Time Series Analysis. A Heuristic Primer.

Class Notes for 12.864 Inference from Data and Models.

March 30, 2005

Carl Wunsch

Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology
Cambridge MA 02139 USA, Telephone: (617) 253-5937, Fax: (617) 253-4464
E-mail address: cwunsch@mit.edu

Contents

1. Preface

Time series analysis is a sub-field of statistical estimation methods. It is a mature subject with a long history and very large literature. For anyone dealing with processes evolving in time and/or space, it is an essential tool, but one usually given short-shrift in oceanographic, meteorological and climate courses. It is difficult to overestimate the importance of a zero-order understanding of these statistical tools for anyone involved in studying climate change, the nature of a current meter record, or even the behavior of a model.

There are many good textbooks in this field, and the refusal of many investigators to invest the time and energy to master a few simple elements is difficult to understand. These notes are not meant to be a substitute for a serious textbook; rather they are intended, partly through a set of do-it-yourself exercises, to communicate some of the basic concepts, which should at least prevent the reader from the commonest blunders now plaguing much of the literature. Many of the examples used here are oceanographic, or climate-related in origin, but no particular knowledge of these fields is required to follow the arguments.

Two main branches of time series analysis exist. Branch 1 is focussed on methodologies applied in the time domain (I will use "time" as a generic term for both time or space dimensions.) and the second branch employs frequency (wavenumber) domain analysis tools. The two approaches are intimately related and equivalent and the differences should not be overemphasized, but one or the other sometimes proves more convenient or enlightening in a particular situation. Frequency domain methods employ (mostly) Fourier series and transforms. *Algorithmically,* one can identify two distinct eras: those before and after the (re-) discovery of the Fast Fourier transform (FFT) algorithm about 1966. For numerical purposes, with some very narrow exceptions (described later), the pre-FFT computer implementations are now obsolete and there is no justification for their continued use.

Out of the huge literature on time-series analysis, I would recommend Bracewell (1978) for its treatment of Fourier analysis, Percival and Walden (1993) for spectra, and Priestley (1981) as a general broad reference incorporating both mathematical and practical issues. (Percival and Walden do not treat coherence, whereas Priestley does). Among the older books (pre-FFT), Jenkins and Watts (1968) is outstanding and still highly useful for the basic concepts. For time-domain methods, Box, Jenkins and Reinsel (1994) is generally regarded as the standard. Another comprehensive text is Hamilton (1994), with a heavy economics emphasis, but covering some topics not contained in the other books. Study of one or more of these books is essential to anyone seriously trying to master time series methods.

Additional copies of these notes can be obtained through MIT's OpenCourseware project (http://ocw.mit.edu). Please report to me the inevitable remaining errors you may encounter (cwunsch@mit.edu).

CHAPTER 1

Frequency Domain Formulation

1. Fourier Transforms and Delta Functions

"Time" is the physical variable, written as t, although it may well be a spatial coordinate. Let $x(t), y(t)$, etc. be real, continuous, well-behaved functions. The meaning of "well-behaved" is not so-clear. For Fourier transform purposes, it classically meant among other requirements, that

$$\int_{-\infty}^{\infty} |x(t)|^2 < \infty. \tag{1.1}$$

Unfortunately such useful functions as $x(t) = \sin(2\pi t/T)$, or

$$\begin{aligned} x(t) = H(t) &= 0, t < 0 \\ &= 1, t \geq 0 \end{aligned} \tag{1.2}$$

are excluded (the latter is the unit step or Heaviside function). We succeed in including these and other useful functions by admitting the existence and utility of Dirac δ-functions. (A textbook would specifically exclude functions like $\sin(1/t)$. In general, such functions do not appear as physical signals and I will rarely bother to mention the rigorous mathematical restrictions on the various results.)

The Fourier transform of $x(t)$ will be written as

$$\mathcal{F}(x(t)) \equiv \hat{x}(s) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i s t} dt. \tag{1.3}$$

It is often true that

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(s) e^{2\pi i s t} ds \equiv \mathcal{F}^{-1}(\hat{x}(s)). \tag{1.4}$$

Other conventions exist, using radian frequency ($\omega = 2\pi s$), and/or reversing the signs in the exponents of (1.3, 1.4). All are equivalent (I am following Bracewell's convention).

Exercise. The Fourier transform pair (1.3, 1.4) is written in complex form. Re-write it as cosine and sine transforms where all operations are real. Discuss the behavior of $\hat{x}(s)$ when $x(t)$ is an even and odd function of time.

Define $\delta(t)$ such that

$$x(t_0) = \int_{-\infty}^{\infty} x(t) \delta(t_0 - t) dt \tag{1.5}$$

It follows immediately that

$$\mathcal{F}(\delta(t)) = 1 \tag{1.6}$$

and therefore that

$$\delta(t) = \int_{-\infty}^{\infty} e^{2\pi i s t} ds = \int_{-\infty}^{\infty} \cos(2\pi s t)\, ds. \tag{1.7}$$

Notice that the $\delta-$function has units; Eq. (1.5) implies that the units of $\delta(t)$ are $1/t$ so that the equation works dimensionally.

Definition. A "sample" value of $x(t)$ is $x(t_m)$, the value at the specific time $t = t_m$.

We can write, in seemingly cumbersome fashion, the sample value as

$$x(t_m) = \int_{-\infty}^{\infty} x(t)\, \delta(t_m - t)\, dt \tag{1.8}$$

This expression proves surprisingly useful.

Exercise. With $x(t)$ real, show that

$$\hat{x}(-s) = \hat{x}(s)^* \tag{1.9}$$

where * denotes the complex conjugate.

Exercise. a is a constant. Show that

$$\mathcal{F}(x(at)) = \frac{1}{|a|}\hat{x}\left(\frac{s}{a}\right). \tag{1.10}$$

This is the scaling theorem.

Exercise. Show that

$$\mathcal{F}(x(t-a)) = e^{-2\pi i a s}\hat{x}(s). \tag{1.11}$$

(shift theorem).

Exercise. Show that

$$\mathcal{F}\left(\frac{dx(t)}{dt}\right) = 2\pi i s \hat{x}(s). \tag{1.12}$$

(differentiation theorem).

Exercise. Show that

$$\mathcal{F}(x(-t)) = \hat{x}(s)^* \tag{1.13}$$

(time-reversal theorem)

Exercise. Find the Fourier transforms of $\cos 2\pi s_0 t$ and $\sin 2\pi s_0 t$. Sketch and describe them in terms of real, imaginary, even, odd properties.

Exercise. Show that if $x(t) = x(-t)$, that is, x is an "even-function", then

$$\hat{x}(s) = \hat{x}(-s), \tag{1.14}$$

and that it is real. Show that if $x(t) = -x(-t)$, (an "odd-function"), then

$$\hat{x}(s) = \hat{x}(-s)^*, \tag{1.15}$$

and it is pure imaginary.

Note that any function can be written as the sum of an even and odd-function

$$\begin{aligned} x(t) &= x_e(t) + x_o(t) \\ x_e(t) &= \frac{1}{2}\left(x(t) + x(-t)\right),\ x_o(t) = \frac{1}{2}\left(x(t) - x(-t)\right). \end{aligned} \tag{1.16}$$

Thus,

$$\hat{x}(s) = \hat{x}_e(s) + \hat{x}_o(s). \tag{1.17}$$

There are two fundamental theorems in this subject. One is the proof that the transform pair (1.3,1.4) exists. The second is the so-called convolution theorem. Define

$$h(t) = \int_{-\infty}^{\infty} f(t')\, g(t - t')\, dt' \tag{1.18}$$

where $h(t)$ is said to be the "convolution" of f with g. The convolution theorem asserts:

$$\hat{h}(s) = \hat{f}(s)\, \hat{g}(s). \tag{1.19}$$

Convolution is so common that one often writes $h = f * g$. Note that it follows immediately that

$$f * g = g * f. \tag{1.20}$$

Exercise. Prove that (1.19) follows from (1.18) and the definition of the Fourier transform. What is the Fourier transform of $x(t)\, y(t)$?

Exercise. Show that if,

$$h(t) = \int_{-\infty}^{\infty} f(t')\, g(t + t')\, dt' \tag{1.21}$$

then

$$\hat{h}(s) = \hat{f}(s)^*\, \hat{g}(s) \tag{1.22}$$

$h(t)$ is said to be the "cross-correlation" of f and g, written here as $h = f \otimes g$. Note that $f \otimes g \neq g \otimes f$.

If $g = f$, then (1.21) is called the "autocorrelation" of f (a better name is "autocovariance", but the terminology is hard to displace), and its Fourier transform is,

$$\hat{h}(s) = \left|\hat{f}(s)\right|^2 \tag{1.23}$$

and is called the "power spectrum" of $f(t)$.

Exercise: Find the Fourier transform and power spectrum of

$$\Pi(t) = \begin{cases} 1, & |t| \leq 1/2 \\ 0, & |t| > 1/2. \end{cases} \tag{1.24}$$

Now do the same, using the scaling theorem, for $\Pi(t/T)$. Draw a picture of the power spectrum.

One of the fundamental Fourier transform relations is the Parseval (sometimes, Rayleigh) relation:

$$\int_{-\infty}^{\infty} x(t)^2\, dt = \int_{-\infty}^{\infty} |\hat{x}(s)|^2\, ds. \tag{1.25}$$

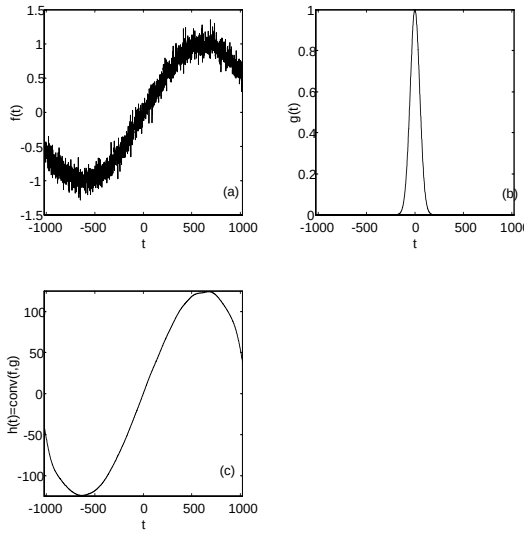

FIGURE 1. An effect of convolution is for a "smooth" function to reduce the high frequency oscillations in the less smooth function. Here a noisy curve (a) is convolved with a smoother curve (b) to produce the result in (c), where the raggedness of the the noisy function has been greatly reduced. (Technically here, the function in (c) is a low-pass filter. No normalization has been imposed however; consider the magnitude of (c) compared to (a).)

Exercise. Using the convolution theorem, prove (1.25).

Exercise. Using the definition of the $\delta-$function, and the differentiation theorem, find the Fourier transform of the Heaviside function $H(t)$. Now by the same procedure, find the Fourier transform of the sign function,

$$\text{signum}(t) = \text{sgn}(t) = \begin{cases} -1, & t < 0 \\ 1, & t > 0 \end{cases}, \tag{1.26}$$

and compare the two answers. Can both be correct? Explain the problem. (Hint: When using the differentiation theorem to deduce the Fourier transform of an integral of another function, one must be aware of integration constants, and in particular that functions such as $s\delta(s) = 0$ can always be added to a result without changing its value.) Solution:

$$\mathcal{F}(\text{sgn}(t)) = \frac{-i}{\pi s}. \tag{1.27}$$

Often one of the functions $f(t)$, $g(t)$ is a long "wiggily" curve, (say) $g(t)$ and the other, $f(t)$ is comparatively simple and compact, for example as shown in Fig. 1 The act of convolution in this situation tends to subdue the oscillations in and other structures in $g(t)$. In this situation $f(t)$ is usually called a "filter", although which is designated as the filter is clearly an arbitrary choice. Filters exist for and are designed for, a very wide range of purposes. Sometimes one wishes to change the frequency

content of $g(t)$, leading to the notion of high-pass, low-pass, band-pass and band-rejection filters. Other filters are used for prediction, noise suppression, signal extraction, and interpolation.

Exercise. Define the "mean" of a function to be,

$$m = \int_{-\infty}^{\infty} f(t)\, dt, \tag{1.28}$$

and its "variance",

$$(\Delta t)^2 = \int_{-\infty}^{\infty} (t-m)^2 f(t)\, dt. \tag{1.29}$$

Show that

$$\Delta t \Delta s \geq \frac{1}{4\pi}. \tag{1.30}$$

This last equation is known as the "uncertainty principle" and occurs in quantum mechanics as the Heisenberg Uncertainty Principle, with momentum and position being the corresponding Fourier transform domains. You will need the Parseval Relation, the differentiation theorem, and the Schwarz Inequality:

$$\left| \int_{-\infty}^{\infty} f(t)\, g(t)\, dt \right|^2 \leq \int_{-\infty}^{\infty} |f(t)|^2\, dt \int_{-\infty}^{\infty} |g(t)|^2\, dt. \tag{1.31}$$

The uncertainty principle tells us that a narrow function must have a broad Fourier transform, and vice-versa with "broad" being defined as being large enough to satisfy the inequality. Compare it to the scaling theorem. Can you find a function for which the inequality is actually equality?

1.1. The Special Role of Sinusoids. One might legitimately inquire as to why there is such a specific focus on the sinusoidal functions in the analysis of time series? There are, after all, many other possible basis functions (Bessel, Legendre, etc.). One important motivation is their role as the eigenfunctions of extremely general linear, time-independent systems. Define a a linear system as an operator $\mathcal{L}(\cdot)$ operating on any input, $x(t)$. $\mathcal{L}$ can be a physical "black-box" (an electrical circuit, a pendulum, etc.), and/or can be described via a differential, integral or finite difference, operator. $\mathcal{L}$ operates on its input to produce an output:

$$y(t) = \mathcal{L}(x(t), t). \tag{1.32}$$

It is "time-independent" if $\mathcal{L}$ does not depend explicitly on t, and it is linear if

$$\mathcal{L}(ax(t) + w(t), t) = a\mathcal{L}(x(t), t) + \mathcal{L}(w(t), t) \tag{1.33}$$

for any constant a. It is "causal" if for $x(t) = 0, t < t_0$, $\mathcal{L}(x(t)) = 0, t < t_0$. That is, there is no response prior to a disturbance (most physical systems satisfy causality).

