coefficients much easier.

5.7.5 Exploratory Data Analysis

Suppose we have a state vector of some system for which we have a large sample. EOF/PC 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. If the data set shows that much of the variance can be explained by a few EOFs, then the analysis can be focused by limiting the analysis to these few structures and their variations within the sample.

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

$$w(x,t) = \cos(2\pi x/L) \times \cos(2\pi t/T) \tag{5.46}$$

The EOF analysis will show one nonzero eigenvalue corresponding to a wave with wavelength L in space and period T in time. A traveling wave has a formula as **(5.47)**.

$$\begin{aligned} w(x,t) &= \cos(2\pi x/L - 2\pi t/T) \\ &= \cos(2\pi x/L) \times \cos(2\pi t/T) + \sin(2\pi x/L) \times \sin(2\pi t/T) \end{aligned} \tag{5.47}$$

Representing a traveling wave requires two EOFs and corresponding PCs, each 90-degrees out of phase. These would have the same eigenvalue, since their amplitudes are equal.

5.8 Rotation of Empirical Orthogonal Functions

EOF analysis enforces orthogonality on the eigenvectors and their corresponding principal components. Sometimes the orthogonality constraint will cause structures to have significant amplitude all over the domain (e.g. spatial domain) of the analysis, when physical reasoning suggests that the structures should 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).

The procedure begins by selecting a subset of the eigenvectors, say those that explain 70% of the variance, and discarding the rest. The constraint of orthogonality is relaxed and replaced with another constraint that is designed to make the EOFs as "simple" as possible. Relaxing one orthogonality is an orthogonal rotation, and relaxing both orthogonalities is an oblique rotation. Simplicity of structure is defined to occur when most of the elements of the eigenvector are either of order one (absolute value) or zero, but not in between. The selected eigenvectors are rotated until the criterion is maximized.

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

$$Q = \frac{1}{Mm} \sum_{i=1}^{M} \sum_{j=1}^{m} \left(b_{ij}^2 - \overline{b^2} \right) \tag{5.48}$$

where

$$\overline{b^2} = \frac{1}{Mm} \sum_{i=1}^{M} \sum_{j=1}^{m} b_{ij}^2 \tag{5.49}$$

The quantity **(5.48)** to be maximized can be simplified to,

$$Q = \frac{1}{Mm} \sum_{i=1}^{M} \sum_{j=1}^{m} b_{ij}^2 - \overline{b^2} \tag{5.50}$$

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.

One often used criterion for defining simplicity is the Varimax Method. The simplicity of an individual rotated eigenvector is defined as the variance of its squared loadings, b_{ij}, where i is the loading index for an eigenvector and j denotes the eigenvector.

$$V_j = \frac{1}{M}\sum_{i=1}^{M}\left(b_{ij}^2\right)^2 - \frac{1}{M^2}\left(\sum_{j=1}^{M} b_{ij}^2\right)^2 \qquad j = 1, 2, ..., m \tag{5.51}$$

When the variance V_j is at a maximum the jth rotated eigenvector has its greatest simplicity in the sense that its loading tends toward unity or zero. The criterion of simplicity of the complete rotated eigenvector matrix is defined as the maximization of the sum of the simplicities of the individual eigenvectors,

$$V = \sum_{j=1}^{m} V_j \tag{5.52}$$

Equation (5.52) is called the raw Varimax criterion. It is sometimes useful to weight the individual eigenvectors with a weight h_j, for example by the variance explained. The final normalized Varimax criterion is then,

$$V = M\sum_{j=1}^{m}\sum_{i=1}^{M}\left\{\frac{b_{ij}}{h_j}\right\}^4 - \sum_{j=1}^{m}\left\{\sum_{i=1}^{M}\frac{b_{ij}^2}{h_j}\right\}^2 \tag{5.53}$$

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. Many other criteria closely related to these can be found in available software packages.

5.8.1 The Eight Physical Variables Example

An interesting example of the use of EOF rotation in factor analysis is the 'Eight Physical Variables Example' that looks at the correlations between eight measures of human anatomy. The eight physical variables are listed in the table below.

Eight Physical Variables	
1. Height	5. Weight
2. Arm Span	6. Bitrochantric Diameter
3. Forearm Length	7. Chest Girth
4. Lower Leg Length	8. Chest Width

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.

$$\mathbf{C} = \begin{bmatrix} 1.00 & .846 & .805 & .859 & .473 & .398 & .301 & .382 \\ .846 & 1.00 & .881 & .826 & .376 & .326 & .277 & .415 \\ .805 & .881 & 1.00 & .801 & .380 & .319 & .237 & .345 \\ .859 & .826 & .801 & 1.00 & .436 & .329 & .327 & .365 \\ .473 & .376 & .380 & .436 & 1.00 & .762 & .731 & .629 \\ .398 & .326 & .319 & .329 & .762 & 1.00 & .583 & .577 \\ .301 & .277 & .237 & .327 & .731 & .583 & 1.00 & .539 \\ .382 & .425 & .345 & .365 & .629 & .577 & .539 & 1.00 \end{bmatrix}$$

The correlation matrix shows that all the physical variables are positively correlated. The eigenvalue spectrum derived from eigenanalysis of the correlation matrix of the eight physical variables shows that the

first two eigenvectors explain 85% and 12% of the correlation, respectively, so that only these two vectors are considered and included in the rotation.

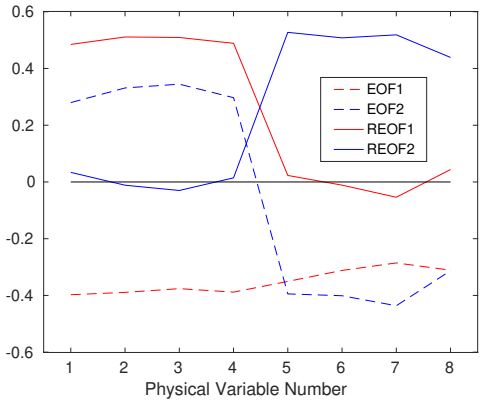

Figure 5.4 Eigenvectors and orthogonally rotated eigenvectors for the Eight Physical Variables data set. Varimax rotation was used.

Fig. 5.4 shows the eigenvectors derived from the correlation matrix of the eight physical variables. The rotated EOFs obtained from an orthogonal varimax rotation are also shown. The first eigenvector indicates that all the physical variables tend to go up and down together, so long people are also broad and heavy. The second eigenvector contains the implication that the people who are long, also tend to be thin. These indications are somewhat contrary to the basic correlation matrix that shows the correlation between the first 4 and the last 4 physical variables are positive, but weak. The structure of the eigenvectors is heavily constrained by the requirement of orgthogonality. Rotation of the eigenvectors gives an alternative interpretation. The rotated eigenvectors lead to an interpretaion of the data in which one factor is length, or the "bone factor", and a second factor is related to width and weight, or the "flesh factor". This interpretation of the data recognizes the strong positive correlation among the length variables and among the weight variables, but does not impose an artificial negative correlation between the length and weight variables. The second interpretation seems more acceptable, and explains the data just as well. In this case the bone and flesh factors are still orthogonal in the parameter domain, since an orthogonal rotation was used.

5.8.2 EOF Analysis of Red Noise

Another example of the use of rotated EOFs is red noise that is autocorrelated in space and time, but has no other structure. Suppose we take a sample of such a space-time series that has 50 spatial grid points and 400 sample times. The autocorrelation in space is 0.9 and the autocorrelation in time is 0.2, so that all the samples are essentially independent. We perform EOF analysis on the standardized data, so that each point has unit variance in time. Before doing the EOF analysis, though, let's consider in **Fig. 5.5**a a one point correlation map between grid point 25 and all the other grid points. This plot indicates that grid point 25 is well correlated with adjacent data points, but not correlated at all with distant points. Now let's do EOF analysis of this data set.

