CHAPTER 2

Time Domain Methods

Time domain methods do not employ any form of transform space to describe a time series (although it is commonly the case that one can best understand their structures by analyzing them in the frequency domain). The names most associated with these techniques are Wiener, and Box and Jenkins. As with the frequency domain methods, one can begin the discussion in continuous time and it was one of Wiener's great contributions to show how to deal with that case. But continuous time representations raise all sorts of complex mathematical issues that disappear when a time series is made discrete, and so for present purposes, we will begin with the discrete case of a uniformly sampled time series x_t.

1. Representations-1

As with Fourier methods, much of the purpose of these methods is to find efficient representations of stochastic processes whose interpretation can lead to physical insights. For notational simplicity, we will assume that $\Delta t = 1$. Consider the simple rule (actually a difference equation of similar form to (7.1) above),

$$x_{m+1} = ax_m + \theta_m \tag{1.1}$$

where a is a constant and θ_m is a zero-mean white noise process of variance σ_θ^2. Starting with $x_0 = 0$, (1.1) permits simple generation of realizations of x_m depending upon the particular run of random numbers θ_m (Fig. 28). We can compute the autocovariance of x_m :

$$R(0) =< x_m^2 >=< (ax_{m-1} + \theta_{m-1})^2 >= a^2 R(0) + \sigma_\theta^2 \tag{1.2}$$

where we used $< x_{m-1}\theta_{m-1} >= 0$, and the assumption that the time-series was wide-sense stationary ($< x_{m-1}^2 >=< x_m^2 >= R(0)$). So,

$$R(0) = \frac{\sigma_\theta^2}{(1-a^2)}. \tag{1.3}$$

Evidently, there would be a problem if $a = 1$, and in fact, $|a| < 1$ proves to be necessary for the time-series to be stationary. Similarly,

$$R(1) =< x_{m+1}x_m >=< (ax_m + \theta_{m+1})\, x_m >= aR(0). \tag{1.4}$$

Exercise. Find $R(2), ..., R(m)$ for x_t in (1.1).

If one knew $R(0)$ and $R(1)$, Eqs. (1.3, 1.4) would fully determine a, σ_θ^2 and they in turn fully determine everything there is to know about it. Before asking how one might determine $R(0), R(1)$, let us ask where an equation such as (1.1) might arise?

Consider a simple differential system

$$\frac{dx(t)}{dt} = Ax(t) + g(t) \tag{1.5}$$

where A is constant and θ is any externally imposed forcing. Equations like this are used to describe, e.g., a local change in a heat content anomaly, $x(t)$, as the result of conduction from a reservoir, heat loss by radiation, and external sources g. Forming simple one-sided time differences, (1.5) becomes

$$x(m\Delta t + \Delta t) = \Delta t(A+1)x(m\Delta t) + \Delta t g(m\Delta t) \tag{1.6}$$

or,

$$x_{m+1} = \Delta t(A+1)x_m + \Delta t g_m \tag{1.7}$$

which is of the form (1.1) with $a = \Delta t(A+1)$. Two types of problem exist. In one, g_m is known, and one seeks a; in the other type, $g_m = \theta_m$ is unknown and believed to be a white noise process..

In the second type of problem one has observations of x_t and the question is what the best estimates of a, σ_θ^2 are. Let us try least-squares by minimizing,

$$J = \sum_{m=0}^{N-1} (x_{m+1} - ax_m)^2. \tag{1.8}$$

The argument here would be that (1.1) can be regarded as an equation which forecasts x_{m+1} from x_m, and minimizing the unpredictable part, θ_m, would give the best possible forecast system. The normal equations for (1.8) are just one equation in one unknown,

$$a\sum_{m=0}^{N-1} x_m^2 = \sum_{m=0}^{N-2} x_{m+1}x_m. \tag{1.9}$$

Divide both sides of this equation by N,and we see that it can be written as

$$a\tilde{R}(0) = \tilde{R}(1), \tag{1.10}$$

where we recognize

$$\frac{1}{N}\sum_{m=0}^{N-1} x_m^2, \tag{1.11}$$

as an *estimate* of the true autocovariance $R(0)$, and similarly for $R(1)$. Given the resulting estimate of a, call it $\tilde{a}$, one can substitute into (1.8) and compute the estimate $\tilde{\sigma}_\theta^2$.

A more general form of the representation of a time-series is,

$$x_{m+1} = a_1 x_m + a_2 x_{m-1} + ... + a_M x_{m-M+1} + \theta_{m+1}, \tag{1.12}$$

which is called an "autoregressive process of order M" or AR(M), so that (1.1) is an AR(1) process. To determine the coefficients a_i we can proceed again by least-squares, to find the minimum of

$$J = \sum_{m=0}^{N-1} \left(x_{m+1} - a_1 x_m - a_2 x_{m-1} - ... - a_M x_{m-M+1} \right)^2 \tag{1.13}$$

and forming the normal equations,

$$\begin{aligned} a_1 \tilde{R}(0) + a_2 \tilde{R}(1) + a_3 \tilde{R}(2) + .. + a_M \tilde{R}(M-1) &= \tilde{R}(1) \\ a_1 \tilde{R}(1) + a_2 \tilde{R}(0) + a_3 \tilde{R}(1) + .. + a_M \tilde{R}(M-2) &= \tilde{R}(2) \\ &... \\ a_1 \tilde{R}(M-1) + a_1 \tilde{R}(M-2) + a_3 \tilde{R}(M-3) . + . + a_M \tilde{R}(0) &= \tilde{R}(M) \end{aligned} \tag{1.14}$$

where we used $\tilde{R}(-k) = \tilde{R}(k)$. Equations (1.14) are usually known as the Yule-Walker equations. Solving them produces an estimate of the vector of unknowns $\mathbf{a} = [a_1, ...a_M]^T$ and the value of J is the estimate of σ_θ^2. If (1.14) is written in matrix form

$$\tilde{\mathbf{R}}\mathbf{a} = \mathbf{b} \tag{1.15}$$

one sees that $\tilde{\mathbf{R}}$ is a covariance matrix having the special property that all diagonals have the same values:

$$\tilde{\mathbf{R}} = \begin{Bmatrix} \tilde{R}(0) & \tilde{R}(1) & \tilde{R}(2) & . & \tilde{R}(M-1) \\ \tilde{R}(1) & \tilde{R}(0) & \tilde{R}(1) & . & \tilde{R}(M-2) \\ \tilde{R}(2) & \tilde{R}(1) & \tilde{R}(0) & . & \tilde{R}(M-3) \\ . & . & . & . & . \\ \tilde{R}(m-1) & \tilde{R}(m-2) & \tilde{R}(m-3) & . & \tilde{R}(0) \end{Bmatrix} \tag{1.16}$$

A matrix with constant diagonals is called "Toeplitz", and the special form of (1.15) permits the system of equations to be solved without a matrix inversion, using an extremely fast recursive algorithm called the Levinson (or sometimes, Levinson-Derber) algorithm. This possibility is less important today than it was in the days before fast computers, but if M is extremely large, or very large numbers of systems have to be solved, the possibility can remain important.