Consider a general time-invariant linear system, subject to a complex periodic input:

$$y(t) = \mathcal{L}\left(e^{2\pi i s_0 t}\right). \tag{1.34}$$

Suppose we introduce a time shift,

$$y(t + t_0) = \mathcal{L}\left(e^{2\pi i s_0 (t+t_0)}\right). \tag{1.35}$$

Now set $t = 0$, and

$$y(t_0) = \mathcal{L}\left(e^{2\pi i s_0 t_0}\right) = e^{2\pi i s_0 t_0}\mathcal{L}\left(e^{2\pi i_0 s t=0}\right) = e^{2\pi i s_0 t_0}\mathcal{L}(1). \tag{1.36}$$

Now $\mathcal{L}(1)$ is a constant (generally complex). Thus (1.36) tells us that for an input function $e^{2\pi i s_0 t_0}$, with both s_0, t_0 completely arbitrary, the output must be another pure sinusoid—at exactly the same period—subject only to a modfication in amplitude and phase. This result is a direct consequence of the linearity and time-independence assumptions. Eq. (1.36) is also a statement that any such exponential is thereby an eigenfunction of $\mathcal{L}$, with eigenvalue $\mathcal{L}(1)$. It is a very general result that one can reconstruct arbitrary linear operators from their eigenvalues and eigenfunctions, and hence the privileged role of sinusoids; in the present case, that reduces to recognizing that the Fourier transform of $y(t_0)$ would be that of $\mathcal{L}$ which would thereby be fully determined. (One can carry out the operations leading to (1.36) using real sines and cosines. The algebra is more cumbersome.)

1.2. Causal Functions and Hilbert Transforms. Functions that vanish before $t = 0$ are said to be "causal". By a simple shift in origin, any function which vanishes for $t < t_0$ can be reduced to a causal one, and it suffices to consider only the special case, $t_0 = 0$. The reason for the importance of these functions is that most physical systems obey a causality requirement that they should not produce any output, before there is an input. (If a mass-spring oscillator is at rest, and then is disturbed, one does not expect to see it move before the disturbance occurs.) Causality emerges as a major requirement for functions which are used to do prediction—they cannot operate on future observations, which do not yet exist, but only on the observed past.

Consider therefore, any function $x(t) = 0, t < 0$. Write it as the sum of an even and odd-function,

$$x(t) = \begin{cases} x_e(t) + x_o(t) = \frac{1}{2}(x(t) + x(-t)) + \frac{1}{2}(x(t) - x(-t)) \\ = 0, \ t < 0, \end{cases} \tag{1.37}$$

but neither $x_e(t)$, nor $x_o(t)$ vanishes for $t < 0$, only their sum. It follows from (1.37) that

$$x_o(t) = sgn(t)\, x_e(t) \tag{1.38}$$

and that

$$x(t) = (1 + sgn(t))\, x_e(t). \tag{1.39}$$

Fourier transforming (1.39), and using the convolution theorem, we have

$$\hat{x}(s) = \hat{x}_e(s) + \frac{-i}{\pi s} * \hat{x}_e(s) \tag{1.40}$$

using the Fourier transform for $sgn(t)$.

Because $\hat{x}_e(s)$ is real, the imaginary part of $\hat{x}(s)$ must be

$$\mathrm{Im}(\hat{x}(s)) = \hat{x}_o(s) = \frac{-1}{\pi s} * \hat{x}_e(s) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\hat{x}_e(s')}{s - s'} ds'. \tag{1.41}$$

Re-writing (1.39) in the obvious way in terms of $x_o(t)$, we can similarly show,

$$\hat{x}_e(s) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\hat{x}_o(s')}{s-s'}ds'. \tag{1.42}$$

Eqs. (1.41, 1.42) are a pair, called Hilbert transforms. Causal functions thus have intimately connected real and imaginary parts of their Fourier transforms; knowledge of one determines the other. These relationships are of great theoretical and practical importance. An oceanographic application is discussed in Wunsch (1972).

The Hilbert transform can be applied in the time domain to a function $x(t)$, whether causal or not. Here we follow Bendat and Piersol (1986, Chapter 13). Define

$$x^H(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(t')}{t-t'}dt' \tag{1.43}$$

and $x(t)$ can be recovered from $x^H(t)$ by the inverse Hilbert transform (1.42). Eq. (1.43) is the convolution

$$x^H(t) = x(t) * \frac{1}{\pi t} \tag{1.44}$$

and by the convolution theorem,

$$\hat{x}^H(s) = \hat{x}(s)\left(-i\,\mathrm{sgn}(s)\right) \tag{1.45}$$

using the Fourier transform of the signum function. The last expression can be re-written as

$$\hat{x}^H(s) = \hat{x}(s)\begin{cases} \exp(-i\pi/2), & s<0 \\ \exp(i\pi/2), & s>0 \end{cases}, \tag{1.46}$$

that is, the Hilbert transform in time is equivalent to phase shifting the Fourier transform of $x(t)$ by $\pi/2$ for positive frequencies, and by -$\pi/2$ for negative ones. Thus $x^H(t)$ has the same frequency content of $x(t)$, but is phase shifted by 90°. It comes as no surprise therefore, that if e.g., $x(t) = \cos(2\pi s_0 t)$, then $x^H(t) = \sin(2\pi s_0 t)$. Although we do not pursue it here (see Bendat and Piersol, 1986), this feature of Hilbert transformation leads to the idea of an "analytic signal",

$$y(t) = x(t) + ix^H(t) \tag{1.47}$$

which proves useful in defining an "instantaneous frequency", and in studying the behavior of wave propagation including the idea (taken up much later) of complex empirical orthogonal functions.

Writing the inverse transform of a causal function,

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(s)\, e^{2\pi i s t} ds, \tag{1.48}$$

one might, if $\hat{x}(s)$ is suitably behaved, attempt to evaluate this transform by Cauchy's theorem, as

$$x(t) = \begin{cases} 2\pi i \sum(\text{residues of the lower half-}s\text{-plane, }), & t<0 \\ 2\pi i \sum(\text{residues of the upper half-}s\text{-plane, }), & t>0 \end{cases} \tag{1.49}$$

Since the first expression must vanish, if $\hat{x}(s)$ is a rational function, it cannot have any poles in the lower-half-$s-$plane; this conclusion leads immediately so-called Wiener filter theory, and the use of Wiener-Hopf methods.

1.3. Asymptotics. The gain of insight into the connections between a function and its Fourier transform, and thus developing intuition, is a very powerful aid to interpreting the real world. The scaling theorem, and its first-cousin, the uncertainty principle, are part of that understanding. Another useful piece of information concerns the behavior of the Fourier transform as $|s| \to \infty$. The classical result is the Riemann-Lebesgue Lemma. We can write

$$\hat{f}(s) = \int_{-\infty}^{\infty} f(t)\, e^{-2\pi i s t} dt. \tag{1.50}$$

where here, $f(t)$ is assumed to satisfy the classical conditions for existence of the Fourier transform pair. Let $t' = t - 1/(2s)$, (note the units are correct) then by a simple change of variables rule,

$$\hat{f}(s) = -\int_{-\infty}^{\infty} f\left(t' + \frac{1}{2s}\right) e^{-2\pi i s t'} dt' \tag{1.51}$$

$(\exp(-i\pi) = -1)$ and taking the average of these last two expressions, we have,

$$\begin{aligned} \left|\hat{f}(s)\right| &= \left|\frac{1}{2}\int_{-\infty}^{\infty} f(t)\, e^{-2\pi i s t} dt - \frac{1}{2}\int_{-\infty}^{\infty} f\left(t + \frac{1}{2s}\right) e^{-2\pi i s t} dt\right| \\ &\leq \frac{1}{2}\int_{-\infty}^{\infty} \left|f(t) - f\left(t + \frac{1}{2s}\right)\right| dt \to 0, \text{ as } s \to \infty \end{aligned} \tag{1.52}$$

because the difference between the two functions becomes arbitrarily small with increasing $|s|$.

This result tells us that for classical functions, we are assured that for sufficiently large $|s|$ the Fourier transform will go to zero. It doesn't however, tell us how fast it does go to zero. A general theory is provided by Lighthill (1958), which he then builds into a complete analysis system for asymptotic evaluation. He does this essentially by noting that functions such as $H(t)$ have Fourier transforms which for large $|s|$ are dominated by the contribution from the discontinuity in the first derivative, that is, for large $s, H(s) \to 1/s$ (compare to $signum(t)$). Consideration of functions whose first derivatives are continuous, but whose second derivatives are discontinuous, shows that they behave as $1/s^2$ for large $|s|$; in general if the $n-th$ derivative is the first discontinuous one, then the function behaves asymptotically as $1/|s|^n$. These are both handy rules for what happens and useful for evaluating integrals at large s (or large distances if one is going from Fourier to physical space). Note that even the δ-function fits: its $0-$th derivative is discontinuous (that is, the function itself), and its asymptotic behavior is $1/s^0 =$ constant; it does not decay at all as it violates the requirements of the Riemann-Lebesgue lemma.

2. Fourier Series and Time-Limited Functions

Suppose $x(t)$ is periodic:

$$x(t) = x(t+T) \tag{2.1}$$

Define the complex Fourier *coefficients* as

$$\alpha_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t) \exp\left(\frac{-2\pi i n t}{T}\right) dt \tag{2.2}$$

Then under very general conditions, one can represent $x(t)$ in a Fourier Series:

$$x(t) = \sum_{n=-\infty}^{\infty} \alpha_n \exp\left(\frac{2\pi i n t}{T}\right). \tag{2.3}$$

Exercise. Write $x(t)$ as a Fourier cosine and sine series.

The Parseval Theorem for Fourier series is

$$\frac{1}{T}\int_{-T/2}^{T/2} x(t)^2 \, dt = \sum_{n=-\infty}^{\infty} |a_n|^2, \tag{2.4}$$

and which follows immediately from the orthogonality of the complex exponentials over interval T.

Exercise. Prove the Fourier Series versions of the shift, differentiation, scaling, and time-reversal theorems.

Part of the utility of $\delta-$functions is that they permit us to do a variety of calculations which are not classically permitted. Consider for example, the Fourier transform of a periodic function, e.g., any $x(t)$ as in Eq. (2.3),

$$\hat{x}(s) = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \alpha_n e^{(2\pi i n t/T)} e^{-2\pi i s t} dt = \sum_{n=-\infty}^{\infty} \alpha_n \delta(s - n/T), \tag{2.5}$$

ignoring all convergence issues. We thus have the nice result that a periodic function has a Fourier transform; it has the property of vanishing except precisely at the usual Fourier series frequencies where its value is a $\delta-$function with amplitude equal to the complex Fourier series coefficient at that frequency.

Suppose that instead,

$$x(t) = 0, \ |t| \geq T/2 \tag{2.6}$$

that is, $x(t)$ is zero except in the finite interval $-T/2 \leq t \leq T/2$ (this is called a "time-limited" function). The following elementary statement proves to be very useful. Write $x(t)$ as a Fourier series in $|t| < T/2$, and as zero elsewhere:

$$x(t) = \begin{cases} \sum_{n=-\infty}^{\infty} \alpha_n \exp(2\pi i n t/T), & |t| \leq T/2 \\ 0, & |t| > T/2 \end{cases} \tag{2.7}$$

where

$$\alpha_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t) \exp\left(-\frac{2\pi i n t}{T}\right) dt, \tag{2.8}$$

as though it were actually periodic. Thus as defined, $x(t)$ corresponds to some different, periodic function, in the interval $|t| \leq T/2$, and is zero outside. $x(t)$ is perfectly defined by the special sinusoids with frequency $s_n = n/T$, $n = 0, \pm 1, ... \infty$.

The function $x(t)$ isn't periodic and so its Fourier transform can be computed in the ordinary way,

$$\hat{x}(s) = \int_{-T/2}^{T/2} x(t) e^{-2\pi i s t} dt. \tag{2.9}$$

and then,

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(s) e^{2\pi i s t} ds. \tag{2.10}$$

We observe that $\hat{x}(s)$ is defined at *all* frequencies s, on the continuum from 0 to $\pm\infty$. If we look at the special frequencies $s = s_n = n/T$, corresponding to the Fourier series representation (2.7), we observe that

$$\hat{x}(s_n) = T\alpha_n = \frac{1}{1/T}\alpha_n. \tag{2.11}$$

That is, the Fourier transform at the special Fourier series frequencies, differs from the corresponding Fourier series coefficient by a constant multiplier. The second equality in (2.11) is written specifically to show that the Fourier transform value $\hat{x}(s)$, can be thought of as an amplitude density per unit frequency, with the α_n being separated by $1/T$ in frequency.

The information content of the representation of $x(t)$ in (2.7) must be the same as in (2.10), in the sense that $x(t)$ is perfectly recovered from both. But there is a striking difference in the apparent efficiency of the forms: the Fourier series requires values (a real and an imaginary part) at a countable infinity of frequencies, while the Fourier transform requires a value on the line continuum of all frequencies. One infers that the sole function of the infinite continuum of values is to insure what is given by the second line of Eq. (2.7): that the function vanishes outside $|t| \leq T/2$. Some thought suggests the idea that one ought to be able to calculate the Fourier transform at *any* frequency, from its values at the special Fourier series frequencies, and this is both true, and a very powerful tool.