The spectrum of eigenvalues of our red noise data set is shown in **Fig. 5.5**b. Most of the variance is explained by the first few eigenvectors, and the North Test suggests that the largest eigenvalues are distinct and the associated eigenvectors should be robust and reproducible. This is true because the data is highly autocorrelated in space, so that slowly varying functions should explain a lot of the variance. Beyond telling us that the data are correlated in space, however, these eigenvectors are not physically meaningful, since we

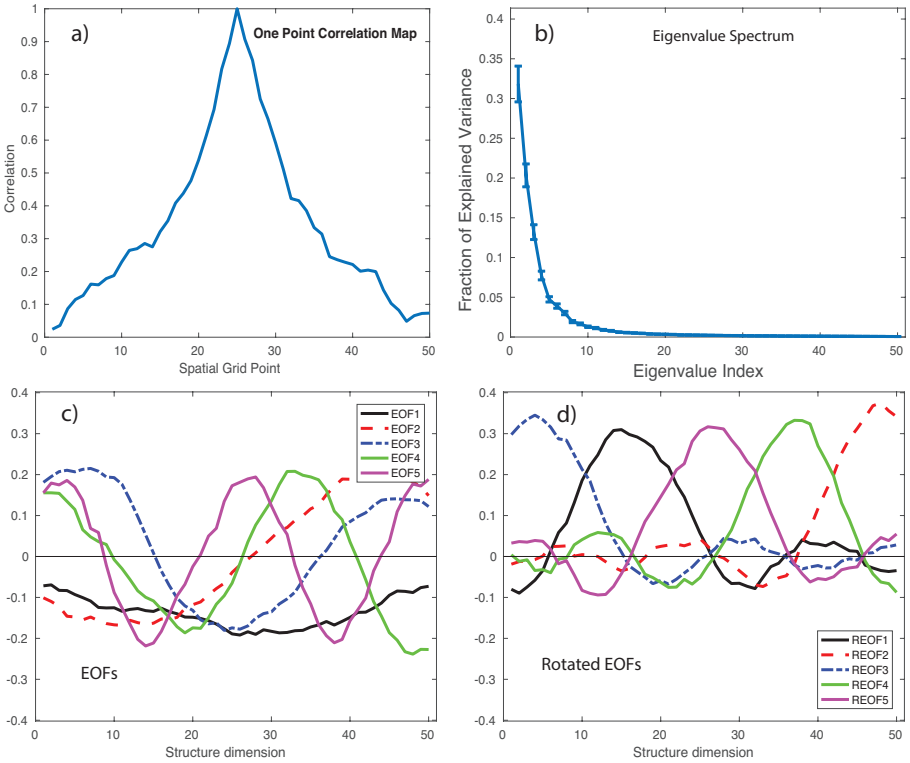

Figure 5.5 a) Autocorrelation of grid point 25 with all the other grid points of a sample of red noise with a spatial autocorrelation of 0.9. b) Eigenvalue spectrum for red noise data set with North uncertainty bars. c) First five eigenvectors. d) First five rotated eigenvectors. An orthogonal varimax rotation was used.

know the data are just red noise in space. An assessment of the effective number of degrees of freedom in this data set using **(5.43)** indicates that the data set has fewer than 6 spatial degrees of freedom despite having 50 grid points.

The raw eigenvectors are shown in **Fig. 5.5**c. The first eigenvector is of one sign everywhere, like a constant first term in a Fourier series. It suggests that all the points go up and down together, but we know from **Fig. 5.5**a that the data at distant points are in fact not correlated. The second eigenvector looks like a sine wave with a wavelength equal to the size of the model domain. Here EOF analysis is just constructing the functions of a Fourier series that is orthogonal within the spatial domain.

The retained eigenvectors that explain 70% of the total variance were rotated using an orthogonal Varimax criterion. The rotated eigenvectors shown in **Fig. 5.5**d are localized in space and span a region that is similar in scale to the autocorrelation shown in **Fig. 5.5**a. In this case then, the rotated EOFs are much more representative of the data than are the raw EOFS. They show a sequence of blobs that are localized in space, but near zero elsewhere, which is a good representation of red noise. These rotated EOFs are also orthogonal in the space dimension, so there is no shared variance between them. They each explain about an equal fraction of the total variance. The negative values could be brought much closer to zero by using an oblique rotation.

5.8.3 Wintertime 500hPa Height Example

As a more realistic example of the use of rotation of EOFs, consider the wintertime anomalies of 500hPa height with time scales longer than 30 days. To perform this analysis we first remove the climatological annual cycle, or it would dominate the analysis. We want to explore the structure of the slowly varying anomalies

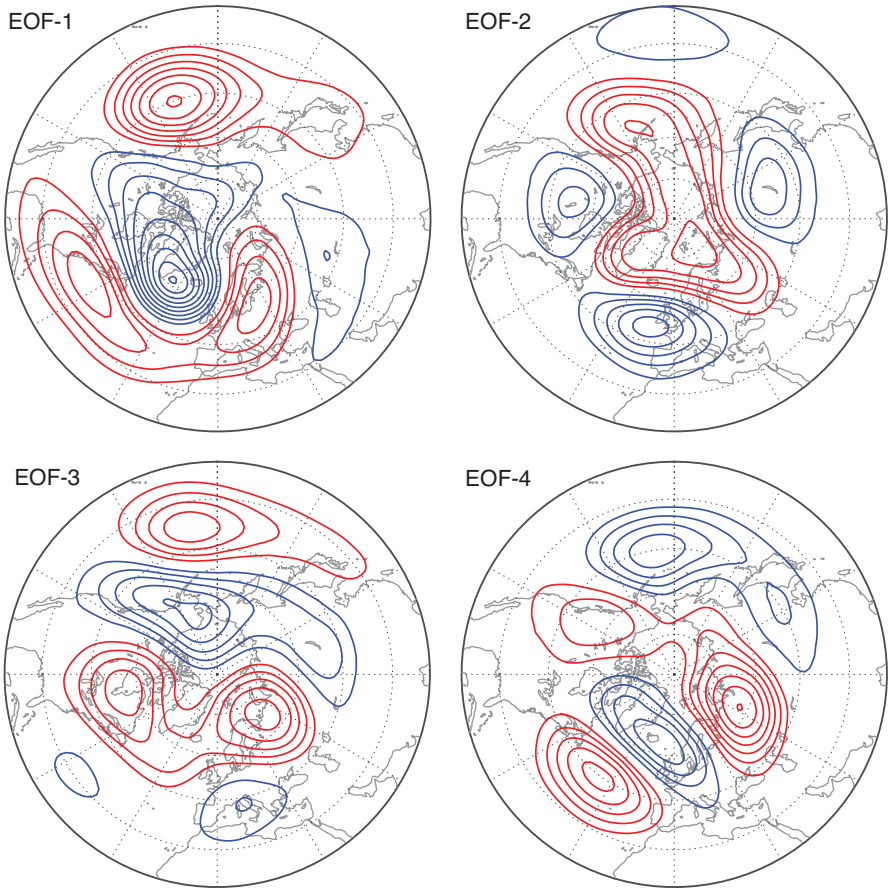

Figure 5.6 First four EOFs of the wintertime 500hPa height anomalies north of 20N.

of the 500hPa height field, which characterizes the structure of the mid-troposphere. To focus on the winter season we consider only the period from October 1 to March 31. To focus on the Northern hemisphere, we consider only the region poleward of 20N. Because the data are in latitude-longitude coordinates, we weight the variance by cosine of latitude, so so that the more closely spaced data in high latitudes do not bias the analysis toward high latitudes. The raw anomalies are used, unstandardized, so that regions of large variance have a bigger impact on the structures obtained.