If g_m is a known time-series, one can proceed analogously by minimizing via least-squares, the objective function

$$J = \sum_{m=0}^{N-1} \left(x_{m+1} - a x_m - g_m \right)^2 \tag{1.17}$$

with respect to a. Higher order generalizations are obvious, and details are left to the reader.

2. Geometric Interpretation

There is a geometric interpretation of the normal equations. Let us define vectors (of nominally infinite length) as

$$\mathbf{x}_r = \left[...x_{r-1}, x_{r-1}, \underset{\uparrow}{x_r}, x_{r+1}, x_{r+2,...} \right]^T \tag{2.1}$$

$$\boldsymbol{\theta}_r = \left[...\theta_{r-1}, \theta_{r-1}, \underset{\uparrow}{\theta_r}, \theta_{r+1}, \theta_{r+2,...} \right]^T \tag{2.2}$$

where the arrow denotes the time origin (these vectors are made up of the elements of the time series, "slid over" so that element r lies at the time origin). Define the inner (dot) products of these vectors in the usual way, ignoring any worries about convergence of infinite sums. Let us attempt to expand vector $\mathbf{x}_r$ in terms of $M-$ *past* vectors:

$$\mathbf{x}_r = a_1 \mathbf{x}_{r-1} + a_2 \mathbf{x}_{r-2} + ... + a_M \ \mathbf{x}_{r-M} + \varepsilon_r \tag{2.3}$$

where ε_r is the residual of the fit. Best fits are found by making the residuals orthogonal to the expansion vectors:

$$\mathbf{x}_{r-i}^T \left(\mathbf{x}_r - a_1 \mathbf{x}_{r-1} - a_2 \mathbf{x}_{r-2} - ... - a_M \ \mathbf{x}_{r-M} \right) = 0, 1 \leq i \leq M \tag{2.4}$$

which produces, after dividing all equations by N, and taking the limit as $N \to \infty$,

$$a_1 R(0) + a_2 R(1) + ... + a_M R(M-1) = R(1) \tag{2.5}$$

$$a_1 R(1) + a_2 R(0) + ... + R(M-2) = R(2) \tag{2.6}$$

$$..., \tag{2.7}$$

that is precisely the Yule-Walker equations, but with the theoretical values of R replacing the estimated ones. One can build this view up into a complete vector space theory of time series analysis. By using the actual finite length vectors as approximations to the infinite length ones, one connects this theoretical construct to the one used in practice. Evidently this form of time series analysis is equivalent to the study of the expansion of a vector in a set of (generally non-orthogonal) vectors.

3. Representations-2

An alternative canonical representation of a time series is

$$x_m = \sum_{k=0}^{M} a_k \theta_{m-k}, \tag{3.1}$$

(M can be infinite), which called a moving average process of order M (an MA(M)). Here again θ_t is zero-mean white noise of variance σ_θ^2. Notice that only positive indices a_k exist, so that x_m involves only

the *past* values of θ_k. We can determine these a_k by the same least-squares approach of minimizing an error,

$$J = \sum_{m=0}^{N-1} \left(x_m - \sum_{k=0}^{M} a_k \theta_{m-k} \right)^2, \tag{3.2}$$

and leading to a set of normal equations, again producing a Toeplitz matrix. As with the AR problem, the real issue is deciding how large M should be.

Exercise. Derive the normal equations for (3.2) and for $J =< \left(x_m - \sum_{k=0}^{M} a_k \theta_{m-k} \right)^2 >$.

Various theorems exist to show that any stationary discrete time series can be represented with arbitrary accuracy as either an MA or AR form. Given one form, it is easy to generate the other. Consider for example (3.1). We recognize that x_m is being represented as the convolution of the finite sequence a_k with the sequence θ_m. Taking the Fourier (or z) transform of x_m produces,

$$\hat{x}(z) = \hat{a}(z)\,\hat{\theta}(z) \tag{3.3}$$

or,

$$\hat{\theta}(z) = \frac{\hat{x}(z)}{\hat{a}(z)}. \tag{3.4}$$

Assuming that $\hat{a}(z)$ has a stable, causal, convolution inverse, such that $\hat{b}(z) = 1/\hat{a}(z)$, we can write, by taking the inverse transform of (3.4)

$$\theta_m = \sum_{k=0}^{L} b_k x_{m-k} \tag{3.5}$$

Normalizing $b_0 = 1$, by dividing both sides of the last equation, we can recognize that (3.5) is in exactly the form of (1.12).

Exercise. Convert the moving average process $x_m = \theta_m - 1/3\theta_{m-1} + 1/4\theta_{m-2}$ into an AR process. What order is the AR process?

Because of the reciprocal nature of the vectors $\mathbf{a}, \mathbf{b}$ in the AR and MA forms, it is generally true that a finite length $\mathbf{a}$ generates a formally infinite length $\mathbf{b}$, and vice-versa (although in practice, one may well be able to truncate the formally infinite representation without significant loss of accuracy). The question arises as to whether a combined form, usually called an autoregressive-moving-average process (or ARMA), might produce the *most efficient* representation? That is, one might try to represent a given time series as

$$x_m - a_1 x_{m-1} - a_2 x_{m-2} - ... - a_N x_{m-N} = \theta_t + b_1 \theta_{m-1} + ... + b_M \theta_{m-M} \tag{3.6}$$

in such a way that the fewest possible coefficients are required. Taking the z transform of both sides of (3.6), we obtain

$$\hat{x}(z)\,\hat{a}(z) = \hat{\theta}(z)\,\hat{b}(z) \tag{3.7}$$

(defining $a_0 = b_0 = 1$). One can again use least-squares in either time or frequency domains. The major issues are once again the best choice of M, N and this problem is discussed at length in the various references.

Exercise. An ARMA is given as

$$x_m - \frac{1}{2}x_{m-1} = \theta_m - \frac{1}{8}\theta_{m-1} \tag{3.8}$$

Convert it to (a) an AR, and (b) a MA.

If one has the simplest AR,

$$x_m = ax_{m-1} + \theta_m \tag{3.9}$$

and takes its Fourier or z-transform, we have

$$\hat{x}(z)(1 - az) = \hat{\theta}(z) \tag{3.10}$$

and dividing

$$\hat{x}(z) = \frac{\hat{\theta}(z)}{1 - az} \tag{3.11}$$

The Taylor Series about the origin is

$$\hat{x}(z) = \hat{\theta}(x)\left(1 + az + az^2 + az^3 + ...\right) \tag{3.12}$$

which converges on $|z| = 1$ if and only if $|a| < 1$ and the corresponding MA form is

$$x_m = \sum_{k=0}^{\infty} a_k \theta_{m-k} \tag{3.13}$$

where the magnitude of the contribution from the remote past of θ_m diminishes to arbitrarily small values. If we nonetheless take the limit $a \to 1$, we have a process

$$x_m = \sum_{k=0}^{\infty} \theta_{m-k} \tag{3.14}$$

with an apparent infinite memory of past random forcing. Note that the AR equivalent of (3.14) is simply

$$x_m = x_{m-1} + \theta_m \tag{3.15}$$

—a more efficient representation.

This last process is not stationary and is an example of what is sometimes called an ARIMA (autoregressive integrated moving average).

Exercise: Show that the process (3.14 or 3.15) has a variance which grows with time.