Let us compute the Fourier transform of $x(t)$, using the form (2.7):

$$\hat{x}(s) = \int_{-T/2}^{T/2} \sum_{n=-\infty}^{\infty} \alpha_n \exp\left(\frac{2\pi i n t}{T}\right) \exp(2\pi i s t)\, dt \tag{2.12}$$

and assuming we can interchange the order of integration and summation (we can),

$$\begin{aligned} \hat{x}(s) &= T \sum_{n=-\infty}^{\infty} \alpha_n \frac{\sin(\pi T(n/T - s))}{\pi T(n/T - s)} \\ &= \sum_{n=-\infty}^{\infty} \hat{x}(s_n) \frac{\sin(\pi T(n/T - s))}{\pi T(n/T - s)}, \end{aligned} \tag{2.13}$$

using Eq. (2.11). Notice that as required $\hat{x}(s) = \hat{x}(s_n) = T\alpha_n$,when $s = s_n$, but in between these values, $\hat{x}(s)$ is a weighted (interpolated) linear combination of all of the Fourier Series components.

Exercise. Prove by inverse Fourier transformation that any sum

$$\hat{x}(s) = \sum_{n=-\infty}^{\infty} \beta_n \frac{\sin(\pi T(n/T - s))}{\pi T(n/T - s)}, \tag{2.14}$$

where β_n are arbitrary constants, corresponds to a function vanishing $t > |T/2|$, that is, a time-limited function.

The surprising import of (2.13) is that the Fourier transform of a time-limited function can be perfectly reconstructed from a knowledge of its values at the Fourier series frequencies alone. That means, in turn, that a knowledge of the countable infinity of Fourier coefficients can reconstruct the original function exactly. Putting it slightly differently, *there is no purpose in computing a Fourier transform at frequency intervals closer than* $1/T$ *where* T *is either the period, or the interval of observation.*

3. The Sampling Theorem

We have seen that a time-limited function can be reconstructed from its Fourier coefficients. The reader will probably have noticed that there is symmetry between frequency and time domains. That is to say, apart from the assignment of the sign of the exponent of $\exp(2\pi ist)$, the s and t domains are essentially equivalent. For many purposes, it is helpful to use not t, s with their physical connotations, but abstract symbols like q, r. Taking the lead from this inference, let us interchange the t, s domains in the equations (2.6, 2.13), making the substitutions $t \to s, s \to t, T \to 1/\Delta t, \hat{x}(s) \to x(t)$.. We then have,

$$\hat{x}(s) = 0, s \geq 1/2\Delta t \tag{3.1}$$

$$\begin{aligned} x(t) &= \sum_{m=-\infty}^{\infty} x(m\Delta t) \frac{\sin(\pi(m - t/\Delta t))}{\pi(m - t/\Delta t)} \\ &= \sum_{m=-\infty}^{\infty} x(m\Delta t) \frac{\sin((\pi/\Delta t)(t - m\Delta t))}{(\pi/\Delta t)(t - m\Delta t)}. \end{aligned} \tag{3.2}$$

This result asserts that a function *bandlimited* to the frequency interval $|s| \leq 1/2\Delta t$ (meaning that its Fourier transform vanishes for all frequencies outside of this *baseband*) can be perfectly reconstructed by samples of the function at the times $m\Delta t$. This result (3.1,3.2) is the famous Shannon sampling theorem. As such, it is an *interpolation statement.* It can also be regarded as a statement of information content: all of the information about the bandlimited continuous time series is contained in the samples. This result is actually a remarkable one, as it asserts that a continuous function with an uncountable infinity of points can be reconstructed from a countable infinity of values.

Although one should never use (3.2) to interpolate data in practice (although so-called *sinc* methods are used to do numerical integration of analytically-defined functions), the implications of this rule are very important and can be stated in a variety of ways. In particular, let us write a general bandlimiting form:

$$\hat{x}(s) = 0, s \geq s_c \tag{3.3}$$

If (3.3) is valid, it *suffices* to sample the function at *uniform* time intervals $\Delta t \leq 1/2s_c$ (Eq. 3.1 is clearly then satisfied.).

Exercise. Let $\Delta t = 1$. $x(t)$ is measured at all times, and found to vanish, except for $t = m = 0, 1, 2, 3$ and the values are $[1, 2, -1, -1]$. Calculate the values of $x(t)$ at intervals $\Delta t/10$ from $-5 \leq t \leq 5$ and plot it. Find the Fourier transform of $x(t)$.

The consequence of the sampling theorem for discrete observations in time is that there is no purpose in calculating the Fourier transform for frequencies larger in magnitude than $1/(2\Delta t)$. Coupled with the result for time-limited functions, we conclude that *all of the information about a finite sequence of N observations at intervals Δt and of duration, $(N-1)\Delta t$ is contained in the baseband $|s| \leq 1/2\Delta t$, at frequencies $s_n = n/(N\Delta t)$.*

There is a theorem (owing to Paley and Wiener) that a time-limited function cannot be band-limited, and vice-versa. One infers that a truly time-limited function must have a Fourier transform with non-zero values extending to arbitrarily high frequencies, s. If such a function is sampled, then some degree of aliasing is inevitable. For a truly band-limited function, one makes the required interchange to show that it must actually extend with finite values to $t = \pm\infty$. Some degree of aliasing of real signals is therefore inevitable. Nonetheless, such aliasing can usually be rendered arbitrarily small and harmless; the need to be vigilant is, however, clear.

3.1. Tapering, Leakage, Etc. Suppose we have a continuous cosine $x(t) = \cos(2\pi p_1 t/T_1)$. Then the true Fourier transform is

$$\hat{x}(s) = \frac{1}{2}\{\delta(s - p_1) + \delta(s + p_1)\}. \tag{3.4}$$

If it is observed (continuously) over the interval $-T/2 \leq t \leq T/2$, then we have the Fourier transform of

$$x_\Pi(t) = x(t)\,\Pi(t/T) \tag{3.5}$$

and which is found immediately to be

$$\hat{x}_\Pi(s) = \frac{T}{2}\left\{\frac{\sin(\pi T(s - p_1))}{(\pi T(s - p_1))} + \frac{\sin(\pi T(s + p_1))}{(\pi T(s + p_1))}\right\} \tag{3.6}$$

The function

$$sinc(Ts) = \sin(\pi T s)/(\pi T s), \tag{3.7}$$

plotted in Fig. (2), is ubiquitous in time series analysis and worth some study. Note that in (3.6) there is a "main-lobe" of width $2/T$ (defined by the zero crossings) and with amplitude maximum T. To each side of the main lobe, there is an infinite set of diminishing "sidelobes" of width $1/T$ between zero crossings. Let us suppose that p_1 in (3.4) is chosen to be one of the special frequencies $s_n = n/T, T = N\Delta t$, in particular, $p_1 = p/T$. Then (3.6) is a sum of two *sinc* functions centered at $s_p = \pm p/T$. A very important feature is that each of these functions vanishes identically at all other special frequencies $s_n, n \neq p$. If

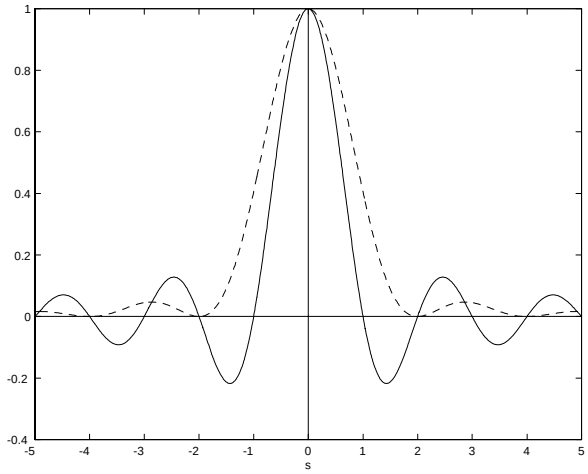

FIGURE 2. The function, $sinc(sT) = \sin(\pi sT)/(\pi sT)$, (solid line) which is the Fourier transform of a pure exponential centered at that corresponding frequency. Here $T = 1$. Notice that the function crosses zero whenever $s = m$, which corresponds to the Fourier frequency separation. The main lobe has width 2, while successor lobes have width 1, with a decay rate only as fast as $1/|s|$. The function $sinc^2(s/2)$ (dotted line) decays as $1/|s|^2$, but its main lobe appears, by the scaling theorem, with twice the width of that of the $sinc(s)$ function.

we confine ourselves, as the inferences of the previous section imply, to computing the Fourier transform at only these special frequencies, we would see only a large value T at $s = s_p$ and zero at every other such frequency. (Note that if we convert to Fourier coefficients by division by $1/T$, we obtain the proper values.) The Fourier transform does not vanish for the continuum of frequencies $s \neq s_n$, but it could be obtained from the sampling theorem.

Now suppose that the cosine is no longer a Fourier harmonic of the record length. Then computation of the Fourier transform at s_n no longer produces a zero value; rather one obtains a finite value from (3.6). In particular, if p_1 lies halfway between two Fourier harmonics, $n/T \leq p_1 \leq (n+1)/T$, $|\hat{x}(s_n)|, |\hat{x}(s_{n+1})|$ will be approximately equal, and the absolute value of the remaining Fourier coefficients will diminish roughly as $1/|n-m|$. The words "approximately" and "roughly" are employed because there is another *sinc* function at the corresponding negative frequencies, which generates finite values in the positive half of the $s-$axis. The analyst will not be able to distinguish the result (a single pure Fourier frequency *in between* s_n, s_{n+1}) from the possibility that there are *two* pure frequencies present at s_n, s_{n+1}. Thus we have what is sometimes called "Rayleigh's criterion": that to separate, or "resolve" two pure sinusoids, at frequencies p_1, p_2, their frequencies must differ by

$$|p_1 - p_2| \geq \frac{1}{T}, . \tag{3.8}$$

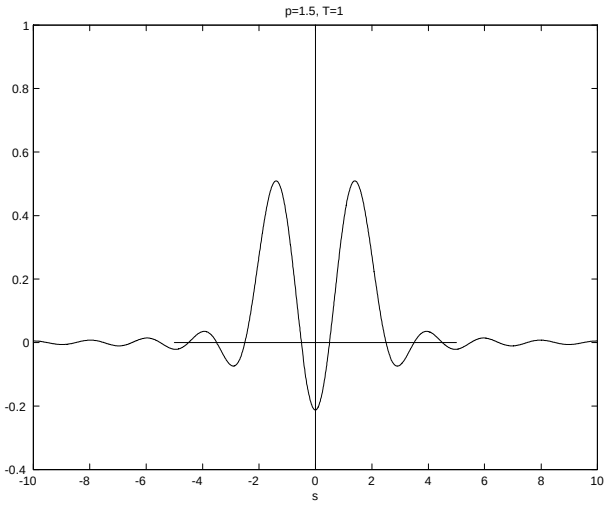

FIGURE 3. Interference pattern from a cosine, showing how contributions from positive and negative frequencies add and subtract. Each vanishes at the central frequency plus $1/T$ and at all other intervals separated by $1/T$.

or precisely by a Fourier harmonic; see Fig. 3. (The terminology and criterion originate in spectroscopy where the main lobe of the *sinc* function is determined by the width, L, of a physical slit playing the role of T.)

The appearance of the *sinc* function in the Fourier transform (and series) of a finite length record has some practical implications (note too, that the sampling theorem involves a sum over *sinc* functions). Suppose one has a very strong sinusoid of amplitude A, at frequency p, present in a record, $x(t)$ whose Fourier transform otherwise has a magnitude which is much less than A. If one is attempting to estimate $\hat{x}(s)$ apart from the sinusoid, one sees that the influence of A (from both positive and negative frequency contributions) will be additive and can seriously corrupt $\hat{x}(s)$ even at frequencies far from $s = p$. Such effects are known as "leakage". There are basically three ways to remove this disturbance. (1) Subtract the sinusoid from the data prior to the Fourier analysis. This is a very common procedure when dealing, e.g., with tides in sealevel records, where the frequencies are known in advance to literally astronomical precision, and where $|\hat{x}(s_p)|^2 \approx |A^2|$ may be many orders of magnitude larger than its value at other frequencies. (2) Choose a record length such that $p = n/T$; that is, make the sinusoid into a Fourier harmonic and rely on the vanishing of the *sinc* function to suppress the contribution of A at all other frequencies. This procedure is an effective one, but is limited by the extent to which finite word length computers can compute the zeros of the *sinc* and by the common problem (e.g., again for tides) that several pure frequencies are present simultaneously and not all can be rendered simultaneously as Fourier harmonics. (3) Taper the record. Here one notes that the origin of the leakage problem is that the *sinc* diminishes only as $1/s$ as one moves away from the central frequency. This slow reduction is in turn

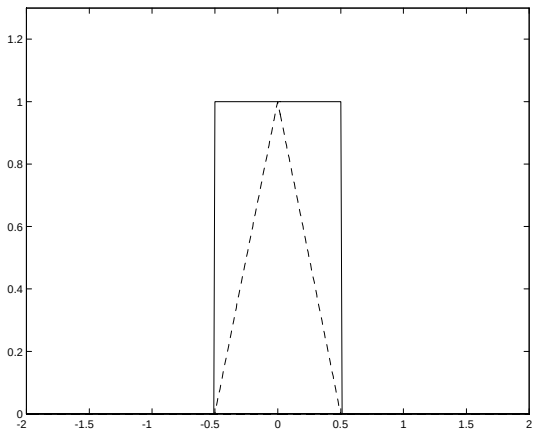

FIGURE 4. "Tophat", or $\Pi(t)$ (solid) and "triangle" or $\Lambda(t/2)$. A finite record can be regarded as the product $x(t)\Pi(t/T)$, giving rise to the *sinc* pattern response. If this finite record is tapered by multiplying it as $x(t)\Lambda(t/(2T))$, the Fourier transform decays much more rapidly away from the central frequency of any sinusoids present.

easily shown to arise because the Π function in (3.5) has finite steps in value (recall the Riemann-Lebesgue Lemma.)