The eigenvalue spectrum is smoothly decreasing with eigenvalue number, and 70% of the variance is explained by the first 11 eigenvectors, indicating that the low-frequency variability of the 500hPa height is strongly autocorrelated and of large spatial scale compared to the grid spacing of the basic data. The first 11 eigenvectors are rotated, but for economy of presentation we show only the first four in **Fig. 5.6**. These generally have some amplitude everywhere in the Northern Hemisphere, and tend to show strong connections between the Atlantic and the Pacific Ocean basins.

When the eigenvectors are rotated, they become more local and tend to look more meteorological (**Fig. 5.7**). The first rotated EOF corresponds to the North Atlantic Oscillation, which was discovered in local correlation maps before good global analyses were available. The second and third rotated EOFs correspond to known patterns that reflect the propagation of Rossby waves along great circle routes. The fourth rotated EOF appears to be a Rossby wave train propagating across Eurasia. In this case the rotation to produce more localized structures leads to a much more sensible physical interpretation than the raw EOFs, which try very hard to maximize the variance explained over the whole domain with a set of orthogonal structures. Rotation produces structures more similar to what you get from one-point correlation maps.

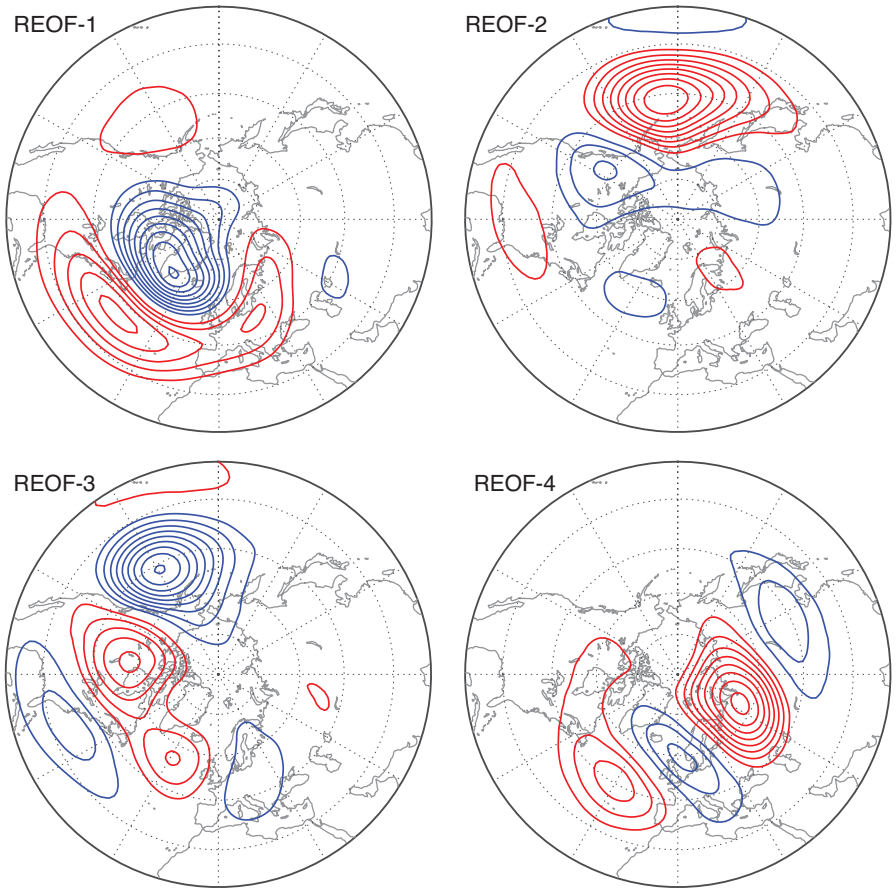

Figure 5.7 First four rotated EOFs of the wintertime 500hPa height anomalies north of 20N. Orthogonal varimax rotation was used.

5.9 Maximum Covariance Analysis

EOF analysis searches for structures that explain the maximum amount of variance in some data matrix. The data matrix is presumed to have a structure dimension, for example spatial location, and a sampling dimension, for example time. In Maximum Covariance Analysis (MCA) two data matrices with different structures, or state spaces, are considered, but they share a common sampling dimension. For example, one could consider the fields of sea surface temperature and surface chlorophyll content, measured at the same set of times. Or the sampling dimension could be a collection of hospital patients, and the two state vectors could be their "Eight Physical Variables" and their cholesterol data (say, 3 numbers). Suppose the sample of patients is N. One could address the relationship between body dimensions and cholesterol by making an augmented state vector consisting of the 11 numbers that include their 8 physical variables and their 3 cholesterol values. EOF analysis of the 11 x N data matrix might determine if particular combinations of variables vary together and in that way explain a lot of the combined variance. One might expect that if the physical variables and the cholesterol variables vary together, then they should show up in structures that efficiently explain the variance of the augmented or combined state vector. This is a form of augmented EOF analysis that tries to explain the maximum amount of variance over a combined data set. Where the variables have different units, one would normally standardize the data to unit variance before the conducting the EOF analysis.

In MCA one looks for structures in the data set that are well correlated with structures in the other data set. To analyze the relationship between the eight physical variables and the cholesterol data using MCA, one first computes the covariance (or correlation) matrix between the 8 x N and 3 x N data sets to make an

8x3 covariance matrix. Then SVD analysis of this covariance matrix yields structures that are well correlated between body dimensions and cholesterol levels, if such exist. 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? The singular values tell you the amount of covariance that is explained by each pair of structures.

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

5.9.1 MCA Mathematics

Let us suppose we have two data matrices $\mathbf{X}$ and $\mathbf{Y}$ 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.

$$\mathbf{X}\mathbf{Y}^{\mathsf{T}}/N = \mathbf{C}_{XY} \tag{5.54}$$

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. If the two fields have different units, then the correlation matrix should be used.

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.

$$\mathbf{C}_{XY} = \mathbf{U}\,\mathbf{V}^{\mathsf{T}} \tag{5.55}$$

The columns of $\mathbf{U}$ (MxM)are the column space of $\mathbf{C}_{XY}$ and represent the structures in $\mathbf{X}$ that are highly correlated with $\mathbf{Y}$. The columns of $\mathbf{V}$ are the row space of $\mathbf{C}_{XY}$ and are those structures in the Y space that best explain the covariance. The singular values are down the diagonal of the matrix and give the amount of covariance explained by each pair of left and right singular vectors. 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 $\mathbf{X}$ and $\mathbf{Y}$. The number of non-zero singular values will be less than or equal to the smaller of M or L, $k \leqslant K \leqslant M \cap L$.

$$\|\mathbf{C}_{XY}\|^2 = \sum_{i=1}^{M}\sum_{j=1}^{L}\left(\overline{x_i y_j}\right)^2 = \sum_{k=1}^{K}\sigma_k^2 \tag{5.56}$$

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.