Despite the non-stationarity, (3.15) is a very simple rule to implement. The resulting time series has a number of very interesting properties, some of which are described by Wunsch (1999) and Stephenson et al. (2000) including some discussion of their applicability as a descriptor of climate change.

Exercise. Using a pseudo-random number generator, form a 10,000 point realization of (3.15). Calculate the mean and variance as a function of sample length N. How do they behave? What is the true mean and variance? Find the power density spectrum of the realization and describe it. Compare the results to realizations from (3.9) with $a = 0.9999, 0.99, 0.9$ and describe the behaviors of the sample averages and variance with sample length and the changes in the spectrum.

Exercise. We can generalize the various representations to vector processes. Let

$$\mathbf{x}_m = [x_1(m), x_2(m), ...x_L(m)]^T$$

be a vector time series of dimension L. Then a vector MA form is

$$\mathbf{x}_m = \sum_{k=0}^{K} \mathbf{A}_k \, \boldsymbol{\theta}_{m-k}, \tag{3.16}$$

where the $\mathbf{A}_k$ are matrices, and $\boldsymbol{\theta}_m$ are vectors of white noise elements. Find the normal equations for determining $\mathbf{A}_k$ and discuss any novel problems in their solution. Discuss the question of whether the $\mathbf{A}_k$ should be square matrices or not. Define a vector AR form, and find the normal equations. What might the advantages be of this approach over treating each of the scalar elements $x_j(m)$ on its own?

4. Spectral Estimation from ARMA Forms

Suppose that one has determined the ARMA form (3.6). Then we have

$$\hat{x}(z) = \frac{\hat{\theta}(z)\,\hat{b}(z)}{\hat{a}(z)} \tag{4.1}$$

or setting $z = \exp(-2\pi i s)$,

$$< \hat{x}(\exp(-2\pi i s))\hat{x}(\exp(-2\pi i s))^* >= \Phi(s) = \frac{\left|\hat{b}(\exp(-2\pi i s))\right|^2}{\left|\hat{a}(\exp(-2\pi i s))\right|^2}\sigma_\theta^2. \tag{4.2}$$

If a, b are short sequences, then the calculation in (4.2) of the power density spectrum of the time series can be done essentially analytically. In particular, if $\hat{b} = 1$, so that one has a pure AR, the result is called the "all-pole" method, the power density spectrum being completely determined by the positions of the zeros of $\hat{a}(z)$ in the complex z plane. Under some circumstances, e.g., when the time series is made up of two pure frequencies differing in frequency by Δs in the presence of a white noise background, separation of the two lines can be achieved even if the record length is such that $\Delta s < 1/T$ that is, in violation of the Rayleigh criterion. This possibility and related considerations lead to what is commonly known as maximum entropy spectral estimation.

Exercise. Let $x_m = \sin(2\pi s_1 m) + \sin(2\pi s_2 m) + \theta_m, m = 0, 1, ...N$. Find an AR representation of x_m and use it to calculate the corresponding power density spectrum.

A considerable vogue developed at one time involving use of "exotic" methods of spectral representation, including, especially the maximum entropy method. Over time, the fashion has nearly disappeared because the more astute users recognized that maximum entropy etc. methods are dangerous: they can

give seemingly precise and powerful results apparently unavailable in the Fourier methods. But these results are powerful precisely because they rely upon the accuracy of the AR or ARMA etc. model. The sensitivity of e.g., (4.2) to the zero positions in $\hat{a}(z)$ means that if the pure pole representation is not the correct one, the appearance of spectral peaks may be spurious. The exotic methods began to fade with the realization that many apparent peaks in spectra were the result of an incorrect model. Tukey (1984) and others, have characterized ARMA-based methods as "covert", meaning that they hide a whole series of assumptions, and recommend reliance instead on the "overt" or non-parametric Fourier methods which are robust and hide nothing. This is good advice except for individuals who know exactly what they are doing. (Percival and Walden discuss these various methods at length.)

5. Karhunen-Loève Theorem and Singular Spectrum Analysis

The $N \times N$, $\mathbf{R}$ matrix in Eq. (1.16) is square and symmetric. It is an important result of linear algebra that such matrices have an orthogonal decomposition

$$\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{5.1}$$

where $\mathbf{\Lambda}$ is a diagonal matrix, $diag\{\lambda_i\}$ of the eigenvalues of $\mathbf{R}$ and $\mathbf{V}$ is the matrix of eigenvectors $\mathbf{V} = \{\mathbf{v}_i\}$, such that

$$\mathbf{R}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{5.2}$$

and they are orthonormal $\mathbf{v}_i^T \mathbf{v}_j = \delta_{ij}$. (It follows that $\mathbf{V}^{-1} = \mathbf{V}^T$.)

Let us write a time-series as an expansion in the $\mathbf{v}_q$ in the form

$$x_n = \text{element } n \text{ of } \left[\sum_{q=1}^{N} \alpha_q \sqrt{\lambda_q}\mathbf{v}_q\right] \tag{5.3}$$

or more succinctly, if we regard x_n as an $N-$element vector, $\mathbf{x}$,

$$\mathbf{x} = \left[\sum_{q=1}^{N} \alpha_q \sqrt{\lambda_q}\mathbf{v}_q\right]. \tag{5.4}$$

Here a_q are unit variance, uncorrelated random variates, e.g., $G(0,1)$. We assert that such a time-series has covariance matrix $\mathbf{R}$, and therefore must have the corresponding (by the Wiener-Khinchin Theorem) power density spectrum. Consider

$$\begin{aligned} R_{ij} &= < x_i x_j > = \sum_{q=1}^{N}\sum_{r=1}^{N} < \alpha_q \alpha_r > \sqrt{\lambda_q \lambda_r} v_{iq} v_{jr} \\ &= \sum_{q=1}^{N} \lambda_q v_{iq} v_{jq} \end{aligned} \tag{5.5}$$

by the covariance properties of α_q. But this last equation is just

$$R_{ij} = \left\{\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\right\}_{ij} \tag{5.6}$$

which is what is required.

Exercise, Confirm (5.6).

Thus (5.4) gives us another way of synthesizing a time series from its known covariance. That the decomposition (5.4) can always be constructed (in continuous time) for a stationary time series is called the Karhunen-Loève Theorem (see Davenport and Root, 1958). Because it is based upon a decomposition of the covariance matrix, it is evidently a form of empirical orthogonal function synthesis, or if one prefers, an expansion in principal components (see e. g., Jolliffe, 1986).

The relative importance of any orthogonal structure, $\mathbf{v}_i$, to the structure of x_n, is controlled by the magnitude of $\sqrt{\lambda_i}$. Suppose one has been successful in obtaining a physical interpretation of one or more of the important $\mathbf{v}_i$. Each of these vectors is itself a time series. One can evidently compute a power density spectrum for them, either by Fourier or more exotic methods. The idea is that the spectra of the dominant $\mathbf{v}_i$ are meant to be informative about the spectral properties of processes underlying the full time series. This hypothesis is a plausible one, but will only be as valid as the reality of the underlying physical structure attributed to $\mathbf{v}_i$. (At least two issues exist, determining when λ_i is significantly different from zero, and obtaining a physical interpretation of the components. Principal components are notoriously difficult to relate to normal modes and other physically generated structures.) This subject has come to be known as "singular spectrum analysis" (e.g. Vautard and Ghil, 1989) and its powers (and statistical information such as confidence limits) are still murky.