Suppose we "taper" $x_\Pi(t)$, by multiplying it by the triangle function (see Bracewell, 1978, and Fig. 4),

$$\Lambda(t) = 1 - |t|, t \leq 1 \tag{3.9}$$

$$= 0, |t| > 1 \tag{3.10}$$

whose first derivative, rather than the function itself is discontinuous. The Fourier transform

$$\hat{\Lambda}(s) = \frac{\sin^2(\pi s)}{(\pi s)^2} = sinc^2(s) \tag{3.11}$$

is plotted in Fig. 2). As expected, it decays as $1/s^2$. Thus if we Fourier transform

$$x_\Lambda(t) = x(t)\Lambda(t/(T/2)) \tag{3.12}$$

the pure cosine now gives rise to

$$\hat{x}_\Lambda(s) = \frac{T}{2}\left\{\frac{\sin^2((\pi/2)T(s-p_1))}{((\pi/2)T(s-p_1))^2} + \frac{\sin^2((\pi/2)T(s+p_1))}{((\pi/2)T(s+p_1))^2}\right\} \tag{3.13}$$

and hence the leakage diminishes much more rapidly, whether or not we have succeeded in aligning the dominant cosine. A price exists however, which must be paid. Notice that the main lobe of $\mathcal{F}(\Lambda(t/(T/2)))$ has width not $2/T$, but $4/T$, that is, it is twice as wide as before, and the resolution of the analysis would be 1/2 of what it was without tapering. Thus tapering the record prior to Fourier analysis incurs a trade-off between leakage and resolution.

One might sensibly wonder if some intermediate function between the Π and Λ functions exists so that one diminishes the leakage but without incurring a resolution penalty as large as a factor of 2. The answer is "yes"; much effort has been made over the years to finding tapers $w(t)$, whose Fourier transforms $\hat{W}(s)$ have desirable properties. Such taper functions are called "windows". A common one tapers the ends by multiplying by half-cosines at either end, cosines whose periods are a parameter of the analysis. Others go under the names of Hamming, Hanning, Bartlett, etc. windows.

Later we will see that a sophisticated choice of windows leads to the elegant recent theory of multitaper spectral analysis. At the moment, we will only make the observation that the Λ taper and all other tapers, has the effect of throwing away data near the ends of the record, a process which is always best regarded as perverse: one should not have to discard good data for a good analysis procedure to work.

Although we have discussed leakage etc. for continuously sampled records, completely analogous results exist for sampled, finite, records. We leave further discussion to the references.

Exercise. Generate a pure cosine at frequency s_1, and period $T_1 = 2\pi/s_1$. Numerically compute its Fourier transform, and Fourier series coefficients, when the record length, T =integer $\times T_1$, and when it is no longer an integer multiple of the period.

4. Discrete Observations

4.0.1. *Sampling.* The above results show that a band-limited function can be reconstructed perfectly from an infinite set of (perfect) samples. Similarly, the Fourier transform of a time-limited function can be reconstructed perfectly from an infinite number of (perfect) samples (the Fourier Series frequencies). In observational practice, functions must be both band-limited (one cannot store an infinite number of Fourier coefficients) and time-limited (one cannot store an infinite number of samples). Before exploring what this all means, let us vary the problem slightly. Suppose we have $x(t)$ with Fourier transform $\hat{x}(s)$ and we sample $x(t)$ at uniform intervals $m\Delta t$ without paying attention, initially, as to whether it is band-limited or not. What is the relationship between the Fourier transform of the sampled function and that of $x(t)$? That is, the above development does not tell us how to compute a Fourier transform from a set of samples. One could use (3.2), interpolating before computing the Fourier integral. As it turns out, this is unnecessary.

We need some way to connect the sampled function with the underlying continuous values. The $\delta-$function proves to be the ideal representation. Eq. (2.13) produces a single sample at time t_m. The quantity,

$$x_{III}(t) = x(t) \sum_{n=-\infty}^{\infty} \delta(t - n\Delta t), \tag{4.1}$$

vanishes except at $t = q\Delta t$ for any integer q. The value associated with $x_{III}(t)$ at that time is found by integrating it in an infinitesimal interval $-\varepsilon + q\Delta t \leq t \leq \varepsilon + q\Delta t$ and one finds immediately that $x_{III}(q\Delta t) = x(q\Delta t)$. Note that *all* measurements are integrals over some time interval, no matter how

short (perhaps nanoseconds). Because the $\delta-$function is infinitesimally broad in time, the briefest of measurement integrals is adequate to assign a value.[1].

Let us Fourier analyze $x_{III}(t)$, and evaluate it in two separate ways:

(I) Direct sum.

$$\begin{aligned}\hat{x}_{III}(s) &= \int_{-\infty}^{\infty} x(t) \sum_{m=-\infty}^{\infty} \delta(t - m\Delta t)\, e^{-2\pi i s t} dt \\ &= \sum_{m=-\infty}^{\infty} x(m\Delta t)\, e^{-2\pi i s m \Delta t}. \end{aligned} \tag{4.2}$$

(II) By convolution.

$$\hat{x}_{III}(s) = \hat{x}(s) * \mathcal{F}\left(\sum_{m=-\infty}^{\infty} \delta(t - m\Delta t) \right). \tag{4.3}$$

What is $\mathcal{F}\left(\sum_{m=-\infty}^{\infty} \delta(t - m\Delta t)\right)$? We have, by direct integration,

$$\mathcal{F}\left(\sum_{m=-\infty}^{\infty} \delta(t - m\Delta t) \right) = \sum_{m=-\infty}^{\infty} e^{-2\pi i m s \Delta t} \tag{4.4}$$

What function is this? The right-hand-side of (4.4) is clearly a Fourier series for a function periodic with period $1/\Delta t$, in s. I assert that the periodic function is $\Delta t \delta(s)$, and the reader should confirm that computing the Fourier series representation of $\Delta t \delta(s)$ in the $s-$domain, with period $1/\Delta t$ is exactly (4.4). But such a periodic $\delta-$function can also be written[2]

$$\Delta t \sum_{n=-\infty}^{\infty} \delta(s - n/\Delta t) \tag{4.5}$$

Thus (4.3) can be written

$$\begin{aligned}\hat{x}_{III}(s) &= \hat{x}(s) * \Delta t \sum_{n=-\infty}^{\infty} \delta(s - n/\Delta t) \\ &= \int_{-\infty}^{\infty} \hat{x}(s')\, \Delta t \sum_{n=-\infty}^{\infty} \delta(s - n/\Delta t - s')\, ds' \\ &= \Delta t \sum_{n=-\infty}^{\infty} \hat{x}\left(s - \frac{n}{\Delta t}\right) \end{aligned} \tag{4.6}$$

We now have two apparently very different representations of the Fourier transform of a sampled function. (I) Asserts two important things. The Fourier transform can be computed as the naive discretization of the complex exponentials (or cosines and sines if one prefers) times the sample values. Equally important, the result is a periodic function with period $1/\Delta t$. (Figure 5). Form (II) tells us that

[1] $\delta-$functions are meaningful only when integrated. Lighthill (1958) is a good primer on handling them. Much of the book has been boiled down to the advice that, if in doubt about the meaning of an integral, "integrate by parts".

[2] Bracewell (1978) gives a complete discussion of the behavior of these othewise peculiar functions. Note that we are ignoring all questions of convergence, interchange of summation and integration etc. Everything can be justified by appropriate limiting processes.

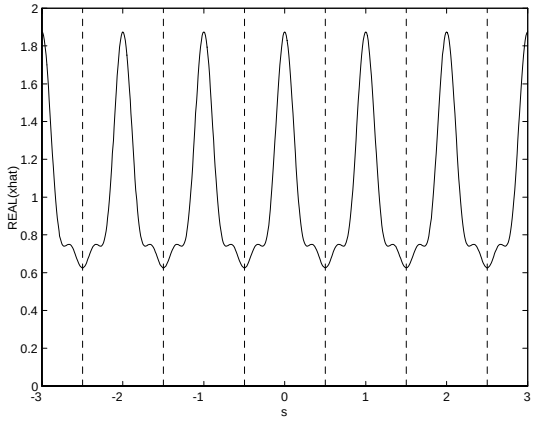

FIGURE 5. Real part of the periodic Fourier transform of a sampled function. The baseband is defined as $-1/2\Delta t \leq s \leq 1/2\Delta t$, (here $\Delta t = 1$), but any interval of width $1/\Delta t$ is equivalent. These intervals are marked with vertical lines.

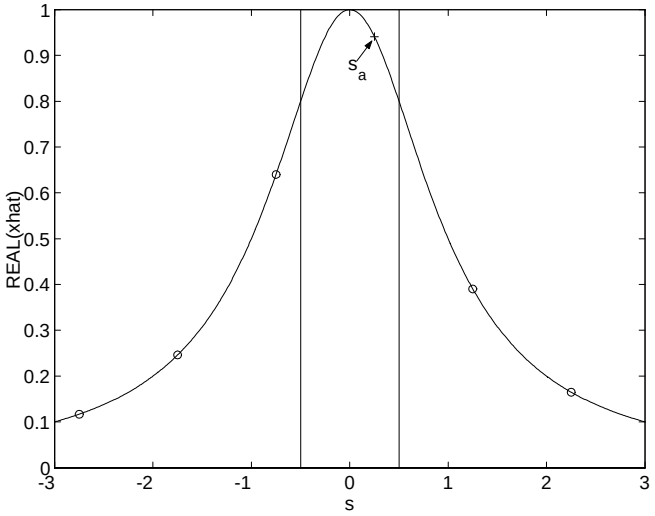

FIGURE 6. s_a is the position where all Fourier transform amplitudes from the Fourier transform values indicated by the dots (Eq. 4.6) will appear. The baseband is indicated by the vertical lines and any non-zero Fourier transform values outside this region will alias into it.

the value of the Fourier transform at a particular frequency s is not in general equal to $\hat{x}(s)$. Rather it is the sum of *all* values of $\hat{x}(s)$ separated by frequency $1/\Delta t$. (Figure 6). This second form is clearly periodic with period $1/\Delta t$, consistent with (I).

Because of the periodicity, we can confine ourselves for discussion, to one interval of width $1/\Delta t$. By convention we take it symmetric about $s = 0$, in the range $-1/(2\Delta t) \leq s \leq 1/(2\Delta t)$ which we call the

baseband. We can now address the question of when $\hat{x}_{III}(s)$ in the baseband will be equal to $\hat{x}(s)$? The answer follows immediately from (4.6): if, and only if, $\hat{x}(s)$ vanishes for $s \geq |1/2\Delta t|$. That is, the Fourier transform of a sampled function will be the Fourier transform of the original continuous function only if the original function is bandlimited and Δt is chosen to be small enough such that $\hat{x}(|s| > 1/\Delta t) = 0$. We also see that *there is no purpose in computing* $\hat{x}_{III}(s)$ *outside the baseband:* the function is perfectly periodic. We could use the sampling theorem to interpolate our samples before Fourier transforming. But that would produce a function which vanished outside the baseband—and we would be no wiser.

Suppose the original continuous function is

$$x(t) = A \sin(2\pi s_0 t). \tag{4.7}$$

It follows immediately from the definition of the $\delta-$function that

$$\hat{x}(s) = \frac{i}{2}\{\delta(s + s_0) - \delta(s - s_0)\}. \tag{4.8}$$

If we choose $\Delta t < 1/2s_0$, we obtain the $\delta-$functions in the baseband at the correct frequency. We ignore the $\delta-$functions outside the baseband because we know them to be spurious. But suppose we choose, either knowing what we are doing, or in ignorance, $\Delta t > 1/2s_0$. Then (4.6) tells us that it will appear, spuriously, at

$$s = s_a = s_0 \pm m/\Delta t,\ |s_a| \leq 1/2\Delta t \tag{4.9}$$

thus determining m. The phenomenon of having a periodic function appear at an incorrect, lower frequency, because of insufficiently rapid sampling, is called "aliasing" (and is familiar through the stroboscope effect, as seen for example, in movies of turning wagon wheels).

5. Aliasing

Aliasing is an elementary result, and it is pervasive in science. Those who do not understand it are condemned—as one can see in the literature—to sometimes foolish results (Wunsch, 2000). If one understands it, its presence can be benign. Consider for example, the TOPEX/POSEIDON satellite altimeter (e.g., Wunsch and Stammer, 1998), which samples a fixed position on the earth with a return period (Δt) of 9.916 days=237.98 hours (h). The principle lunar semi-diurnal tide (denoted M_2) has a period of 12.42 hours. The spacecraft thus aliases the tide into a frequency (from 4.9)

$$|s_a| = \left| \frac{1}{12.42\text{h}} - \frac{n}{237.98\text{h}} \right| < \frac{1}{2 \times 237.98\text{h}}. \tag{5.1}$$

To satisfy the inequality, one must choose $n = 19$, producing an alias frequency near $s_a = 1/61.6$days, which is clearly observed in the data. (The TOPEX/POSEIDON orbit was very carefully designed to avoid aliasing significant tidal lines (there are about 40 different frequencies to be concerned about) into geophysically important frequencies such as those corresponding to the mean (0 frequency), and the annual cycle (see Parke, et al., 1987)).

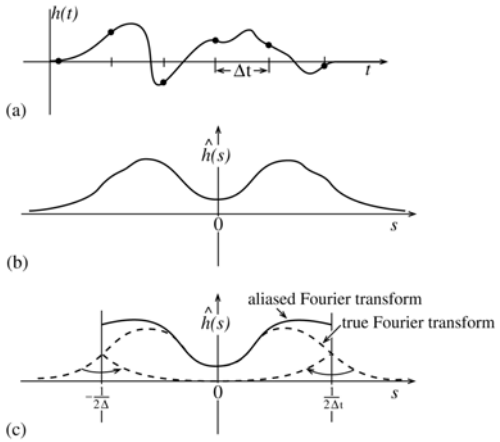

FIGURE 7. A function (a) with Fourier transform as in (b) is sampled as shown at intervals Δt, producing a corrupted (aliased) Fourier transform as shown in (c). Modified after Press et al. (1992)

Exercise. Compute the alias period of the principal solar semidiurnal tide of period 12.00 hours as sampled by TOPEX/POSEIDON, and for both lunar and solar semidiurnal tides when sampled by an altimeter in an orbit which returns to the same position every 35.00 days.