As in EOF analysis we can project the left an right singular vectors onto the data to express the initial data in the coordinate space of the optimal directions for expressing covariance.

$$\mathbf{X}^* = \mathbf{U}^{\mathsf{T}}\mathbf{X} \quad ; \quad \mathbf{Y}^* = \mathbf{V}^{\mathsf{T}}\mathbf{Y} \tag{5.57}$$

Using (5.54), (5.54) and the orthonormality of the singular vectors, it is easy to show that the covariance matrix of the amplitude loadings of the left and right singular vectors computed in (5.56) is diagonal and equal to the singular value matrix of the original covariance matrix.

$$\mathbf{C}_{X^*Y^*} = \mathbf{X}^*\mathbf{Y}^{*T}/N = \tag{5.58}$$

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 total 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.

5.9.2 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 and call it the normalized root mean squared covariance. 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 the covariance using MCA.

$$RMSC = \left(\frac{\sum_{i=1}^{M} \sum_{j=1}^{L} \overline{x_i y_j}^2}{\left(\sum_{i=1}^{M} \overline{x_i^2} \right) \left(\sum_{j=1}^{L} \overline{y_j^2} \right)} \right)^{1/2} \tag{5.59}$$

The normalized root mean square covariance should be on the order of 0.1 or greater to indicate well-correlated fields so that MCA is justified.

5.9.3 Heterogeneous and Homogeneous Regression Maps

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.

As in EOF/PC analysis, the dimensional amplitude of the left and right singular vector patterns can be obtained by regressing the original data onto the normalized loading vectors (5.56). MCA analysis is a little different than EOF analysis, since to get the structure of the left field, you can project the left field data onto the expansion coefficient of the right singular vector, and vice versa. These are heterogeneous regression maps.

$$\mathbf{D}_{XY} = \mathbf{X}\widetilde{\mathbf{Y}}^{*T}/N/quad \quad \mathbf{D}_{YX} = \mathbf{Y}\widetilde{\mathbf{X}}^{*T}/N \tag{5.60}$$

Where here the tilde indicates that the time series of the loading vectors has been normalized to unit variance. One can also construct homogeneous regression maps by regressing the left variable onto the normalized left loading vectors.

$$\mathbf{D}_{XX} = \mathbf{X}\widetilde{\mathbf{X}}^{*T}/N \qquad \mathbf{D}_{YY} = \mathbf{Y}\widetilde{\mathbf{Y}}^{*T}/N \tag{5.61}$$

Heteorgeneous regressions show the amplitude and structure of the patterns that best explain the covariance between two data sets. The homogenous regressions 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. If they are similar then the structures that explain the variance well also explain the covariance.

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.

5.9.4 Statistical significance of MCA

MCA 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 sensitivity of the results to the specfic sample chosen. 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.

Many caveats and criticisms have been offered for MCA analysis. Newman and Sardeshmukh(1995) showed that MCA would reveal a linear operator $\mathsf{y} = \mathsf{L}\mathsf{x}$ only under the very restrictive condition that the operator was orthogonal, $\mathsf{L}^{\mathsf{T}} = \mathsf{L}^{-1}$.

Cherry(1996) recommended extreme caution in applying MCA, since it tends 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. 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.

5.10 Canonical Correlation Analysis

Principal component analysis and MCA analysis can be performed in sequence, and we can call the result Canonical Correlation Analysis (CCA) (e.g. Barnett and Preisendorfer; 1987), First performing EOF analysis, and then truncating the EOF expansion will reduce noise and make the search for correlation less subject to noise. The raw state variables can be first subjected to EOF/PC analysis, and new state variables formed from the subset the EOF/PCs that explains 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. One then calculates the SVD from the correlation 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.

The first step in CCA is to perform EOF analysis on the original data for both the left and right fields and construct the time series of the PCs, which are the amplitudes of the EOFs at each sampling time for each data set. Of course this step presumes that 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 data are next truncated by retaining only those PCs that explain a lot of variance, thus reducing 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 trade-off 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. 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.

The retained PC time series are next normalized to 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 truncated and normalized PC 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.

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.

**********end Dennis **********

5.11 Cluster Analysis

Cluster analysis is a popular technique in data mining and its goal is to group a set of observations into a number of distinct clusters. The end result is that each observation belongs to only one cluster, and this clustering is determined based on a particular cost function. Determining the optimal set of clusters is often an iterative process that begins with an initial guess.

Recall that the principal components in EOF analysis tell you how much a given observation looks like a particular EOF pattern. While, an observation may look predominately like one particular EOF (thus, the principal component for this EOF is large), it likely still has non-zero principal component values for the other EOFs. In cluster analysis, this is not the case, as each observation is tied to only one cluster.

5.11.1 k-*means Clustering*

k-means clustering is one of the more popular clustering techniques in Earth science due to its relative simplicity. The aim is to group N observations in k clusters in which each observation belongs to the cluster with the nearest center. This results in the data being partitioned into Voronoi cells. The optimal distribution of the cluster centers is the distribution that minimizes the within cluster sum of squares (i.e. the sum of the distance functions of each point to the cluster center). Mathematically, this is written as:

$$\underset{\mathbf{S}}{\operatorname{argmin}} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{S}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{5.62}$$

where $\mathbf{S} = S_1, S_2, ..., S_k$ and denotes the set of k clusters, $\mathbf{x}$ denotes the set of observations, $\boldsymbol{\mu}_i$ denotes the center of cluster S_i defined as the mean of all points in S_i, and argmin specifies that we are looking for the minimum over varying options of $\mathbf{S}$.

While the idea behind k-means is straightforward, actually computing the most optimal solution is very difficult and time intensive, especially for large data sets. Thus, most applications of k-means clustering employ heuristic algorithms that converge quickly to a local optimal solution that may or may not be the global optimal solution. Thus, in practice, one typically runs the k-means clustering algorithm many times (with each instance starting with a unique initial guess) and takes the iteration that produced the smallest minimum within cluster sum of squares (i.e. **(5.62)**).

The standard algorithm for k-means clustering is Lloyd's algorithm, which begins with an initial guess of the cluster centers $\boldsymbol{\mu}$ and then iteratively refines this guess until the it converges. **Fig. 5.8** shows an example of this iterative process.

To apply k-means clustering, two decisions that must be made:

1. **How many clusters do you want?** This is a decision that must be made by the user, and is not determined by the clustering algorithm. One will obtain different results based on the choice of k as seen in **Fig. 5.9**. There are methods such as the gap statistic that can be used to help determine which k to use (Tibshirani et al. 2001).
2. **How should you seed your initial guess?** There are many options available for how to best make your initial guess at the cluster centers. In addition, running k-means clustering many times, and choosing the optimal solution out of these is a way to avoid your final solution being too dependent on the initial guess.