6. Wiener and Kalman Filters

6.1. The Wiener Filter. The theory of filtering of stationary time series for a variety of purposes was constructed by Norbert Wiener in the 1940s for continuous time processes in a notable feat of mathematics (Wiener, 1949). The work was done much earlier, but was classified until well after World War II). In an important paper however, Levinson (1947) showed that in discrete time, the entire theory could be reduced to least squares and so was mathematically very simple. This approach is the one used here. Note that the vector space method sketched above is fully equivalent too.

The theory of Wiener filters is directed at operators (filters) which are causal. That is, they operate only upon the past and present of the time series. This requirement is essential if one is interested in forecasting so that the future is unavailable. (When the future is available, one has a "smoothing" problem.) The immediate generator of Wiener's theory was the need during the Second World War for determining where to aim anti-aircraft guns at dodging airplanes. A simple (analogue) computer in the gunsight could track the moving airplane, thus generating a time history of its movements and some rough autocovariances. Where should the gun be aimed so that the shell arrives at the position of the airplane with smallest error? Clearly this is a forecasting problem. In the continuous time formulation, the requirement of causality leads to the need to solve a so-called Wiener-Hopf problem, and which can be mathematically tricky. No such issue arises in the discrete version, unless one seeks to solve the problem in the frequency domain where it reduces to the spectral factorization problem alluded to in Chapter 1.

The Wiener Predictor

Consider a time series x_m with autocovariance $R_{xx}(\tau)$ (either from theory or from prior observations). We assume that $x_m, x_{m-1}, ...,$ are available; that is, that enough of the past is available for practical purposes (formally the infinite past was assumed in the theory, but in practice, we do not, and cannot use, filters of infinite length). We wish to forecast one time step into the future, that is, seek a filter to construct

$$\tilde{x}_{m+1} = a_0 x_m + a_1 x_{m-1} + .. + a_M x_{m-M} = \sum_{k=0}^{M} a_k x_{m-k}, \tag{6.1}$$

such that the a_i are fixed. Notice that the causality of a_k permits it only to work on present (time m) and past values of x_p. We now minimize the "prediction error" for all times:

$$J = \sum_m \left(\tilde{x}_{m+1} - x_{m+1}\right)^2. \tag{6.2}$$

This is the same problem as the one leading to (1.14) with the same solution. The prediction error is just

$$\begin{aligned} P &= < \left(\tilde{x}_{m+1} - x_{m+1}\right)^2 > \\ &= R(0) - 2\sum_{k=0}^{M} a_k R(k+1) + \sum_{k=1}^{M}\sum_{l=1}^{M} a_k a_l R(k-l) \leq R(0). \end{aligned} \tag{6.3}$$

Notice that if x_m is a white noise process, $R(\tau) = \sigma_{xx}^2 \delta_{\tau 0}$, $a_i = 0$, and the prediction error is σ_{xx}^2. That is to say, the best prediction one can make is $\tilde{x}_{m+1} = 0$ and Eq. (6.3) reduces to $P = R(0)$, the full variance of the process and there is no prediction skill at all. These ideas were applied by Wunsch (1999) to the question of whether one could predict the NAO index with any skill through linear means. The estimated autocovariance of the NAO is shown in Fig. 1 (as is the corresponding spectral density). The conclusion was that the autocovariance is so close to that of white noise (the spectrum is nearly flat), that while there was a very slight degree of prediction skill possible, it was unlikely to be of any real interest. (The NAO is almost white noise.) Colored processes can be predicted with a skill depending directly upon how much structure their spectra have.

Serious attempts to forecast the weather by Wiener methods were made during the 1950s. They ultimately foundered with the recognition that the atmosphere has an almost white noise spectrum for periods exceeding a few days. The conclusion is not rigidly true, but is close enough that what linear predictive skill could be available is too small to be of practical use, and most sensible meteorologists abandoned these methods in favor of numerical weather prediction (which however, is still limited for related reasons, to skillful forecasts of no more than a few days). It is possible that spatial structures within the atmosphere, possibly obtainable by wavenumber filtering, would have a greater linear prediction possibility. This may well be true, but they would therefore contain only a fraction of the weather variance, and one again confronts the issue of significance. To the extent that the system is highly non-linear (non-Gaussian), it is possible that a non-linear filter could do better than a Wiener one. It is possible to show however, that for a Gaussian process, no non-linear filter can do any better than the Wiener one.

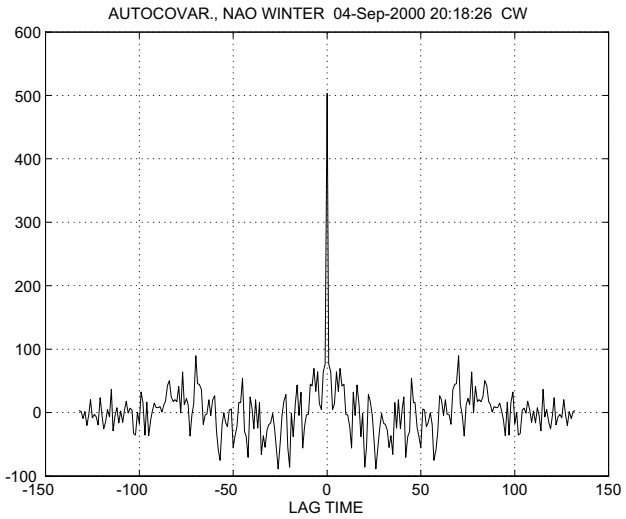

FIGURE 1. The estimated autocovariance of the North Atlantic Oscillation Index (NAO). Visually, and quantitatively, the autocovariance is dominated by the spike at the origin, and differs little from the autocovariance of a white noise process. By the Wiener filter theory, it is nearly unpredictable. See Wunsch (1999).

Exercise. Find the prediction error for a forecast at $k-$time steps into the future.

Exercise. Consider a vector time series, $\mathbf{x}_m = [h_m, g_m]^T$, where $< h_m g_m > \neq 0$. Generalize the Wiener prediction filter to this case. Can you find a predictive decomposition?

A slightly more general version of the Wiener filter (there are a number of possibilities) is directed at the extraction of a signal from noise. Let there be a time series

$$x_m = S_m + n_m \tag{6.4}$$

where S_m is the signal which is desired, and n_m is a noise field. We suppose that $< S_m n_m >= 0$ and that the respective covariances $R_{SS}(\tau) =< S_m S_{m+\tau} >$, $R_{mn}(\tau) =< n_m n_{m+\tau} >$ are known, at least approximately. We seek a filter, a_m, acting on x_m so that

$$\sum_m a_m x_{k-m} \approx S_k \tag{6.5}$$

as best possible. (The range of summation has been left indefinite, as one might not always demand causality.) More formally, minimize

$$J = \sum_{k=0}^{N-1} \left(S_k - \sum_m a_m x_{k-m} \right)^2 . \tag{6.6}$$

Exercise. Find the normal equations resulting from (6.6). If one takes $S_m = x_{m+1}$, is the result the same as for the prediction filter? Suppose a_m is symmetric (that is acausal), take the Fourier transform of

the normal equations, and using the Wiener-Khinchin theorem, describe how the signal extraction filter works in the frequency domain. What can you say if a_k is forced to be causal?