Exercise. The frequency of the so-called tropical year (based on passage of the sun through the vernal equinox) is $s_t = 1/365.244$d. Suppose a temperature record is sampled at intervals $\Delta t = 365.25$d. What is the apparent period of the tropical year signal? Suppose it is sampled at $\Delta t = 365.00$d (the "common year"). What then is the apparent period? What conclusion do you draw?

Pure sinusoids are comparatively easily to deal with if aliased, as long as one knows their origin. Inadequate sampling of functions with more complex Fourier transforms can be much more pernicious. Consider the function shown in Figure 7a whose Fourier transform is shown in Figure 7b. When sub-sampled as indicated, one obtains the Fourier transform in Fig. 7c. If one was unaware of this effect, the result can be devastating for the interpretation. *Once the aliasing has occurred, there is no way to undo it.* Aliasing is inexorable and unforgiving; we will see it again when we study stochastic functions.

6. Discrete Fourier Analysis

The expression (4.2) shows that the Fourier transform of a discrete process is a function of $\exp(-2\pi i s)$ and hence is periodic with period 1 (or $1/\Delta t$ for general Δt). A finite data length means that all of the information about it is contained in its values at the special frequencies $s_n = n/T$. If we define

$$z = e^{-2\pi i s \Delta t} \tag{6.1}$$

the Fourier transform is

$$\hat{x}(s) = \sum_{m=-T/2}^{T/2} x_m z^m \tag{6.2}$$

We will write this, somewhat inconsistently interchangeably, as $\hat{x}(s), \hat{x}\left(e^{-2\pi i s}\right), \hat{x}(z)$ where the two latter functions are identical; $\hat{x}(s)$ is clearly not the same function as $\hat{x}\left(e^{-2\pi i s}\right)$, but the context should make clear what is intended. Notice that $\hat{x}(z)$ is just a polynomial in z, with negative powers of z multiplying x_n at negative times. That a Fourier transform (or series–which differs only by a constant normalization) is a polynomial in $\exp(-2\pi i s)$ proves to be a simple, but powerful idea.

Definition. We will use interchangeably the terminology "sequence", "series" and "time series", for the discrete function x_m, whether it is discrete by definition, or is a sampled continuous function. Any subscript implies a discrete value.

Consider for example, what happens if we multiply the Fourier transforms of x_m, y_m :

$$\hat{x}(z)\,\hat{y}(z) = \left(\sum_{m=-T/2}^{T/2} x_m z^m\right)\left(\sum_{k=-T/2}^{T/2} y_k z^k\right) = \sum_k \left(\sum_m x_m y_{k-m}\right) z^k = \hat{h}(z). \tag{6.3}$$

That is to say, the product of the two Fourier transforms is the Fourier transform of a new time series,

$$h_k = \sum_{m=-\infty}^{\infty} x_m y_{k-m}, \tag{6.4}$$

which is the rule for polynomial multiplication, and is a discrete generalization of convolution. The infinite limits are a convenience—most often one or both time series vanishes beyond a finite value.

More generally, the algebra of discrete Fourier transforms is the algebra of polynomials. We could ignore the idea of a Fourier transform altogether and simply define a transform which associates any sequence $\{x_m\}$ with the corresponding polynomial (6.2), or formally

$$\{x_m\} \longleftrightarrow \mathcal{Z}(x_m) \equiv \sum_{m=-T/2}^{T/2} x_m z^m \tag{6.5}$$

The operation of transforming a discrete sequence into a polynomial is called a $z-$transform. The $z-$transform coincides with the Fourier transform on the unit circle $|z| = 1$. If we regard z as a general complex variate, as the symbol is meant to suggest, we have at our disposal the entire subject of complex functions to manipulate the Fourier transforms, as long as the corresponding functions are finite on the unit circle. Fig. 8 shows how the complex $s-$plane, maps into the complex $z-$plane, the real line in the former, mapping into the unit circle, with the upper half-$s-$plane becoming the interior of $|z| = 1$

There are many powerful results. One simple type is that any function analytic on the unit circle corresponds to a Fourier transform of a sequence. For example, suppose

$$\hat{x}(z) = Ae^{az} \tag{6.6}$$

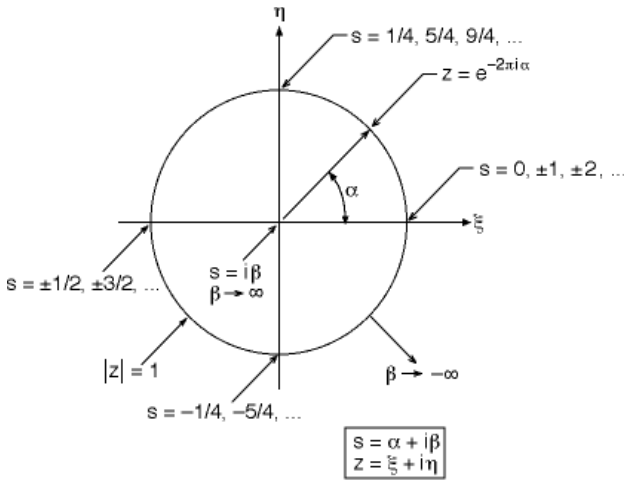

FIGURE 8. Relationships between the complex z and s planes.

Because $\exp(az)$ is analytic everywhere for $|z| < \infty$, it has a convergent Taylor Series on the unit circle

$$\hat{x}(z) = A\left(1 + az + a^2\frac{z^2}{2!} + ...\right) \tag{6.7}$$

and hence $x_0 = A, x_1 = Aa, x_2 = Aa^2/2!, ...$Note that $x_m = 0, m < 0$. Such a sequence, vanishing for negative m, is known as a "causal" one.

Exercise. Of what sequence is $A\sin(bz)$ the $z-$transform? What is the Fourier Series? How about,

$$\hat{x}(z) = \frac{1}{(1-az)(1-bz)}, a > 1, b < 1? \tag{6.8}$$

This formalism permits us to define a "convolution inverse". That is, given a sequence, x_m, can we find a sequence, b_m, such that the discrete convolution

$$\sum_k b_k x_{m-k} = \sum_k b_{m-k} x_k = \delta_{m0} \tag{6.9}$$

where δ_{m0} is the Kronecker delta (the discrete analogue of the Dirac δ)? To find b_m, take the $z-$transform of both sides, noting that $\mathcal{Z}(\delta_{m0}) = 1$, and we have

$$\hat{b}(z)\,\hat{x}(z) = 1 \tag{6.10}$$

or

$$\hat{b}(z) = \frac{1}{\hat{x}(z)} \tag{6.11}$$

But since $\hat{x}(z)$ is just a polynomial, we can find $\hat{b}(z)$ by simple polynomial division.

Example. Let $x_m = 0, m < 0, x_0 = 1, x_1 = 1/2, x_2 = 1/4, x_m = 1/8, ...$ What is its convolution inverse? $\mathcal{Z}(x_m) = 1 + z/2 + z^2/4 + z^3/8 + ...$ So

$$\hat{x}(z) = 1 + z/2 + z^2/4 + z^3/8 + ... = \frac{1}{1 - (1/2)z} \tag{6.12}$$

so $\hat{b}(z) = 1 - (1/2)z$, with $b_0 = 1, b_1 = -1/2, b_m = 0$, otherwise.

Exercise. Confirm by direct convolution that the above b_m is indeed the convolution inverse of x_m.

This idea leads to the extremely important field of "deconvolution". Define

$$h_m = \sum_{n=-\infty}^{\infty} f_n g_{m-n} = \sum_{n=-\infty}^{\infty} g_n f_{m-n}. \tag{6.13}$$

Define $g_m = 0, m < 0$; that is, g_m is causal. Then the second equality in (6.13) is

$$h_m = \sum_{n=0}^{\infty} g_n f_{m-n}, \tag{6.14}$$

or writing it out,

$$h_m = g_0 f_m + g_1 f_{m-1} + g_2 f_{m-2} + ... \tag{6.15}$$

If time $t = m$ is regarded as the "present", then g_n operates only upon the present and earlier (the past) values of f_k; no future values of f_m are required. Causal sequences g_n appear, e.g., when one passes a signal, f_k through a linear system which does not respond to the input before it occurs, that is the system is causal. Indeed, the notation g_n has been used to suggest a Green function. So-called real time filters are always of this form; they cannot operate on observations which have not yet occurred.

In general, whether a $z-$transform requires positive, or negative powers of z (or both) depends only upon the location of the singularities of the function relative to the unit circle. If there are singularities in $|z| < 1$, a Laurent series is required for convergence on $|z| = 1$; if all of the singularities occur for $|z| > 1$, a Taylor Series will be convergent and the function will be causal. If both types of singularities are present, a Taylor-Laurent Series is required and the sequence cannot be causal. When singularities exist on the unit circle itself, as with Fourier transforms with singularities on the real s-axis one must decide through a limiting process what the physics are.

Consider the problem of deconvolving h_m in (6.15) from a knowledge of g_n and h_m, that is one seeks f_k. From the convolution theorem,

$$\hat{f}(z) = \frac{\hat{h}(z)}{\hat{g}(z)} = \hat{h}(z)\,\hat{a}(z). \tag{6.16}$$

Could one find f_k given only the past and present values of h_m? Evidently, that requires a filter $\hat{a}(z)$ which is also causal. Thus it cannot have any poles inside $|z| < 1$. The poles of $\hat{a}(z)$ are evidently the zeros of $\hat{g}(z)$ and so the latter cannot have any zeros inside $|z| < 1$. Because $\hat{g}(z)$ is itself causal, if it is to have a (stable) causal inverse, it cannot have either poles or zeros inside the unit circle. Such a sequence g_m is called "minimum phase" and has a number of interesting and useful properties (see e.g., Claerbout, 1985).

As one example, consider that it is possible to show that for *any* stationary, stochastic process, x_n, that one can always write it as

$$x_n = \sum_{k=0}^{\infty} a_n \theta_{n-k},\ a_0 = 1$$

where a_n is minimum phase and θ_n is white noise, with $< \theta_n^2 >= \sigma_\theta^2$. Let n be the present time. Then one time-step in the future, one has

$$x_{n+1} = \theta_{n+1} + \sum_{k=1}^{\infty} a_n \theta_{n-k}.$$

Now at time n, θ_{n+1} is completely unpredictable. Thus the best possible prediction is

$$\tilde{x}_{n+1} = 0 + \sum_{k=1}^{\infty} a_n \theta_{n-k}. \tag{6.17}$$

with expected error,

$$< \left(\tilde{x}_{n+1} - x_{n+1}\right)^2 >=< \theta_n^2 >= \sigma_\theta^2.$$

It is possible to show that this prediction, given a_n, is the best possible one; no other predictor can have a smaller error than that given by the minimum phase operator a_n. If one wishes a prediction q steps into the future, then it follows immediately that

$$\tilde{x}_{n+q} = \sum_{k=q}^{\infty} a_k \theta_{n+q-k},$$

$$< \left(\tilde{x}_{n+q} - x_{n+q}\right)^2 >= \sigma_\theta^2 \sum_{k=0}^{q} a_k^2$$

which sensibly, has a monotonic growth with q. Notice that θ_k is determinable from x_n and its *past* values only, as the minimum phase property of a_k guarantees the existence of the convolution inverse filter, b_k, such that,

$$\theta_n = \sum_{k=0}^{\infty} b_k x_{n-k},\ b_0 = 1.$$

Exercise. Consider a $z-$transform

$$\hat{h}\left(z\right) = \frac{1}{1 - az} \tag{6.18}$$

and find the corresponding sequence h_m when $a \to 1$ from above, and from below.

It is helpful, sometimes, to have an inverse transform operation which is more formal than saying "read off the corresponding coefficient of z^m). The inverse operator $\mathcal{Z}^{-1}$ is just the Cauchy Residue Theorem

$$x_m = \frac{1}{2\pi i} \oint_{|z|=1} \frac{\hat{x}\left(z\right)}{z^{m+1}} dz. \tag{6.19}$$

We leave all of the details to the textbooks (see especially, Claerbout, 1985).

The discrete analogue of cross-correlation involves two sequences x_m, y_m in the form

$$r_\tau = \sum_{n=-\infty}^{\infty} x_n y_{n+\tau} \tag{6.20}$$

which is readily shown to be the convolution of y_m with the *time-reverse* of x_n. Thus by the discrete time-reversal theorem,

$$\mathcal{F}(r_\tau) = \hat{r}(s) = \hat{x}(s)^* \hat{y}(s). \tag{6.21}$$

Equivalently,

$$\hat{r}(z) = \hat{x}\left(\frac{1}{z}\right)\hat{y}(z). \tag{6.22}$$

$\hat{r}\left(z = e^{-2\pi i s}\right) = \Phi_{xy}(s)$ is known as the cross-power spectrum of x_n, y_n.

If $x_n = y_n$, we have discrete autocorrelation, and

$$\hat{r}(z) = \hat{x}\left(\frac{1}{z}\right)\hat{x}(z). \tag{6.23}$$

Notice that wherever $\hat{x}(z)$ has poles and zeros, $\hat{x}(1/z)$ will have corresponding zeros and poles. $\hat{r}\left(z = e^{-2\pi i s}\right) = \Phi_{xx}(s)$ is known as the power spectrum of x_n. Given any $\hat{r}(z)$, the so-called spectral factorization problem consists of finding two factors $\hat{x}(z)$, $\hat{x}(1/z)$ the first of which has all poles and zeros outside $|z| = 1$, and the second having the corresponding zeros and poles inside. The corresponding x_m would be minimum phase.

Example. Let $x_0 = 1, x_1 = 1/2, x_n = 0, n \neq 0, 1$. Then $\hat{x}(z) = 1 + z/2$, $\hat{x}(1/z) = 1 + 1/(2z)$, $\hat{r}(z) = (1 + 1/(2z))(1 + z/2) = 5/4 + 1/2(z + 1/z)$. Hence $\Phi_{xx}(s) = 5/4 + \cos(2\pi s)$.