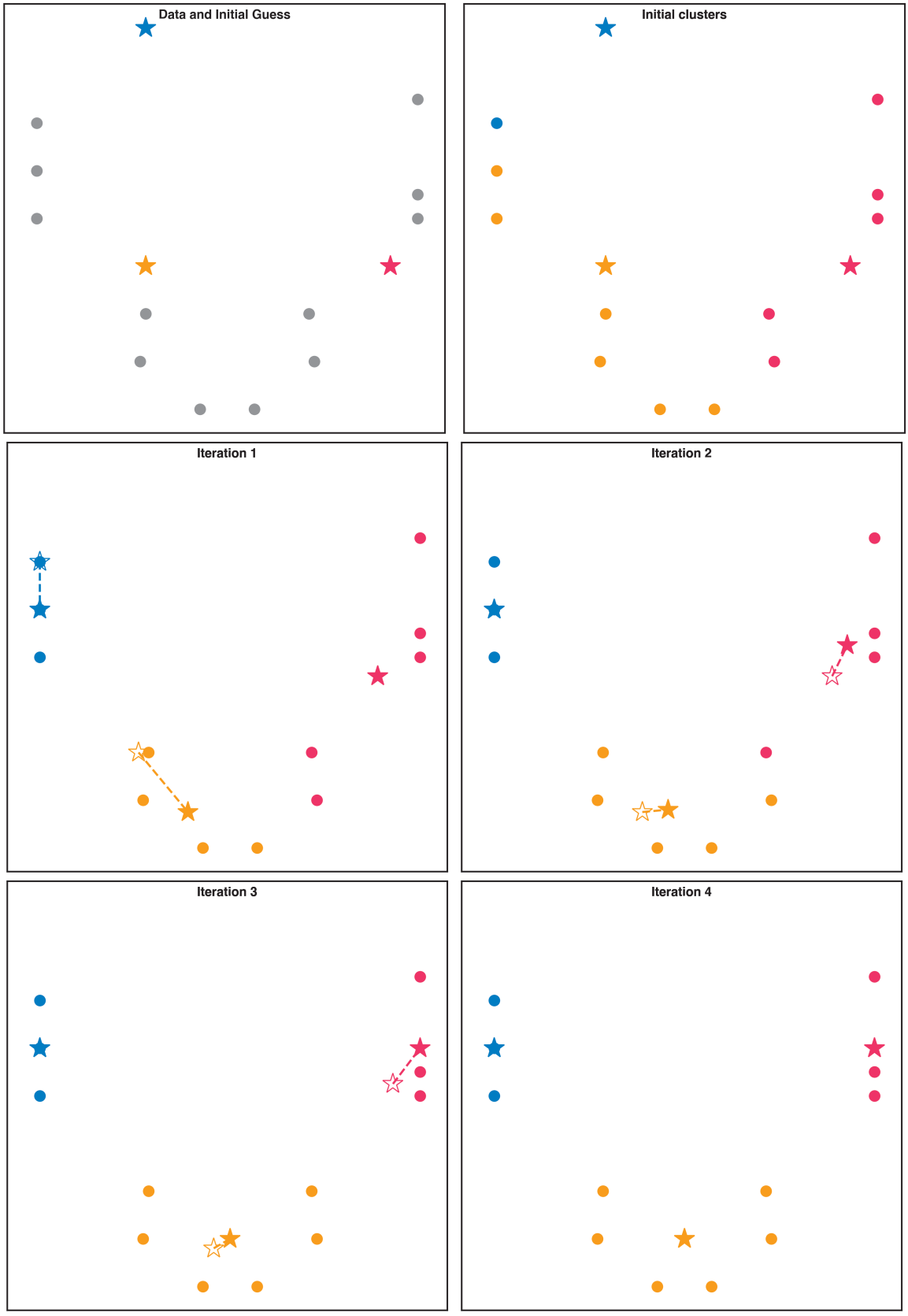

Figure 5.8 An example of the iterative refinement of k-means clustering using Lloyd's algorithm. Cluster centers are shows as stars, and empty stars denote the previous cluster center. The algorithm has converged after the third iteration.

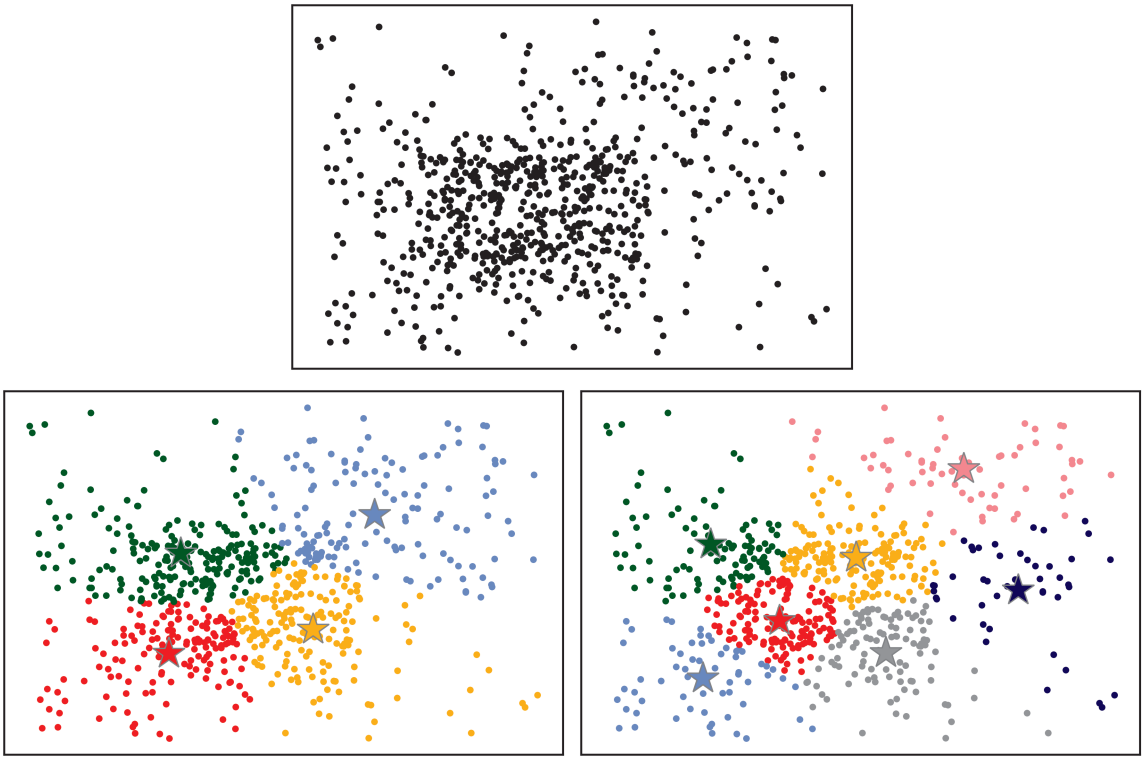

Figure 5.9 Using k-means clustering to identify 4 (bottom left) and 7 (bottom right) clusters in the data. The top panel shows the raw data. The stars denote the centroids of each of the seven clusters. The k-means algorithm was run 20 different times and the iteration with the optimal cost function was retained.

5.11.2 Self-Organizing Maps (SOMs)

A self-organizing map, also known as a Kohonen map, is yet another form of cluster analysis that uses unsupervised machine learning to train an artificial neural network. Under certain conditions it is identical to the k-means clustering method, however, the distinguishing feature of SOMs is that when one cluster center is updated, the neighboring cluster centers can also be updated in the a similar way. This results in a set of cluster centers (or *nodes*) that are organized such that nodes that are most similar are located near one another and nodes that are least similar are further apart when the are organized into an $m \times p$ grid.

Like most artificial neural networks, the SOMs algorithm can be summarized as a two step process: training and mapping.

1. **Training:** The training phase is when the SOM algorithm does the heavy thinking. During training, the SOM nodes are updated by comparing them to input examples (i.e. observations) that are ingested one at a time. Often, the same training data is iterated through multiple times. The result is a set of trained SOM nodes.
2. **Mapping:** After training is complete, the mapping phase involves automatically classifying a new observation as belonging to one of the nodes. In some cases, the training data is distinct from the data you are interested in mapping, while in other cases they may be one and the same.

As with k-means clustering, the user must first determine how many SOM nodes they wish to create to represent their data. Unlike k-means, however, this information is provided as an $m \times p$ grid, and thus, the user must determine both m and p. In the example below we will choose a SOM size of 20×20 giving a total of 400 nodes, however, it is important to note that the resuting SOM nodes are highly dependent on the dimensions chosen.