6.2. The Kalman Filter. R. Kalman (1960) in another famous paper, set out to extend Wiener filters to non-stationary processes. Here again, the immediate need was of a military nature, to forecast the trajectories of ballistic missiles, which in their launch and re-entry phases would have a very different character than a stationary process could describe. The formalism is not very much more complicated than for Wiener theory, but is best left to the references (see e.g., Wunsch, 1966, Chapter 6). But a sketch of a simplified case may perhaps give some feeling for it.

Suppose we have a "model" for calculating how x_m will behave over time. Let us assume, for simplicity, that it is just an AR(1) process

$$x_m = ax_{m-1} + \theta_m. \tag{6.7}$$

Suppose we have an estimate of $x_{m-1,}$ called $\tilde{x}_{m-1}$, with an estimated error

$$P_{m-1} =< \left(\tilde{x}_{m-1} - x_{m-1}\right)^2 > . \tag{6.8}$$

Then we can make a forecast of x_m,

$$\tilde{x}_m\left(-\right) = \tilde{a}x_{m-1} \tag{6.9}$$

because θ_m is unknown and completely unpredictable by assumption. The minus sign in the argument indicates that no observation from time m has been used. The prediction error is now

$$P_m\left(-\right) =< \left(\tilde{x}_m\left(-\right) - x_m\right)^2 >= \sigma_\theta^2 + P_{m-1}, \tag{6.10}$$

that is, the initial error propagates forward in time and is additive to the new error from the unknown θ_m. Now let us further assume that we have a measurement of x_m but one which has noise in it,

$$y_m = Ex_m + \varepsilon_m, \tag{6.11}$$

where $< \varepsilon_m >= 0, < \varepsilon_m^2 >= \sigma_\varepsilon^2$, which produces an estimate y_m/E, with error variance $E^{-2}\sigma_\varepsilon^2$. The observation of x_m ought to permit us to improve upon our forecast of it, $\tilde{x}_m\left(-\right)$. A plausible idea is to *average the measurement with the forecast*, weighting the two inversely as their relative errors:

$$\tilde{x}_m = \frac{\sigma_\theta^2 + P_{m-1}}{\left(\sigma_\theta^2 + P_{m-1}\right) + E^{-2}\sigma_\varepsilon^2} E^{-1} y_m + \frac{E^{-2}\sigma_\varepsilon^2}{\left(\sigma_\theta^2 + P_{m-1}\right) + E^{-2}\sigma_\varepsilon^2} \tilde{x}_m\left(-\right) \tag{6.12}$$

$$= \tilde{x}_m\left(-\right) + \frac{\left(\sigma_\theta^2 + P_{m-1}\right) E^{-1}}{\left(\sigma_\theta^2 + P_{m-1}\right) + E^{-2}\sigma_\varepsilon^2} \left(y_m - E\tilde{x}_m\left(-\right)\right) \tag{6.13}$$

(See Wunsch, 1996, section 3.7). If the new data are very poor relative to the forecast, $\sigma_\varepsilon^2 \to \infty$, the estimate reduces to the forecast. In the opposite limit when $\left(\sigma_\theta^2 + P_{m-1}\right) >> E^{-2}\sigma_\varepsilon^2$, the new data give a much better estimate, and it can be confirmed that $\tilde{x}_m \to y_m/E$ as is also sensible.

The expected error of the average is

$$P_m = \left[\left(E^{-2} \sigma_\varepsilon^2 \right)^{-1} + \left(\sigma_\theta^2 + P_{m-1} \right)^{-1} \right]^{-1} \tag{6.14}$$

$$= \left(\sigma_\theta^2 + P_m \right) - \left(\sigma_\theta^2 + {P_{m-1}}^2 \right)^2 \left[\left(\sigma_\theta^2 + P_{m-1} \right) + E^{-2} \sigma_\varepsilon^2 \right]^{-1}, \tag{6.15}$$

which should be studied in the two above limits. Now we proceed another time-step into the future. The new best forecast is,

$$\tilde{x}_{m+1} = a \tilde{x}_m \tag{6.16}$$

where $\tilde{x}_m$ has error P_m and we proceed just as before, with $m \to m+1$. If there are no observations available at time m, then the forecast cannot be improved, $\tilde{x}_{m+1} = a\tilde{x}_m$, and one keeps going with $\tilde{x}_{m+2} = a\tilde{x}_{m+1}$. The Kalman filter permits one to employ whatever information is contained in a model (perhaps dynamical, as in orbital equations or an ocean circulation model) along with any observations of the elements of the system that come in through time. The idea is extremely powerful and many thousands of papers and books have been written on it and its generalizations.[1] If there is a steady data stream, and the model satisfies certain requirements, one can show that the Kalman filter asymptotically reduces to the Wiener filter—an important conclusion because the Kalman formalism is generally computationally much more burdensome than is the Wiener one. The above derivation contains the essence of the filter, the only changes required for the more general case being the replacement of the scalar state $x(t)$ by a vector state, $\mathbf{x}(t)$, and with the covariances becoming matrices whose reciprocals are inverses and one must keep track of the order of operations.

6.3. Wiener Smoother. "Filtering" in the technical sense involves using the present and past, perhaps infinitely far back, of a time-series so as to produce a best estimate of the value of a signal or to predict the time series. In this situation, the future values of x_m are unavailable. In most oceanographic problems however, we have a stored time series and the formal future is available. It is unsurprising that one can often make better estimates using future values than if they are unavailable (just as interpolation is more accurate than extrapolation). When the filtering problem is recast so as to employ both past and future values, one is doing "smoothing". Formally, in equations such as (6.5, 6.6), one permits the index m to take on negative values and finds the new normal equations.

7. Gauss-Markov Theorem

It is often the case that one seeks a signal in a noise background. Examples would be the determination of a sinusoid in the presence of a background continuum stochastic process, and the determination of a trend in the presence of other processes with a different structure. A perhaps less obvious example, because it is too familiar, is the estimation of the mean value of a time series, in which the mean is the signal and the deviations from the mean are therefore arbitrarily defined as the "noise." The cases one usually encounters in elementary textbooks and classes are some simple variant of a signal in presence of

[1] Students may be interested to know that it widely rumored that Kalman twice failed his MIT general exams (in EECS).

white noise. Unfortunately, while this case is easy to analyze, it usually does not describe the situation one is faced with in practice, when for example, one must try to determine a trend in the presence of red noise processes, ones which may exhibit local trends easily confused with a true secular (deterministic) trend. This problem plagues climate studies.

There are several approaches to finding some machinery which can help one to fall into statistical traps of confusing signal with noise. One widely applicable methodology is called variously "minimum variance estimation", the "Gauss-Markov Theorem", and in a different context, sometimes the "stochastic inverse." Although a more extended discussion is given in Wunsch (1996), or see Liebelt (1967) for a complete derivation, the following heuristic outline may help.