Convolution as a Matrix Operation

Suppose f_n, g_n are both causal sequences. Then their convolution is

$$h_m = \sum_{n=0}^{\infty} g_n f_{m-n} \tag{6.24}$$

or writing it out,

$$\begin{aligned} h_0 &= f_0 g_0 \\ h_1 &= f_0 g_1 + f_1 g_0 \\ h_2 &= f_0 g_2 + f_1 g_1 + f_2 g_0 \\ &\cdots \end{aligned} \tag{6.25}$$

which we can write in vector matrix form as

$$\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ . \end{bmatrix} = \begin{Bmatrix} g_0 & 0 & 0 & 0 & . & 0 \\ g_1 & g_0 & 0 & 0 & . & 0 \\ g_2 & g_1 & g_0 & 0 & . & 0 \\ . & . & . & . & . & . \end{Bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ . \end{bmatrix},$$

or because convolution commutes, alternatively as

$$\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ . \end{bmatrix} = \begin{Bmatrix} f_0 & 0 & 0 & 0 & . & 0 \\ f_1 & f_0 & 0 & 0 & . & 0 \\ f_2 & f_1 & f_0 & 0 & . & 0 \\ . & . & . & . & . & . \end{Bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ . \end{bmatrix},$$

which can be written compactly as

$$\mathbf{h} = \mathbf{G}\mathbf{f} = \mathbf{F}\mathbf{g}$$

where the notation should be obvious. Deconvolution then becomes, e.g.,

$$\mathbf{f} = \mathbf{G}^{-1}\mathbf{h},$$

if the matrix inverse exists. These forms allow one to connect convolution, deconvolution and signal processing generally to the matrix/vector tools discussed, e.g., in Wunsch (1996). Notice that causality was not actually required to write convolution as a matrix operation; it was merely convenient.

Starting in Discrete Space

One need not begin the discussion of Fourier transforms in continuous time, but can begin directly with a discrete time series. Note that some processes are by nature discrete (e.g., population of a list of cities; stock market prices at closing-time each day) and there need not be an underlying continuous process. But whether the process is discrete, or has been discretized, the resulting Fourier transform is then periodic in frequency space. If the duration of the record is finite (and it could be physically of limited lifetime, not just bounded by the observation duration; for example, the width of the Atlantic Ocean is finite and limits the wavenumber resolution of any analysis), then the resulting Fourier transform need be computed only at a finite, countable number of points. Because the Fourier sines and cosines (or complex exponentials) have the somewhat remarkable property of being exactly orthogonal not only when integrated over the record length, but also of being exactly orthogonal when *summed* over the same interval, one can do the entire analysis in discrete form.

Following the clear discussion in Hamming (1973, p. 510), let us for variety work with the real sines and cosines. The development is slightly simpler if the number of data points, N, is even, and we confine the discussion to that (if the number of data points is in practice odd, one can modify what follows, or simply add a zero data point, or drop the last data point, to reduce to the even number case). Define $T = N\Delta t$ (notice that the basic time duration is not $(N-1)\,\Delta t$ which is the true data duration, but has

one extra time step. Then the sines and cosines have the following orthogonality properties:

$$\sum_{p=0}^{N-1} \cos\left(\frac{2\pi k}{T} p\Delta t\right) \cos\left(\frac{2\pi m}{T} p\Delta t\right) = \begin{cases} 0 & k \neq m \\ N/2 \ , & k = m \neq 0, N/2 \\ N & k = m = 0, N/2 \end{cases} \tag{6.26}$$

$$\sum_{p=0}^{N-1} \sin\left(\frac{2\pi k}{T} p\Delta t\right) \sin\left(\frac{2\pi m}{T} p\Delta t\right) = \begin{cases} 0 & k \neq m \\ N/2 \ , & k = m \neq 0, N/2 \end{cases} \tag{6.27}$$

$$\sum_{p=0}^{N-1} \cos\left(\frac{2\pi k}{T} p\Delta t\right) \sin\left(\frac{2\pi m}{T} p\Delta t\right) = 0. \tag{6.28}$$

Zero frequency, and the Nyquist frequency, are evidently special cases. Using these orthogonality properties the expansion of an arbitrary sequence at data points $m\Delta t$ proves to be:

$$x_m = \frac{a_0}{2} + \sum_{k=1}^{N/2-1} a_k \cos\left(\frac{2\pi k m\Delta t}{T}\right) + \sum_{k=1}^{N/2-1} b_k \sin\left(\frac{2\pi k m\Delta t}{T}\right) + \frac{a_{N/2}}{2} \cos\left(\frac{2\pi N m\Delta t}{2T}\right), \tag{6.29}$$

where

$$a_k = \frac{2}{N} \sum_{p=0}^{N-1} x_p \cos\left(\frac{2\pi k p\Delta t}{T}\right), k = 0, ..., N/2 \tag{6.30}$$

$$b_k = \frac{2}{N} \sum_{p=0}^{N-1} x_p \sin\left(\frac{2\pi k p\Delta t}{T}\right), k = 1, ...N/2 - 1. \tag{6.31}$$

The expression (6.29) separates the 0 and Nyquist frequencies and whose sine coefficients always vanish; often for notational simplicity, we will assume that a_0, a_N vanish (removal of the mean from a time series is almost always the first step in any case, and if there is significant amplitude at the Nyquist frequency, one probably has significant aliasing going on.). Notice that as expected, it requires $N/2 + 1$ values of a_k and $N/2 - 1$ values of b_k for a total of N numbers in the frequency domain, the same total numbers as in the time-domain.

Exercise. Write a computer code to implement (6.30,6.31) directly. Show numerically that you can recover an arbitrary sequence x_p.

The complex form of the Fourier *series,* would be

$$x_m = \sum_{k=-N/2}^{N/2} \alpha_k e^{2\pi i k m\Delta t/T} \tag{6.32}$$

$$\alpha_k = \frac{1}{N} \sum_{p=0}^{N-1} x_p e^{-2\pi i k p\Delta t/T}. \tag{6.33}$$

This form follows from multiplying (6.32) by $\exp(-2\pi i m r\Delta t/T)$, summing over m and noting

$$\sum_{m=0}^{N-1} e^{(k-r)2\pi i m\Delta t/T} = \begin{cases} N, & k = r \\ \left(1 - e^{(2\pi i(k-r))}\right) / \left(1 - e^{(2\pi i(k-r)/N)}\right) = 0, & k \neq r \end{cases}. \tag{6.34}$$

The last expression follows from the finite summation formula for geometric series,

$$\sum_{j=0}^{N-1} ar^j = a\frac{1-r^N}{1-r}. \tag{6.35}$$

The Parseval Relationship becomes

$$\frac{1}{N}\sum_{m=0}^{N-1} x_m^2 = \sum_{k=-N/2}^{N/2} |\alpha_k|^2 . \tag{6.36}$$

The number of complex coefficients α_k appears to involve $2(N/2)+1 = N+1$ complex numbers, or $2N+2$ values, while the x_m are only N real numbers. But it follows immediately that if x_m are real, that $\alpha_{-k} = \alpha_k^*$, so that there is no new information in the negative index values, and $\alpha_0, \alpha_{N/2} = \alpha_{-N/2}$ are both real so that the number of Fourier series values is identical. Note that the Fourier transform values,$\hat{x}(s_n)$ at the special frequencies $s_n = 2\pi n/T$, are

$$\hat{x}(s_n) = N\alpha_n, \tag{6.37}$$

so that the Parseval relationship is modified to

$$\sum_{m=0}^{N-1} x_m^2 = \frac{1}{N}\sum_{k=-N/2}^{N/2} |\hat{x}(s_n)|^2 . \tag{6.38}$$

To avoid negative indexing issues, many software packages redefine the baseband to lie in the positive range $0 \le k \le N$ with the negative frequencies appearing after the positive frequencies (see, e.g., Press et al., 1992, p. 497). Supposing that we do this, the complex Fourier transform can be written in vector/matrix form. Let $z_n = e^{-2\pi i s_n t}$, then

$$\begin{bmatrix} \hat{x}(s_0) \\ \hat{x}(s_1) \\ \hat{x}(s_2) \\ . \\ \hat{x}(s_m) \\ . \end{bmatrix} = \begin{Bmatrix} 1 & 1 & 1 & . & 1 & . \\ 1 & z_1^1 & z_1^2 & . & z_1^N & . \\ 1 & z_2^1 & z_2^2 & . & z_2^N & . \\ . & . & . & . & . & . \\ 1 & z_m^1 & z_m^2 & . & z_m^N & . \\ . & . & . & . & . & . \end{Bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ . \\ x_q \\ . \end{bmatrix}. \tag{6.39}$$

or,

$$\hat{\mathbf{x}} = \mathbf{B}\mathbf{x}, \tag{6.40}$$

The inverse transform is thus just

$$\mathbf{x} = \mathbf{B}^{-1}\hat{\mathbf{x}}, \tag{6.41}$$

and the entire numerical operation can be thought of as a set of simultaneous equations, e.g., (6.41), for a set of unknowns $\hat{\mathbf{x}}$.

The relationship between the complex and the real forms of the Fourier series is found simply. Let $\alpha_n = c_n + id_n$, then for real x_m, (6.32) is,

$$\begin{aligned} x_m &= \sum_{n=0}^{N/2} (c_n + id_n)(\cos(2\pi nm/T) + i\sin(2\pi nm/T)) + (c_n - id_n)(\cos(2\pi nm/T) - i\sin(2\pi nm/T)) \\ &= \sum_{n=0}^{N/2} \{2c_n \cos(2\pi nm/T) - 2d_n \sin(2\pi nm/T)\}, \end{aligned} \tag{6.42}$$

so that,

$$a_n = 2\,\mathrm{Re}(\alpha_n),\ b_n = -2\,\mathrm{Im}(\alpha_n) \tag{6.43}$$

and when convenient, we can simply switch from one representation to the other.

Software that shifts the frequencies around has to be used carefully, as one typically rearranges the result to be physically meaningful (e.g., by placing negative frequency values in a list preceding positive frequency values with zero frequency in the center). If an inverse transform is to be implemented, one must shift back again to whatever convention the software expects. Modern software computes Fourier transforms by a so-called fast Fourier transform (FFT) algorithm, and not by the straightforward calculation of (6.30, 6.31). Various versions of the FFT exist, but they all take account of the idea that many of the operations in these coefficient calculations are done repeatedly, if the number, N of data points is not prime. I leave the discussion of the FFT to the references (see also, Press et al., 1992), and will only say that one should avoid prime N, and that for very large values of N, one must be concerned about round-off errors propagating through the calculation.

Exercise. Consider a time series $x_m, -T/2 \le m \le T/2$, sampled at intervals Δt. It is desired to interpolate to intervals $\Delta t/q$, where q is a positive integer greater than 1. Show (numerically) that an extremely fast method for doing so is to find $\hat{x}(s), |s| \le 1/2\Delta t$, using an FFT, to extend the baseband with zeros to the new interval $|s| \le q/(2\Delta t)$, and to inverse Fourier transform back into the time domain. (This is called "Fourier interpolation" and is very useful.).

7. Identities and Difference Equations

$Z-$transform analogues exist for all of the theorems of ordinary Fourier transforms.

Exercise. Demonstrate:

The shift theorem: $\mathcal{Z}(x_{m-q}) = z^q \hat{x}(z)$.

The differentiation theorem: $\mathcal{Z}(x_m - x_{m-1}) = (1-z)\hat{x}(z)$. Discuss the influence of a difference operation like this has on the frequency content of $\hat{x}(s)$.

The time-reversal theorem: $\mathcal{Z}(x_{-m}) = \hat{x}(1/z)$.

These and related relationships render it simple to solve many difference equations. Consider the difference equation

$$x_{m+1} - ax_m + bx_{m-1} = p_m \tag{7.1}$$

where p_m is a known sequence and a, b are constant. To solve (7.1), take the $z-$transform of both sides, using the shift theorem:

$$\frac{1}{z}\hat{x}(z) - a\hat{x}(z) + bz\hat{x}(z) = \hat{p}(z) \tag{7.2}$$

and solving,

$$\hat{x}_p(z) = \frac{\hat{p}(z)}{(1/z - a + bz)}. \tag{7.3}$$

If $p_m = 0, m < 0$ (making p_m causal), then the solution (7.3) is both causal and stable only if the zeros of $(1/z - a + z)$ lie outside $|z| = 1$.

Exercise. Find the sequence corresponding to (7.3).

Eq. (7.3) is the particular solution to the difference equation. A second order difference equation in general requires two boundary or initial conditions. Suppose x_0, x_1 are given. Then in general we need a homogeneous solution to add to (7.3) to satisfy the two conditions. To find a homogeneous solution, take $\hat{x}_h(z) = Ac^m$ where A, c are constants. The requirement that $\hat{x}_h(z)$ be a solution to the homogeneous difference equation is evidently $c^{m+1} - ac^m + bc^{m-1} = 0$ or, $c - a + bc^{-1} = 0$, which has two roots, $c_\pm$. Thus the general solution is

$$x_m = \mathcal{Z}^{-1}\left(\hat{x}_p(z)\right) + A_+ c_+^m + A_- c_-^m \tag{7.4}$$

where the two constants $A_\pm$ are available to satisfy the two initial conditions. Notice that the roots $c_\pm$ determine also the character of (7.3). This is a large subject, left at this point to the references.[3]

We should note that Box, Jenkins and Reisel (1994) solve similar equations without using $z-$transforms. They instead define forward and backwards difference operators, e.g., $\mathcal{B}(x_m) = x_{m-1}, \mathcal{F}(x_m) = x_{m+1}$. It is readily shown that these operators obey the same algebraic rules as do the $z-$transform, and hence the two approaches are equivalent.

Exercise. Evaluate $(1 - \alpha\mathcal{B})^{-1} x_m$ with $|\alpha| < 1$.