SOM training begins by first initializing the $m \cdot p$ nodes. This can be done either with random data, or using the principal components of the observations. Then, the algorithm proceeds by ingesting the training data either one observation at a time (*online training*) or as a single group (*batch training*). Then the observations are compared to each of the $m \cdot p$ nodes and the *best match unit* (BMU) is identified as the node with the smallest Euclidean distance to the observation in question. We will denote this BMU as n_i to signify the i^{th} node, where i could take a value from 1 to $m \cdot p$. Finally, the BMU and neighboring nodes are updated in the following manner:

$$n_j(t+1) = n_j(t+1) + \alpha(t) h_{ci}(t)[x(t) - n_i(t)] \tag{5.63}$$

where $1 \leqslant j \leqslant m \cdot p$, t is the training time, α is the learning rate parameter and h_{ci} is the neighborhood function. Specific descriptions of these inputs are given below.

- **training time (t):** how many iterations have been performed
- **learning rate parameter (α):** how strongly to update the nodes given the observation; typically decreases with training time
- **neighborhood function (h):** a function describing how many surrounding nodes to update besides the BMU; typically decreases with training time

The learning rate parameter and the neighborhood function must all be chosen by the user. Often, the learning rate parameter starts large, and then decreases linearly with training time. There are many neighborhood functions to choose from, but some common ones are a 2D Guassian or the Epanechikov function. It is typically considered good practice to initially set your neighborhood function parameters to include the entire SOM-space to ensure that all nodes are updated. This can be modified at later training times.

Typically, the SOM algorithm is implemented in multiple stages, with varying combinations of the above parameters.

Since SOM analysis has a large number of free parameters chosen by the user, there are many possible SOMs that can be computed from a single set of observations. One question is, how do I choose which one to use? Two key metrics should be considered:

- **quantization error:** the average distance between each observation and its BMU
- **topographic error:** the fraction of all observations who's first and second BMUs are *not* adjacent units

One wants to minimize both the quantization error and the topographic error, and typically, the topographic error is kept near zero while the quantization error is minimized. Thus, it is good practice to perform SOM analysis multiple times on the same data set, using different paramater values, and then choose the best based on the quantization and topographic errors.

To demonstrate SOM analysis and how the results can be visualized, we use hourly observations from Christman Field in Fort Collins, Colorado from January 1 - December 31, 2016. The data input into the algorithm is a 2D matrix X with dimensions (8784, 6). That is, 8784 hourly observation periods and 6 variables: temperature (degrees F), relative humidity (%), wind speed (mph), pressure (mb), solar radiation (W/m^2) and precipitation (mm). SOM training was performed in two phases: rough and finetune, each of which had different training lengths and initial and final radii of influence (for use in the neighborhood function). The SOM nodes were initialized with random data and a 20× 20 grid was chosen. The resulting SOM nodes are plotted in **Fig. 5.10**, where each panel denotes a different variable and each grid denotes a different SOM. For example, we can see that times of very strong wind are often associated with low relative humidity and precipitation. **Fig. 5.11** displays the frequency of occurrence of each of the SOM nodes (i.e. how many observations had that particular node as their BMU). We can see that the nodes on the left hand side of our grid tend to occur most frequently, and from **Fig. 5.10**, these periods are defined by low wind, high relative humidity, generally lo temperatures on average. [1]

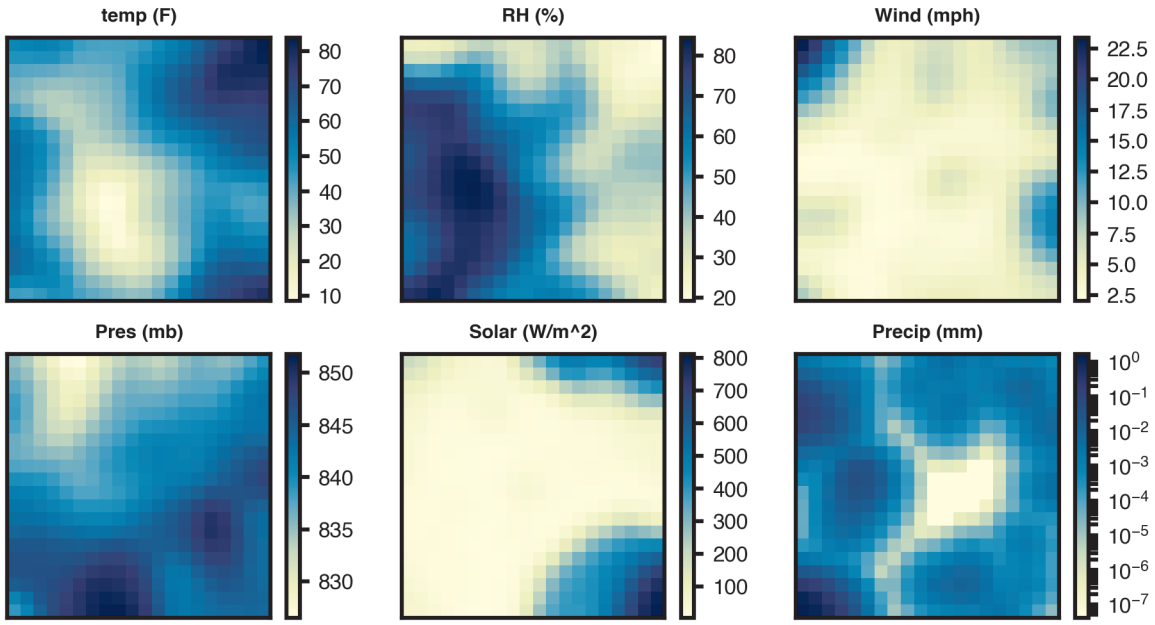

Figure 5.10 SOM weights for each variable displayed in a 20 × 20 grid.

[1] This example as created using the SOMPY code available here: https://github.com/sevamoo/SOMPY/tree/master/sompy

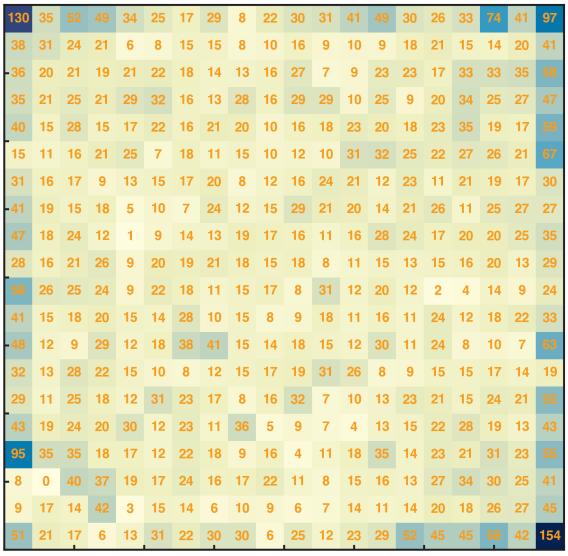

Figure 5.11 The frequency of occurrence of each of the 400 SOM patterns. Darker colors denote more frequent, and the number of occurrences out of the 8784 observations is written in orange.