Let there be some set of N unknown parameters, written here as a vector, $\mathbf{s}$. We suppose that they have zero mean, $< \mathbf{s} >= \mathbf{0}$, and that there is a known covariance for $\mathbf{s}$, $\mathbf{R}_{ss} = \left\langle \mathbf{s}\mathbf{s}^T \right\rangle$. Let there be a set of M measurements, written here also as another vector $\mathbf{y}$, also with zero mean, and known covariance $< \mathbf{y}\mathbf{y}^T >= \mathbf{R}_{yy}$. Finally, we will assume that there is a known covariance between the measurements and the unknowns: $\mathbf{R}_{sy} =< \mathbf{s}\mathbf{y}^T >$. Given these covariances (correlations), what can we say about $\mathbf{s}$, given $\mathbf{y}$, if we try to estimate any of the elements s_i, as a linear combination of the measurements:

$$\tilde{s}_i = \sum_{j=1}^{M} B_{ij} y_j. \tag{7.1}$$

The question is, how should the weights, B_{ij} be chosen? Arbitrarily, but reasonably, we seek B_{ij} such that the variance of the estimated s_i about the true value is as small as possible, that is,

$$\left\langle (\tilde{s}_i - s_i)^2 \right\rangle = \left\langle \left(\sum_{j=1}^{m} B_{ij} y_j - s_i \right)^2 \right\rangle, \; 1 \leq i \leq N \tag{7.2}$$

should be a minimum. Before proceeding, note that B_{ij} is really a matrix, and each row will be separately determinable for each element s_i. This simple observation permits us to re-write (7.2) in a matrix-vector form. Minimize the *diagonal elements* of:

$$< (\tilde{\mathbf{s}} - \mathbf{s}) (\tilde{\mathbf{s}} - \mathbf{s})^T >=< (\mathbf{B}\mathbf{y} - \mathbf{s}) (\mathbf{B}\mathbf{y} - \mathbf{s})^T >\equiv \mathbf{P}. \tag{7.3}$$

The important point here is that, in (7.3), we are meant to minimize the N separate diagonal elements, each separately determining a row of $\mathbf{B}$; but we can use the notation to solve for all rows simultaneously.

At this stage, one expands the matrix product in (7.3), and uses the fact that quantities such as $< \mathbf{s}\mathbf{s}^T >= \mathbf{R}_{ss}$ are known. One can show without too much difficulty (it involves invoking the properties of positive definite matrices) that the minimum of the diagonals is given by the unique choice,

$$\mathbf{B} = \mathbf{R}_{sy}\mathbf{R}_{yy}^{-1}, \tag{7.4}$$

with the first row being the solution for $\tilde{s}_1$, etc.

The result (7.4) is general and abstract. Let us now consider a special case in which the measurements y_q are some linear combination of the parameters, corrupted by noise, that is, $y_q = \sum_{l=1}^{N} E_{ql} s_l + n_q$, which

can also be written generally as,

$$\mathbf{Es} + \mathbf{n} = \mathbf{y}. \tag{7.5}$$

With this assumption, we can evaluate

$$\mathbf{R}_{sy} =< \mathbf{s}\left(\mathbf{Es} + \mathbf{n}\right)^T >= \mathbf{R}_{ss}\mathbf{E}^T, \tag{7.6}$$

assuming $< \mathbf{sn}^T >= 0$, and

$$\mathbf{R}_{yy} = \mathbf{E}\mathbf{R}_{ss}\mathbf{E}^T + \mathbf{R}_{nn} \tag{7.7}$$

where $\mathbf{R}_{nn} =< \mathbf{nn}^T >$. Then one has immediately,

$$\mathbf{B} = \mathbf{R}_{ss}\left(\mathbf{E}\mathbf{R}_{ss}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1}, \tag{7.8}$$

and

$$\tilde{\mathbf{s}} = \mathbf{R}_{ss}\left(\mathbf{E}\mathbf{R}_{ss}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1}\mathbf{y}. \tag{7.9}$$

There is one further, extremely important step: how good is this estimate? This question is readily answered by substituting the value of $\mathbf{B}$ back into the expression (7.3) for the actual covariance about the true value. We obtain immediately,

$$\mathbf{P} = \mathbf{R}_{ss} - \mathbf{R}_{ss}\mathbf{E}^T\left(\mathbf{E}\mathbf{R}_{ss}\mathbf{E}^T + \mathbf{R}_{nn}\right)\mathbf{E}\mathbf{R}_{ss}. \tag{7.10}$$

One of the benefits of the general approach is that we have obtain the complete matrix $\mathbf{P}$, which gives us not only the variances (uncertainties) of each of the $\tilde{s}_i$ about the true value, but also the covariances of these errors or uncertainties in each, with all the others—they do after all, depend upon the same data—so that it is no surprise that they would have correlated errors.)

A special case, written out in Wunsch (1996), and which is particularly illuminating is the simple problem of determining a mean value (so that $\mathbf{s}$ is a scalar), in the presence of a noise field which has an arbitrary correlation. One finds there, that the uncertainty of the mean can be vastly greater than the conventional estimates based upon white noise, if the noise is correlated in time.

REMARK 2. *A common complaint among beginning users of Gauss-Markov and related estimation methods is: "I have no idea what the covariances are. This all becomes completely arbitrary if I just make up something." The answer to this worry is found by examining the statement "I have no idea what the covariances are." If this is really true, it means that an acceptable answer for any element, $\tilde{s}_i$ could have any value at all, including something infinitesimal, 10^{-40}, or astronomical, $\pm 10^{40}$ and one would say "that's acceptable, because I know nothing at all about the solution". The reader may say, "that's not what I really meant". In fact, it is extremely rare to be completely ignorant, and if one is completely ignorant, so that any values at all would be accepted, the investigator ought perhaps to stop and ask if the problem makes any sense? More commonly, one usually knows* something, *e.g., that the parameters are very unlikely to be bigger than about $\pm S_0$. If that is all one is willing to say, then one simply takes*

$$\mathbf{R}_{ss} = S_0^2\mathbf{I}_N \tag{7.11}$$

with something analogous perhaps, for $\mathbf{R}_{nn}$*, which becomes an estimate of the noise magnitude. Letting* $S_0^2 \to \infty$ *is the appropriate limit if one really knows nothing, and one might study 7.10) in that limit. The point is, that the estimation procedure can use whatever information one has, and one need not stipulate anything that is truly unknown. In particular, if one does not know the non-diagonal elements of the covariances, one need not state them. All that will happen is that a solution will be obtained that likely will have a larger uncertainty than one could have obtained had additional information been available. The more information one can provide, the better the estimate. A final comment of course, is that one* must *check that the solution and the errors left are actually consistent with what was postulated for the covariances. If the solution and noise residuals are clearly inconsistent with* $\mathbf{R}_{ss}$*, etc., one should trash the solution and try to understand what was wrong; this is a very powerful test, neglected at one's peril. If used, it can rescue even the naivest user from a truly silly result.*

8. Trend Determination

A common debate in climate and other studies concerns the reality of trends in data. Specifically, one is concerned that an apparent linear or more complex trend should be distinguished, as secular, from local apparent trends which are nothing but the expected short-term behavior of a stochastic process. The synthetic time series displayed above show extended periods where one might be fooled into inferring a trend in climate, when it is nothing but a temporary structure occurring from pure happenstance. For geophysical problems, the existence of rednoise time series makes the problem quite difficult.