8. Circular Convolution

There is one potentially puzzling feature of convolution for discrete sequences. Suppose one has $f_m \neq 0, = 0, 1, 2$, and is zero otherwise, and that $g_m \neq 0, m = 0, 1, 2$, and is zero otherwise. Then $h = f * g$ is,

$$[h_0, h_1, h_2, h_3, h_4, h_5] = [f_0 g_0, f_0 g_1 + f_1 g_0, f_0 g_2 + f_1 g_1 + f_2 g_0, f_1 g_2 + f_2 g_0, f_0 g_2], \tag{8.1}$$

that is, is non-zero for 5 elements. But the product $\hat{f}(z)\,\hat{g}(z)$ is the Fourier transform of only a 3-term non-zero sequence. How can the two results be consistent? Note that $\hat{f}(z), \hat{g}(z)$ are Fourier transforms of two sequences which are numerically indistinguishable from periodic ones with period 2. Thus their product must also be a Fourier transform of a sequence indistinguishable from periodic with period 2.

[3]The procedure of finding a particular and a homogeneous solution to the difference equation is wholly analogous to the treatment of differential equations with constant coefficients.

$\hat{f}(z)\,\hat{g}(z)$ is the Fourier transform of the convolution of two periodic sequences f_m, g_m, not the ones in Eq. (8.1) that we have treated as being zero outside their region of definition. $\mathcal{Z}^{-1}\left(\hat{f}(z)\,\hat{g}(z)\right)$ is the convolutionof two periodic sequences, and which have "wrapped around" on each other—giving rise to their description as "circular convolution". To render circular convolution identical to Eq. (8.1), one should pad f_m, g_m with enough zeros that their lengths are identical to that of h_m before forming $\hat{f}(z)\,\hat{g}(z)$.

In a typical situation however, f_m might be a simple filter, perhaps of length 10, and g_m might be a set of observations, perhaps of length 10,000. If one simply drops the five points on each end for which the convolution overlaps the zeros "off-the-ends", then the two results are virtually identical. An extended discussion of this problem can be found in Press et al. (1992, Section 12.4).

9. Fourier Series as Least-Squares

Discrete Fourier series (6.29) are an exact representation of the sampled function if the number of basis functions (sines and cosines) are taken to equal the number, N, of data points. Suppose we use a number of terms $N' \leq N$, and seek a least-squares fit. That is, we would like to minimize

$$J = \sum_{t=0}^{T-1} \left(x_t - \sum_{m=1}^{[N'/2]} \alpha_m e^{2\pi i m t/T} \right) \left(x_t - \sum_{m=1}^{[N'/2]} \alpha_m e^{2\pi i m t/T} \right)^* . \tag{9.1}$$

Taking the partial derivatives of J with respect to the a_m and setting to zero (generating the least-squares normal equations), and invoking the orthogonality of the complex exponentials, one finds that (1) the governing equations are perfectly diagonal and, (2) the a_m are given by precisely (6.29, 6.30). Thus we can draw an important conclusion: *a Fourier series, whether partial or complete, represents a least-squares fit of the sines and cosines to a time series.* Least-squares is discussed at length in Wunsch (1996).

Exercise. Find the normal equations corresponding to (9.1) and show that the coefficient matrix is diagonal.

Non-Uniform Sampling

This result (9.1) shows us one way to handle a non-uniformly spaced time series. Let $x(t)$ be sampled at arbitrary times t_j. We can write

$$x(t_j) = \sum_{m=1}^{[N'/2]} \alpha_m e^{2\pi i m t_j/T} + \varepsilon_j \tag{9.2}$$

where ε_j represents an error to be minimized as

$$J = \sum_{j=0}^{j_N-1} \varepsilon_j^2 = \sum_{j=0}^{j_N-1} \left(x(t_j) - \sum_{m=1}^{[N'/2]} \alpha_n e^{2\pi i n t_j/T} \right) \left(x(t_j) - \sum_{m=1}^{[N'/2]} \alpha_m e^{2\pi i m t_j/T} \right)^* , \tag{9.3}$$

or the equivalent real form, and the normal equations found. The resulting coefficient matrix is no longer diagonal, and one must solve the normal equations by Gaussian elimination or other algorithm. If the

record length and/or N' is not too large, this is a very effective procedure. For long records, the computation can become arduous. Fortunately, there exists a fast solution method for the normal equations, generally called the Lomb-Scargle algorithm (discussed, e.g., by Press et al., 1992; an application to intermittent satellite sampling of the earth's surface can be seen in Wunsch (1991)). The complexities of the algorithm should not however, mask the underlying idea, which is just least-squares.

Exercise. Generate a uniformly spaced time series, x_t, by specifying the coefficients of its Fourier series. Using the Fourier series, interpolate x_t to generate an irregularly spaced set of values of $x(t_j)$. Using $x(t_j)$ and the normal equations derived from (9.3), determine the Fourier components. Discuss their relationship to the known ones. Study what happens if the $x(t_j)$ are corrupted by the addition of white noise. What happens if the observation times t_j are corrupted by a white noise error? ("White noise" is defined below. For present purposes, it can be understood as the output of an ordinary pseudo-random number generator on your computer.)

Irregular Sampling Theorems

There are some interesting theoretical tools available for the discussion of infinite numbers of *irregularly* spaced perfect samples of band-limited functions. A good reference is Freeman (1965), who gives explicit expressions for reconstruction of band-limited functions from arbitrarily spaced data. Among the useful results are that any regularly spaced sample which is missing, can be replaced by an arbitrary irregularly spaced sample anywhere in the record. The limiting case of the expressions Freeman shows, would suggest that one could take *all* of the samples and squeeze them into an arbitrarily brief time interval. This inference would suggest a strong connection between band-limited functions and analytic functions describable by their Taylor Series (regarding closely spaced samples as being equivalent to first, second, etc. differences). Related results permit one to replace half the samples by samples of the derivatives of the function, etc. The reader is cautioned that these results apply to infinite numbers of perfect samples and their use with finite numbers of inaccurate data has to be examined carefully.

Some useful results about "jittered" sampling can be seen in Moore and Thomson (1991), and Thomson and Robinson (1996); an application to an ice core record is Wunsch (2000).

10. Stochastic Processes

Basic Probability

Probability theory is a large subject that has developed deep philosophical roots and corresponding firmly held differences of opinion about its interpretation. Some of these disputes are very important, but for our purposes we will think of a probability density as a frequency function (histogram) in the limit of an infinite number of trials of an experiment as the bin-sizes go to zero. We assume that such a limit exists, and can be written for a random variable y, as $p_y(Y)$, where the subscript is the variable, and the argument Y is the *values* that y can take on (distinguishing the physical variable, e.g. a temperature, from its numerical value). Sometimes the subscript is dropped when the meaning is otherwise clear.

The use of $p_y(Y)$ in practice means that one is referring to the probability that y lies in the interval $Y \leq y \leq Y + dY$ as $p_y(Y)\,dY$, that is, lying in some differential interval.

Two very useful probability densities are the uniform one over interval L

$$\begin{aligned} p_y(Y) &= \frac{1}{L}, -\frac{L}{2} \leq Y \leq \frac{L}{2} \\ &= \Pi(Y/L) \end{aligned} \tag{10.1}$$

and the normal (or $\sum$),

$$p_y(Y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y-m)^2/\left(2\sigma^2\right)}. \tag{10.2}$$

The latter is also written $G\left(m, \sigma^2\right)$ to denote a normal (Gaussian) density with mean m, and variance σ^2. Both (10.1,10.2) satisfy the necessary requirements, $p \geq 0$, and

$$\int_{-\infty}^{\infty} p_y(Y)\,dY = 1. \tag{10.3}$$

Define a bracket, $< . >$ as the averaging operator with meaning of the integral over all possible values of the argument times the probability density. Thus,

$$< y >= \int_{-\infty}^{\infty} Y p_y(Y)\,dY = m \tag{10.4}$$

(the mean, or center of mass),

$$< y^2 >= \int_{-\infty}^{\infty} Y^2 p_y(Y)\,dY \tag{10.5}$$

the second moment,

$$< (y - < y >)^2 >= \int_{-\infty}^{\infty} (Y - \langle y \rangle)^2 p_y(Y)\,dY = \sigma_y^2 \tag{10.6}$$

the variance (second moment about the mean), etc. An important simple result is that if a is constant (not random), then

$$\langle ay \rangle = a < y > . \tag{10.7}$$

Let $f(y)$ be any function of random variable, y. It follows from the frequency function definition of the probability density, that

$$< f(y) >= \int_{-\infty}^{\infty} f(Y) p_y(Y)\,dY. \tag{10.8}$$

Eq. (10.7) is evidently a special case. Often it is useful to find the probability density of f from that of y. We suppose that the function is invertible so that $y(f)$ is known. Then the line segment dY is mapped into a line-segment dF, by the rule,

$$dY = \frac{dy(F)}{dF} dF, \tag{10.9}$$

we have immediately,

$$p_g(F)\,dF = p_y(Y(F)) \frac{dy(F)}{dF} dF. \tag{10.10}$$

A special case would be $f = ay + b$, or $y = (f - b)/a$, $dy(F)/dF = 1/a$, and thus

$$p_g(F) = p_y((F - b)/a)/a. \tag{10.11}$$

If this result is applied to (10.2), we have

$$p_g(F) = \frac{1}{\sqrt{2\pi}a\sigma} e^{-(F-b-ma)^2/\left(2a^2\sigma^2\right)} \tag{10.12}$$

which is $G\left(b-ma, a^2\sigma^2\right)$. Note that by choosing $b = m/\sigma, a = 1/\sigma$, that the new probability density is $G(0,1)$—a standard form that is the one usually tabulated.

If the derivative dy/dF should be zero or infinity, it implies that the mapping from y to f, or from f to y is not unique, and some care is needed (the differentials are not uniquely mapped). So consider $G(0,1)$,

$$p_y(Y) = \frac{1}{\sqrt{2\pi}} e^{-Y^2/2} \tag{10.13}$$

and suppose $\xi = y^2$. Then clearly the probability that ξ is less than zero is zero, and both $\pm y$ map onto the same value of ξ indicated by,

$$\frac{dy}{d\xi} = \frac{1}{2y} = \frac{1}{2\sqrt{\xi}} \tag{10.14}$$

becoming infinite at $\xi = 0$. We can deduce therefore that,

$$p_\xi(X) = \begin{cases} \frac{1}{\sqrt{2\pi}\sqrt{X}} e^{-X/2}, & X \geq 0 \\ 0, & X < 0 \end{cases} . \tag{10.15}$$

multiplying by 2 to account for the negative Y contributions, too. Probability density (10.15) is known as "chi-square with one degree-of-freedom" usually written χ_1^2. "Degrees-of-freedom" will be defined later. For future reference, note that if $\left\langle y^2 \right\rangle = \sigma^2$, still with zero mean, then Eq. (10.15) becomes

$$p_\xi(X) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}\sqrt{X}} e^{-X/\left(2\sigma^2\right)}, & X \geq 0 \\ 0, & X < 0 \end{cases} \tag{10.16}$$

It can become confusing to keep track of mapping functions $g(y)$ which are not unique, and a more systematic approach than used to find (10.15) is desirable. Introduce the "probability distribution function",

$$P_y(Y) = \int_{-\infty}^{Y} p_y(Y')\, dY', \tag{10.17}$$

which has the properties, $dP_y/dY \geq 0, P_y(\infty) = 1, dP_y/dY = p_y(Y)$. The interpretation of $P_y(Y)$ is as the probability that y is less than or equal to Y.

Then for the above case, with $\xi = y^2$ and y being Gaussian,

$$P_\xi(X) = \text{probability that } \left\{-\sqrt{X} \leq Y \leq \sqrt{X}\right\} = P_y\left(\sqrt{X}\right) - P_y\left(-\sqrt{X}\right) \tag{10.18}$$

$$= \int_{-\infty}^{\sqrt{X}} \frac{1}{\sqrt{2\pi}} e^{-Y^2/2} dY - \int_{-\infty}^{-\sqrt{X}} \frac{1}{\sqrt{2\pi}} e^{-Y^2/2} dY. \tag{10.19}$$

And,

$$p_\xi(X) = \frac{d}{dX}P_\xi(X) = \frac{d}{dX}\left[\int_{-\infty}^{\sqrt{X}} \frac{1}{\sqrt{2\pi}} e^{-Y^2/2} dY - \int_{-\infty}^{-\sqrt{X}} \frac{1}{\sqrt{2\pi}} e^{-Y^2/2} dY\right] \tag{10.20}$$

$$= \left[\frac{1}{\sqrt{2\pi}\sqrt{X}} e^{-\frac{1}{2}X}\right], \quad X \geq 0, \tag{10.21}$$

identical to (10.15). "Leibniz's rule" for differentiation of a variable upper bound of integration was used. This approach is quite general, as long as expressions such as (10.18) can be constructed.

If there are two or more random variables, $\xi_i, i = 1, 2., ...m$ we can discuss their "joint" or "multivariate probability densities", $p_{\xi_1\xi_2...}(\Xi_1, \Xi_2, ...\Xi_m)$; these are to be thought of as derived from the limits of a counting experiment in which $\xi_1, \xi_2..$ are measured many times, and then binned by the values observed. The limit, as the number of such observations goes to infinity and as the bin size goes to zero, is supposed to exist.

If a joint probability density factors,

$$p_{\xi_1\xi_2...}(\Xi_1, \Xi_2, ...\Xi_m) = p_{\xi_1}(\Xi_1) p_{\xi_2}(\Xi_2) ... p_{\xi_m}(\Xi_m) \tag{10.22}$$

then the ξ_i are said to be independent.