Chapter 6
Mapping Data to a Grid and Data Assimilation

6.1 Placing data on a regular grid

In dynamical meteorology, oceanography, and numerical prediction one is often presented with the following problem. Data are available at a number of observation points (usually located near cities or at field stations, along ship cruise tracks, at moorings, or perhaps located by the observation points of an orbiting satellite) that are unevenly distributed over the domain of interest (the globe, for example). In order to compute derivatives of the field variables, as would be required in diagnostic studies or in the initialization of a numerical model, or simply to perform a sensible averaging process, one often requires values of the variables at points on a regular grid. Assigning the best values at the grid points, given data at arbitrarily located stations and perhaps a first guess at regular grid points, is what has traditionally been called objective analysis when done on a computer rather than graphically by hand.

We will use the example of making weather maps from rawinsonde data as the particular example of the mapping problem here. In fact the methods described are applicable to any problem where the data you are given do not fill the domain of interest fully, and/or where the data must be interpolated to a regular grid. The regridding can be in space, in time, or both. You may also find yourself in the position of wanting to plot a continuous function of an observation in two parameter dimensions, and have samples at only a few points. We will proceed through some of the methods in the order that they arose in the history of numerical weather forecasting. In this way we show the weaknesses of some of the most obvious methods such as function fitting, to the correction method, and ultimately to statistically optimized correction methods such as optimum interpolation. Current assimilation schemes in numerical forecast models us a combination of optimum interpolation and use of the governing equations of the model, which we can call *Kalman filtering*, which is discussed in elementary terms in Chapter **??**.

6.1.1 Interpolation with polynomial fits

Let's say we want to estimate the temperature at a point. However, we don't have any observations at that exact location. How might we use our observations to still get an estimate of the temperature at our point? The answer could be to perform some sort of interpolation. Probably one of the most intuitive methods for interpolating is to fit some polynomial to all of our station values, and then use that curve to get the temperature at a location between the observations. For example,

$$\Phi(x, y) = a_0 + a_1 x + a_2 x^2 + b_2 y^2 + 2c_2 xy + \dots \tag{6.1}$$

It turns out this isn't a very good method when you have sparse data due to the unstable nature of the polynomial fit. Removing just one point can wildly change the curve/interpolation in the vicinity of this point and will impact the values at many other points too. The problem gets worse as the order of the polynomial is increased. An example of this is depicted in **Fig. 6.1**. Note how wildly the two curves depart from each other in the vicinity of the missing point. Such problems can be avoided by stepping away from

polynomial fits and rather, utilizing a reasonable "first guess", and then only modifying it when and where data are available. Also, if the new data departs too wildly from the first guess, one suspects that the data are faulty.

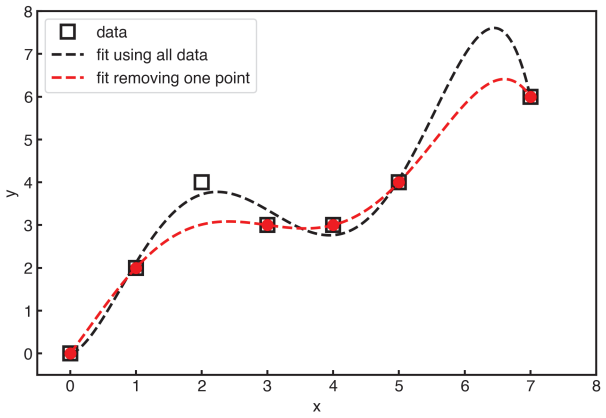

Figure 6.1 Illustration of the unstable nature of polynomial fits when one data point is removed using a 5th order polynomial.

A polynomial fit that actually got adopted by the US National Meteorological Center for its routine operational products was proposed by Flattery (1971). In this scheme, *Hough functions* were used as the interpolating polynomials. These functions are an orthogonal set that are the solutions of the linearized equations for a resting atmosphere (the tidal equations). The idea was that if you expressed the data in terms of actual solutions of the dynamical equations, then your fit between the data points would have some dynamical consistency. The Hough functions are global functions and so all of the observations were used simultaneously to define the global Hough function coefficients and produce a global map. Only the Hough functions describing slowly varying rotational modes were used. The gravity wave modes were zeroed out to produce a well-initialized field. This method replaced Cressman's correction method (Cressman, 1959) for global analyses in about 1972 and was replaced by Optimum Interpolation (see Chapter 6.1.2) in 1978.

This method has some dynamical and mathematical appeal, but is in truth just a glorified polynomial fit and has all of the problems of polynomial fits. First of all, the atmosphere is highly nonlinear and strongly forced by heating, especially in the tropics. The Hough modes chosen were primarily the free, non-divergent Rossby modes, which constitute a large, but not dominant, fraction of the variance. Therefore this aspect of the Flattery method did not buy much. In the tropics, where highly divergent motions forced by heating are important, the analyses constructed with the Flattery method are very much in error, especially in their estimates of divergence, which they set to essentially zero. In addition the Hough function fits are wildly unstable in regions of sparse data, like any polynomial fit. The NMC tropical analyses produced before 1978 are almost totally useless because they were made with the Flattery analysis system. Normal mode fits are still used in numerical initialization schemes to remove fast gravity waves, but this does not really affect the slowly changing meteorological flow. Modern reanalysis data products are based on data assimilation methods that take into account both the data and the model forecast and the uncertainty in both.

6.1.2 Optimum Interpolation

> *"The interpolation which is linear relative to the initial data and whose root-mean- square error is minimum is called the optimum interpolation." - Wiener, 1949*

The difference between optimum interpolation and linear regression is that the coefficients are not determined anew each time. Suppose we consider deviations from some "normal" state. This could be climatology or a

first guess, depending upon the application.

$$\phi^{'} = \phi - \phi_{norm} \quad \phi_{norm} = \overline{\phi} \text{ or a first guess} \tag{6.2}$$

Then we try to approximate the value of ϕ at a grid point, ϕ_g, in terms of a linear combination of the values of ϕ at neighboring station points, ϕ_s.

$$\phi_g^{'} = \sum_{i=1}^{N} p_i \phi_i^{'} \tag{6.3}$$

The coefficients p_i are to be determined by minimizing the mean squared error

$$E = \overline{\left(\phi_g^{'} - \sum_{i=1}^{N} p_i \phi_i^{'}\right)^2} \tag{6.4}$$

We can write the normalized error as

$$\epsilon \equiv \frac{E}{\overline{\phi_g^{'2}}} = 1 - 2\sum_{i=1}^{N} p_i r_{gi} + \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} \tag{6.5}$$

$$\text{where } r_{gi} = \frac{\overline{\phi_g^{'}\phi_i^{'}}}{\overline{\phi_g^{'2}}} \quad r_{ij} = \frac{\overline{\phi_i^{'}\phi_j^{'}}}{\overline{\phi_g^{'2}}} \tag{6.6}$$

Differentation with respect to the coefficients leads to the condition of minimization used to determine them.

$$\frac{\partial \epsilon}{\partial p_i} = -2r_{gi} + 2\sum_{j=1}^{N} p_j r_{ij} = 0 \quad i = 1, 2, \ldots, N \tag{6.7}$$

(6.7) constitutes a system of N linear equations for the N p's. By substituting the conditions **(6.7)** into the expression for the error **(6.5)**, it can be shown that the error obtained after fitting the coefficients is

$$\epsilon = 1 - \sum_{i=1}^{2} r_{gi} p_i \tag{6.8}$$

Note that in this simple example, if one of the observation points, k, coincides with a grid point, then $r_{gk} = 1$, and we expect the regression procedure to return $p_k = 1$ and all the other weights zero. In this case the error is zero, $\epsilon = 0$, since we have assumed the data are perfect. If the station points are uncorrelated with the grid point in question, then $p_i = 0$ and $\epsilon = 1$, the climatic norm. That is, the error will equal the standard deviation, but no worse.