Suppose one has a stationary time series x_q whose power density and corresponding autocovariance $R(\tau)$ are known. From $R(\tau)$ we can make a covariance matrix as in (1.16). We suspect that superimposed upon this time series is a linear secular trend representable as $y_q = a + bq$ and we would like to determine a, b and their uncertainty. Generalizing the discussion in Wunsch (1996, p.188), we regard x_q now as a noise process and $a + bq$ as a signal model. We can represent $a + bq + x_q = g_q$, or

$$\mathbf{Da} + \mathbf{x} = \mathbf{g}, \tag{8.1}$$

where,

$$\mathbf{a} = [a, b]^T, \mathbf{D} = \left\{ \begin{array}{cc} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ . & . \\ 1 & N-1 \end{array} \right\}, \mathbf{g} = [g_0, g_1, ..., g_{N-1}]^T . \tag{8.2}$$

Suppose that no a priori statistical information about a, b is available (that is, we would accept arbitrarily large or small values as the solution). Then the Gauss-Markov Theorem produces a best-estimate

$$\tilde{\mathbf{a}} = \left[\begin{array}{c} \tilde{a} \\ \tilde{b} \end{array} \right] = \left[\mathbf{D}^T \mathbf{R} \mathbf{D}\right]^{-1} \mathbf{D}^T \mathbf{R}^{-1} \mathbf{g} \tag{8.3}$$

with uncertainty

$$\mathbf{P} = < (\tilde{\mathbf{a}} - \mathbf{a})^2 > = \left(\mathbf{D}^T \mathbf{R}^{-1} \mathbf{D}\right)^{-1}. \tag{8.4}$$

Clearly $\mathbf{P}$ depends directly upon the covariance $\mathbf{R}$. If long-temporal correlations are present, apparent, but spurious trends, will probably be found. But the result (8.3) will have large expected uncertainties given by (8.4) and one would not be misled.

Exercise. Generate a time series with power density $\Phi(s) = 1/(5/4 + \cos(2\pi s))$. Add a known trend, and then determine it from the above expressions.

9. EOFs, SVD

A common statistical tool in oceanography, meteorology and climate research are the so-called empirical orthogonal functions (EOFs). Anyone, in any scientific field, working with large amounts of data having covariances, is almost inevitably led to EOFs as an obvious tool for reducing the number of data one must work with, and to help in obtaining insight into the meaning of the covariances that are present. The ubiquity of the tool means, unfortunately, that it has been repeatedly reinvented in different scientific fields, and the inventors were apparently so pleased with themselves over their cleverness, they made no attempt to see if the method was already known elsewhere. The consequence has been a proliferation of names for the same thing: EOFs, principal components, proper orthogonal decomposition, singular vectors, Karhunen-Loève functions, optimals, etc. (I'm sure this list is incomplete.)

The method, and its numerous extensions, is a useful one (but like all powerful tools, potentially dangerous to the innocent user), and a brief discussion is offered here. The most general approach of which I am aware, is that based upon the so-called singular value decomposition (e.g., Wunsch, 1996 and references there). Let us suppose that we have a field which varies, e.g., in time and space. An example (often discussed) is the field of seasurface temperature (SST) in the North Pacific Ocean. We suppose that through some device (ships, satellites), someone has mapped the anomaly of SST monthly over the entire North Pacific Ocean at 1° lateral resolution for 100 years. Taking the width of the Pacific Ocean to be 120° and the latitude range to be 60° each map would have approximately $60 \times 120 = 7200$ gridded values, and there would be 12×100 of these from 100 years. The total volume of numbers would then be about 7200×1200 or about 9 million numbers.

A visual inspection of the maps (something which is *always* the first step in any data analysis), would show that the fields evolve only very slowly from month-to-month in an annual cycle, and in some respects, from year-to-year, and that much, but perhaps not all, of the structure occurs on a spatial scale large compared to the 1° gridding. Both these features suggest that the volume of numbers is perhaps much greater than really necessary to describe the data, and that there are elements of the spatial structure which seem to covary, but with different features varying on different time scales. A natural question then, is whether there is not a tool which could simultaneously reduce the volume of data, and inform

one about which patterns dominated the changes in space and time? One might hope to make physical sense of the latter.

Because there is such a vast body of mathematics available for matrices, consider making a matrix out of this data set. One might argue that each map is already a matrix, with latitude and longitude comprising the rows and columns, but it suits our purpose better to make a single matrix out of the entire data set. Let us do this by making one large column of the matrix out of each map, in some way that is arbitrary, but convenient, e.g., by stacking the values at fixed longitudes in one long column, one under the other (we could even have a random rule for where in the column the values go, as long it is the same for each time—this would just make it hard to figure out what value was where). Then each column is the map at monthly intervals, with 1200 columns. Call this matrix $\mathbf{E}$, which is of dimension M =number of latitudes times the number of longitudes by N, the number of observation times (that is, it is probably not square).

We now postulate that any matrix $\mathbf{E}$ can be written

$$\mathbf{E} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{9.1}$$

that is as the product of three matrices. $\mathbf{\Lambda}$ is a $M \times N$ diagonal matrix (in a generalized sense for a non-square matrix). Matrix $\mathbf{U}$ is square of dimension M, and $\mathbf{V}$ is square of dimension N. $\mathbf{U}, \mathbf{V}$ have the special properties of being "orthogonal",

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}_M, \mathbf{U}^T\mathbf{U} = \mathbf{I}_M, \mathbf{V}\mathbf{V}^T = \mathbf{I}_N, \mathbf{V}^T\mathbf{V} = \mathbf{I}_N \tag{9.2}$$

that is to say, in particular the columns of $\mathbf{U}$ are mutually orthonormal, as are the columns of $\mathbf{V}$ (so are the rows, but that proves less important). $\mathbf{I}_N$ is the identity matrix of dimension N, etc. The matrices $\mathbf{U}, \mathbf{V}$,$\mathbf{\Lambda}$ can be shown, with little difficulty to be determined by the following relations:

$$\mathbf{E}\mathbf{E}^T\mathbf{u}_i = \lambda_i^2\mathbf{u}_i, 1 \le i \le M,\ \mathbf{E}^T\mathbf{E}\mathbf{v}_i = \lambda_i^2\mathbf{v}_i, 1 \le i \le N. \tag{9.3}$$

That is to say, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{E}\mathbf{E}^T$, and the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{E}^T\mathbf{E}$. They are related to each other through the relations,

$$\mathbf{E}\mathbf{v}_i = \lambda_i\mathbf{u}_i, 1 \le i \le N,\ \mathbf{E}\mathbf{u}_i = \lambda_i\mathbf{v}_i, 1 \le i \le M. \tag{9.4}$$

Note that in (9.3,9.4), M, N are in general different, and the only way these relationships can be consistent would be if all of the $\lambda_i = 0, i > \min(M, N)$ (this is the maximum number of non-zero eigenvalues; there may be fewer). By convention, the λ_i and their corresponding $\mathbf{u}_i, \mathbf{v}_i$ are ordered in decreasing value of the λ_i.