6.1.2.1 Adding measurement error

In what we have done so far the observations have been assumed to be perfect. Let us now consider what happens if we explicitly take account of the fact that our observations will always contain some error, δ_i.

$$\phi_i^{'} = \phi_{ia}^{'} + \delta_i \tag{6.9}$$

Let's assume, as is usually reasonable, that the error is unbiased (zero mean) and ucorrelated with the true value, that is,

$$\overline{\phi_{ia}^{'}\delta_i} = 0 \tag{6.10}$$

and that the errors at the various stations where we have data are also uncorrelated

$$\overline{\delta_i \delta_j} = \begin{cases} \overline{\delta^2} \text{ for } i = j \\ 0 \text{ for } i \neq j \end{cases} \tag{6.11}$$

In this case, rather than **(6.5)**, we obtain

$$\epsilon = 1 - 2\sum_{i=1}^{N} p_i r_{gi} + \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} + \eta \sum_{i=1}^{N} p_i^2 \tag{6.12}$$

where r_{ij} is the correlation between the two points and where η is the ratio of the error variance to the measurement variance - in other words, the *signal-to-noise ratio.*

$$\eta = \frac{\overline{\delta^2}}{\overline{\phi_g'^2}} \tag{6.13}$$

Minimization of the error leads to the condition

$$\sum_{j=1}^{N} r_{ij} p_j + \eta p_i = r_{gi} \quad \text{for } i = 1, 2, 3, \ldots, N \tag{6.14}$$

In this case the normalized error is

$$\epsilon = 1 \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} + \eta \sum_{i=1}^{N} p_i^2 \tag{6.15}$$

6.1.2.2 What is the effect of including noise in the measurements?

In order to see how optimum interpolation treats the *a priori* information that the measurements include some error, it is instructive to compare the results **(6.14)** and **(6.15)** with the results **(6.5)** and **(6.7)** obtained assuming perfect data. In the case of perfect data, **(6.7)** gives

$$r_{ij} p_j = r_{gi} \quad \text{or } p_j = r_{ij}^{-1} r_{gi} \tag{6.16}$$

When noise is included we get, rather, the result **(6.14)**, which can be written

$$\{r_i j + \eta \mathbf{I}_{ij}\} p_j = r_{gi} \quad \text{or } p_j = \{r_i j + \eta \mathbf{I}_{ij}\}^{-1} r_{gi} \tag{6.17}$$

where $\mathbf{I}_{ij}$ is the unit matrix. Looking at the right-hand member of the pair of equations in **(6.17)**, it is easy to see that the coefficients p_j will be smaller when the error is large. This is most obvious if we assume that $r_i j$ is diagonal. Thus we see that the inclusion of error makes the coefficients in **(6.3)** smaller and that therefore, by **(6.2)**, the estimate we make will be closer to climatology. If we include error, then Optimum Interpolation will draw more closely to climatology or the first guess and tend to weight new observations less heavily. This is desirable. By putting different values of η_j along the diagonal, one can put information on the confidence one has in individual stations into the analysis scheme and weight more heavily those stations in which one has more confidence.

6.1.2.3 What do we need to make Optimum Interpolation work?

In order to make the above schemes work, we need the correlation matrices r_{ij} and r_{gi}. The first of these is easily calculable from observations, but the second is not since it involves correlations between the station points and the grid points. We do not have data at the grid points, or we would not need an analysis scheme. In practice, not even the r_{ij} are calculated in full generality. It is possible to assume that correlations between

points depend only on the distance between them and not on location or direction (although it would be possible to include directionally dependent (anisotropic) correlations). In this case the single isotropic correlation function can be estimated from station data. This is a crude approximation since correlations between stations depend on the location of the stations and whether longitude or latitude separates them. An example illustrating the anisotropy of correlation functions in 500 hPa geopotential heights is shown in **Fig. 6.2**.

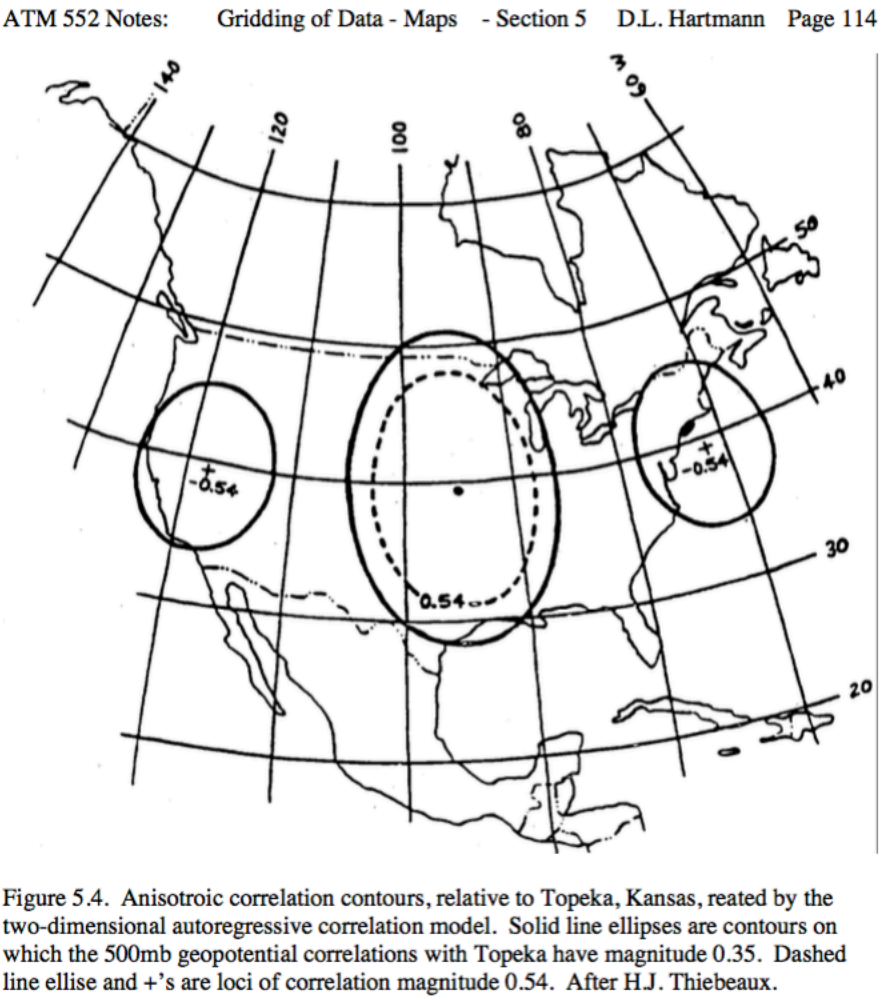

Figure 5.4. Anisotroic correlation contours, relative to Topeka, Kansas, reated by the two-dimensional autoregressive correlation model. Solid line ellipses are contours on which the 500mb geopotential correlations with Topeka have magnitude 0.35. Dashed line ellise and +'s are loci of correlation magnitude 0.54. After HJ. Thiebeaux.

Figure 6.2 Anisotropic correlation contours, relative to Topeka, Kansas, created by a two-dimensional autoregressive correlation model. Solid line ellipses are contours on which the 500 hPa geopotential correlations with Topeka have magnitude 0.35. Dashed lines denote the correlation of 0.54 and +'s are loci of correlation magnitude 0.54. After H.J. Thiebeaux.

Libby: *stopped at Dennis' Chapter 5.4*