Consider $\mathbf{E}^T\mathbf{E}$ in (9.3) This new matrix is formed by taking the dot product of all of the columns of $\mathbf{E}$ with each other in sequence. That is to say, $\mathbf{E}^T\mathbf{E}$ is, up to a normalization factor of $1/M$, the covariance of each anomaly map with every other anomaly map and is thus a covariance matrix of the observations through time and the $\mathbf{v}_i$ are the eigenvectors of this covariance matrix. Alternatively, $\mathbf{E}\mathbf{E}^T$ is the dot product of each row of the maps with each other, and up to a normalization of $1/N$ is the covariance of

the structure at each location in the map with that at every other point on the map; the $\mathbf{u}_i$ are thus the eigenvectors of this covariance matrix.

Consider by way of example, $\mathbf{E}^T\mathbf{E}$. This is a square, non-negative definite matrix (meaning its eigenvalues are all non-negative, a good thing, since the eigenvalues are the λ_i^2, which we might hope would be a positive number). From (9.1, 9.2),

$$\mathbf{E}^T\mathbf{E} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T = \sum_{i=1}^{N} \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^T, \tag{9.5}$$

Eq. (9.5) is an example of the statement that a square, symmetric matrix can be represented exactly in terms of its eigenvectors. Suppose, only $K \leq N$ of the λ_i are non-zero. Then the sum reduces to,

$$\mathbf{E}^T\mathbf{E} = \sum_{j=1}^{K} \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^T = \mathbf{V}_K \mathbf{\Lambda}_K^2 \mathbf{V}_K^T, \tag{9.6}$$

where $\mathbf{\Lambda}_K$ is truncated to its first K rows and columns (is now square) and $\mathbf{V}_K$ contains only the first k columns of $\mathbf{V}$. Now suppose that some of the λ_i are very small compared, e.g., to the others. Let there be K' of them, much larger than the others. The question then arises as to whether the further truncated expression,

$$\mathbf{E}^T\mathbf{E} \sim \sum_{i=1}^{K'} \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^T = \mathbf{V}_{K'} \mathbf{\Lambda}_{K'}^2 \mathbf{V}_{K'}^T, \tag{9.7}$$

is not still a good approximation to $\mathbf{E}^T\mathbf{E}$? Here, $\mathbf{V}_{K'}$ consists only of its first K' columns. The assumption/conclusion that the truncated expansion (9.7) is a good representation of the covariance matrix $\mathbf{E}^T\mathbf{E}$, with $K' << K$ is the basis of the EOF idea. Conceivably K' is as small as 1 or 2, even when there may be hundreds or thousands of vectors $\mathbf{v}_i$ required for an exact result. An exactly parallel discussion applies to the covariance matrix $\mathbf{E}\mathbf{E}^T$ in terms of the $\mathbf{u}_i$.

There are several ways to understand and exploit this type of result. Let us go back to (9.1). Assuming that there are K non-zero λ_i, it can be confirmed (by just writing it out) that

$$\mathbf{E} = \mathbf{U}_K \mathbf{\Lambda}_K \mathbf{V}_K^T = \sum_{i=1}^{K} \lambda_i \mathbf{u}_i \mathbf{v}_i^T \tag{9.8}$$

exactly. This result says that an arbitrary $M \times N$ matrix $\mathbf{E}$ can be represented exactly by at most K pairs of orthogonal vectors, where $K \leq \min(M, N)$. Suppose further, that some of the λ_i are very small compared to the others Then one might suppose that a good approximation to $\mathbf{E}$ is

$$\mathbf{E} \sim \mathbf{U}_K \mathbf{\Lambda}_{K'} \mathbf{V}_{K'}^T = \sum_{i=1}^{K'} \lambda_i \mathbf{u}_i \mathbf{v}_i^T. \tag{9.9}$$

If this is a good approximation, and $K' << K$, and because $\mathbf{E}$ are the actual data, it is possible that only a very small number of orthogonal vectors is required to reproduce all of the significant structure in the data. Furthermore, the covariances of the data are given by simple expressions such as (9.7) in terms of these same vectors.

The factorizations (9.1) or the alternative (9.8) are known as the "singular value decomposition". The λ_i are the "singular values", and the pairs $(\mathbf{u}_i, \mathbf{v}_i)$ are the singular vectors. Commonly, the $\mathbf{v}_i$ are identified as the EOFs, but they can equally well be identified as the $\mathbf{u}_i$; the choice is arbitrary, depending only upon how one seeks to interpret the data.

Eq. (9.9) can be discussed slightly differently. Suppose that one has an arbitrary $\mathbf{E}$. Then if one seeks to represent it in L pairs of orthonormal vectors $(\mathbf{q}_i, \mathbf{r}_i)$

$$\mathbf{E} \approx \sum_{i=1}^{L} \alpha_i \mathbf{q}_i \mathbf{r}_i^T, \tag{9.10}$$

then the so-called Eckart-Young-Mirsky theorem (see references in Wunsch, 1996) states that the best choice (in the sense of making the norm of the difference between the left and right-hand sides as small as possible), is for the$(\mathbf{q}_i, \mathbf{r}_i)$ to be the first L singular vectors, and $\alpha_i = \lambda_i$.

Exercise. Interpret the Karhunen-Loève expansion and singular spectrum analysis in the light of the SVD.

Exercise. (a) Consider a travelling wave $y(r,t) = \sin(kr + \sigma t)$, which is observed at a zonal set of positions, $r_j = (j-1)\Delta r$ at times $t_p = (p-1)\Delta t$. Choose, $k, \sigma, \Delta r, \Delta t$ so that the frequency and wavenumber are resolved by the time/space sampling. Using approximately 20 observational positions and enough observation times to obtain several temporal periods, apply the SVD/EOF analysis to the resulting observations. Discuss the singular vectors which emerge. Confirm that the SVD at rank 2 perfectly reproduces all of the data. The following, e.g., would do (in MATLAB)

```
» x=[0:30]';t=[0:256]';
» [xx,tt]=meshgrid(x,t);
» sigma=2*pi/16;k=2*pi/10;
» Y=sin(k*xx+sigma*tt);
» contourf(Y);colorbar;
```

(b) Now suppose two waves are present: $y(r,t) = \sin(kr + \sigma t) + \sin((k/2)\, r + (\sigma/2)\, t)$. What are the EOFs now? Can you deduce the presence of the two waves and their frequencies/wavenumbers? (c) Repeat the above analysis except take the observation positions r_j to be irregularly spaced. What happens to the EOFs? (d) What happens if you add a white noise to the observations?

REMARK 3. *The very large literature on and the use of EOFs shows the great value of this form of representation. But clearly many of the practitioners of this form of analysis make the often implicit assumption that the various EOFs/singular vectors necessarily correspond to some form of normal mode or simple physical pattern of change. There is usually no basis for this assumption, although one can be lucky. Note in particular, that the double orthogonality (in space and time) of the resulting singular vectors may necessarily require the lumping together of real normal modes, which are present, in various*

linear combinations required to enforce the orthogonality. The general failure of EOFs to correspond to physically interpretable motions is well known in statistics (see, e.g., Jolliffe, 1986). A simple example of the failure of the method to identify physical modes is given in Wunsch (1997, Appendix).

Many extensions and variations of this method are available, including e.g., the introduction of phase shifted values (Hilbert transforms) with complex arithmetic, to display more clearly the separation between standing and travelling modes, and various linear combinations of modes. Some of these are described e.g., by von Storch and Zwiers (1999). Statistics books should be consulted for the determination of the appropriate rank and a discussion of the uncertainty of the results.