Example. The general, $m-$dimensional joint Gaussian probability density is defined as,

$$p_{\xi_1\xi_2...}(\Xi_1, \Xi_2, ...\Xi_m) = \tag{10.23}$$
$$\frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\xi} - \mathbf{m})^T \mathbf{R}^{-1} (\boldsymbol{\xi} - \mathbf{m})\right).$$

Here $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, ...\xi_m]^T$,

$$\mathbf{R} = \left\{ \begin{array}{cccc} <(\xi_1 - m_1)(\xi_1 - m_1)> & <(\xi_1 - m_1)(\xi_2 - m_2)> & . \ . & <(\xi_1 - m_1)(\xi_m - m_m)> \\ <(\xi_2 - m_2)(\xi_1 - m_1)> & <(\xi_2 - m_2)(\xi_2 - m_2)> & . \ . & <(\xi_2 - m_2)(\xi_m - m_m)> \\ . & . & . \ . & . \\ . & . & . \ . & . \\ <(\xi_m - m_m)(\xi_1 - m_1)> & <(\xi_m - m_m)(\xi_2 - m_2)> & . \ . & <(\xi_m - m_m)(\xi_m - m_m)> \end{array} \right\} \tag{10.24}$$

and $|\mathbf{R}|$ is the determinant. $\mathbf{R}$ can be written in a variety of ways including

$$\mathbf{R} = \left\{ \begin{array}{cccc} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & . \ . & \sigma_1\sigma_m\rho_{1m} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & . \ . & \sigma_2\sigma_m\rho_{2m} \\ . & . & . \ . & . \\ . & . & . \ . & . \\ \sigma_m\sigma_1\rho_{m1} & \sigma_m\sigma_2\rho_{m2} & . \ . & \sigma_m^2 \end{array} \right\} \tag{10.25}$$

where the σ_i^2 are the corresponding variances about the mean, and the $\rho_{ij} \equiv <(\xi_i - m_i)(\xi_j - m_j)> /\sigma_i\sigma_j = \rho_{ji}$ are called correlations (discussed below).

The important special case of two normal variables ("bivariate normal") can be written as

$$p_{\xi_1\xi_2}(\Xi_1,\Xi_2) = \tag{10.26}$$
$$\frac{1}{2\pi\sigma_1\sigma_2\left(1-\rho^2\right)^{1/2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{\left(\Xi_1-m_1\right)^2}{\sigma_1^2}-\frac{2\rho\left(\Xi_1-m_1\right)\left(\Xi_2-m_2\right)}{\sigma_1\sigma_2}+\frac{\left(\Xi_2-m_2\right)^2}{\sigma_2^2}\right)\right\},$$

where ρ is defined as $\rho =< \left(\xi_1-m_1\right)\left(\xi_2-m_2\right) > /\sigma_1\sigma_2$, and taken up immediately below. Other variants of (10.26) are possible.

Exercise. Show that if all $\rho_{ij}=0, i\neq j$, that (10.23) reduces to the product of m-univariate normal distributions, hence showing that uncorrelated normal variates are also independent.

The joint expectations about the means (moments about the means) are, for two variates,

$$<\xi_1'\xi_2'>=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\Xi_1-m_1\right)\left(\Xi_2-m_2\right)p_{\xi_1\xi_2}\left(\Xi_1,\Xi_2\right)d\Xi_1 d\Xi_2. \tag{10.27}$$

We will use ξ_i' to denote the variable with its mean removed. If ξ_1',ξ_2' are independent, it follows that $<\xi_1'\xi_2'>=<\xi_1'><\xi_2'>=0$.

If $<\xi_1'\xi_2'>\neq 0$, (suppressing the prime), then they are said to be "correlated" variables. This implies that a knowledge of one of them provides some predictive capacity for the other. We can use the idea of conditional probability, e.g., the probability that ξ_2 takes on a particular value (or range of values) Ξ_2 given that $\xi_1=\Xi_1$, which we write as

$$p_{\xi_2|\xi_1}\left(\Xi_2|\Xi_1\right). \tag{10.28}$$

The textbooks all show that if ξ_1,ξ_2 are independent $p_{\xi_2|\xi_1}\left(\Xi_2|\Xi_1\right)=p_{\xi_2}(\Xi_2)$, that is, knowledge of the value of ξ_1 then contains no predictive information about ξ_2. More generally, suppose that we try to predict ξ_2 from ξ_1 in the form

$$\xi_2 = a\xi_1+\varepsilon \tag{10.29}$$

where by definition ε is independent of ξ_1. Forming $<\xi_2\xi_1>=a<\xi_1^2>+<\varepsilon\xi_1>=a<\xi_1^2>$, or

$$a=\frac{<\xi_2\xi_1>}{<\xi_1^2>}, \tag{10.30}$$

which would vanish if $<\xi_1\xi_2>=0$. Define the "correlation coefficient"

$$\rho=\frac{<\xi_2\xi_1>}{<\xi_1^2>^{1/2}<\xi_2^2>^{1/2}}. \tag{10.31}$$

It is straightforward to show that $|\rho|\leq 1$. (e.g., Priestley, p. 79) We have,

$$a=\rho\frac{<\xi_2^2>^{1/2}}{<\xi_1^2>^{1/2}}, \tag{10.32}$$

and it follows that the fraction of the variance in ξ_2 which is correlated with (predictable by knowledge of ξ_1) is just,

$$<\xi_2^2>\rho^2 \tag{10.33}$$

(the "correlated power"), and the part of the variance which is not predictable, is

$$< \xi_2^2 > \left(1 - \rho^2\right) \tag{10.34}$$

which is the "uncorrelated power". If ξ_1, ξ_2 are independent, they are uncorrelated; if they are uncorrelated, they need not be independent.

Let there be m random variables $\xi_1, ...\xi_m$, and suppose we define m new variables $\eta_{1...}\eta_m$ which are functions of the original variables. Conservation of area rules lead to the conclusion that if the joint probability density of the ξ_i is $p_{\xi_1\xi_2...\xi_m}\left(\Xi_1, \Xi_2, ...\Xi_m\right)$, then the joint probability density for the η_i is

$$p_{\eta_1\eta_2...}\left(\eta_1, \eta_2, ..., \eta_m\right) = \tag{10.35}$$
$$p_{\xi_1\xi_2...\xi_m}\left(\Xi_1\left(\eta_1, ..., \eta_m\right), \Xi_2\left(\eta_1, ..., \eta_m\right), ..., \Xi_m\left(\eta_1, ..., \eta_m\right)\right) \frac{\partial\left(\Xi_1, \Xi_2, ..., \Xi_m\right)}{\partial\left(\eta_1, ..., \eta_m\right)}$$

where

$$\frac{\partial\left(\Xi_1, \Xi_2, ..., \Xi_m\right)}{\partial\left(\eta_1, ..., \eta_m\right)} \tag{10.36}$$

is the Jacobian of the transformation between the two sets of variables.

In one dimension, the most common and useful transformations are of the form $\eta = a\xi + b$, a, b constant. In most cases, we work with canonical probability densities, such that e.g., the mean is zero, and the variance unity. The linear transformation, with Jacobian a or a^{-1} permits one to use these standard densities for random variables with arbitrary means and variances (see Eq. 10.12). All the ideas concerning correlation, predictability and independence are generalizable to more than two variables, through multiple and partial correlation studies, but these are left to the references.

Example. Let x, y be two uncorrelated (and hence independent) Gaussian random variables of zero mean and variance σ^2. Define $r = \sqrt{x^2 + y^2}$ and $\phi = \tan^{-1}\left(y/x\right)$. We seek the joint probability density for r, ϕ, and their univariate probability densities. The joint probability density function is

$$p_{x,y}\left(X, Y\right) = \frac{1}{2\pi\sigma} \exp\left[-\left(X^2 + Y^2\right)/2\sigma^2\right].$$

The Jacobian of the transformation from (x, y) to (r, ϕ) is just r (Cartesian to polar coordinates). Then the joint probability density of the new variables is

$$p_{r,\phi}\left(R, \Phi\right) = \frac{R}{2\pi\sigma} \exp\left[-R^2/2\sigma^2\right]$$

Integrating out the Φ variable over its entire range, $-\pi \leq \phi \leq \pi$, (Φ doesn't actually appear), we have immediately,

$$p_r\left(R\right) = \frac{R}{\sigma} \exp\left(-R^2/2\sigma^2\right), \tag{10.37}$$

and by inspection, it must be true that

$$p_\phi\left(\Phi\right) = \frac{1}{2\pi}.$$

Thus the phase has a uniform distribution $-\pi \leq \phi \leq \pi$, and the amplitude and phase are uncorrelated with each other. Note that the probability density for r is called the "Rayleigh distribution".

Exercise. Find the mean and variance of a Rayleigh distributed variable.

For time series work, the most important $m-$dimensional probability density is the Gaussian or normal one (10.23). As the textbooks all show, the normal probability density has several important special properties. One of them is that it is completely specified by its mean, $\mathbf{m}$, and covariance matrix, $\mathbf{R} = \left\{ < \left(\xi_i - m_i\right)\left(\xi_j - m_j\right) > \right\}$, and that if any pair is uncorrelated, they are also independent.

Adding Independent Variates. Characteristic Functions

A common problem is to determine the probability density of a sum of variates, e.g., $\xi = \xi_1 + \xi_2 + ...$ Obtaining the probability density is best done by undertaking a seeming digression. Consider an arbitrary random variable y with known probability density $p_y(Y)$, and the function $g(y) = e^{iyt}$. By (10.8), its expected value is

$$\phi_y(t) =< e^{iyt} >= \int_{-\infty}^{\infty} p_y(Y)\, e^{iYt} dY, \tag{10.38}$$

that is, up to the absence of a factor of 2π in the numerator of the exponential, the Fourier transform of $p_y(Y)$. This expected value is clearly a function of t (which should *not* be interpreted as time—its use is merely convention, as is the definition (10.38) without 2π where our own convention would put it), which we will denote $\phi_y(t)$, and which is called the "characteristic function". Now suppose we have two *independent* random variables y, x. Consider the characteristic function for $w = x + y$,

$$\phi_w(t) =< e^{i(x+y)t} >=< e^{ixt} >< e^{iyt} >= \int_{-\infty}^{\infty} e^{iXt} p_x(X)\, dX \int_{-\infty}^{\infty} e^{iYt} p_y(Y)\, dY, \tag{10.39}$$

by independence. So

$$\phi_w(t) = \phi_x(t)\, \phi_y(t), \tag{10.40}$$

that is to say, the product of two Fourier transforms. Because characteristic functions are defined as Fourier transforms, it follows that the inverse transform is,

$$p_w(W) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_w(t)\, e^{-iWt} dt \tag{10.41}$$

with the $1/2\pi$ being necessary because of its absence in the complex exponentials. But by the convolution theorem (or its very slight re-derivation with this changed convention), we must have

$$p_w(W) = \int_{-\infty}^{\infty} p_x(W')\, p_y(W - W')\, dW'. \tag{10.42}$$

The solution generalizes in an obvious way to the sum of an arbitrary number of independent variables. If $p_x = p_y$, we have $\phi_w = \phi_y(t)^2$ or for the sum of n such variables, $\phi_w(t) = \phi_y(t)^n$.

The characteristic function of the $G(0,1)$ Gaussian is readily found to be

$$\phi(t) = e^{-t^2/2}, \tag{10.43}$$

and thus the sum of two $G(0,1)$ variables would have a characteristic function, $e^{-t^2/2} e^{-t^2/2} = e^{-t^2}$, whose inverse transform is found to be

$$p(X) = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-X^2}, \tag{10.44}$$

that is to say another Gaussian of zero mean, but variance 2. It follows immediately from (10.43) that a sum of n - $G(0,1)$ variables is a new $G(0,n)$ variable. If X is instead the square of a $G\left(0,\sigma^2\right)$ variable, then Eq. (10.44) becomes

$$p(X) = \frac{1}{\sqrt{2\pi}\sqrt{2}\sigma^2} e^{-X^2/\sigma^2} \tag{10.45}$$

One use of the characteristic function is in its relations to the moments of the corresponding probability density. Let us suppose $\phi(t)$ has a convergent Taylor series about $t = 0$:

$$\phi(t) = 1 + t \left.\frac{d\phi(t)}{dt}\right|_{t=0} + \frac{t^2}{2!} \left.\frac{d^2\phi(t)}{dt^2}\right|_{t=0} + ... \tag{10.46}$$

But from (10.38) we can evaluate the successive derivatives:

$$\begin{aligned}
\left.\frac{d\phi(t)}{dt}\right|_{t=0} &= \int_{-\infty}^{\infty} iY p_y(Y)\, dY = im, \\
\left.\frac{d^2\phi(t)}{dt^2}\right|_{t=0} &= \int_{-\infty}^{\infty} (iY)^2 p_y(Y)\, dY = i^2 m_2, \\
\left.\frac{d^3\phi(t)}{dt^3}\right|_{t=0} &= \int_{-\infty}^{\infty} (iY)^3 p_y(Y)\, dY = i^3 m_3, \\
\left.\frac{d^{(n)}\phi(t)}{dt^{(n)}}\right|_{t=0} &= \int_{-\infty}^{\infty} (iY)^n p_y(Y)\, dY = i^n m_n,
\end{aligned} \tag{10.47}$$

Where m_i are the successive moments of the probability density. Thus the successive moments determine the terms of the Taylor series expansion in (10.46), and hence the probability density itself. These results can be turned into a statement that a knowledge of all of the moments m_i of a random process is usually equivalent to knowledge of the complete probability density. Conversely, knowledge of the characteristic function means that all of the moments are readily generated from its derivatives evaluated at $t = 0$. (There are probability densities whose moments are not finite, e.g., the Cauchy, and the argument fails for them.)

The characteristic function generalizes to multivariate probability densities by introduction of Fourier transforms in several dimensions. Moment generation is achieved by using Taylor series in several dimensions.

Central Limit Theorems (CLT)

Consider the sum of n *independent* variables, ξ_i,

$$\xi = \xi_1 + ... + \xi_n \tag{10.48}$$

all having the same mean, m_1 and variance, σ_1^2, and with the same, but arbitrary, probability density, $p_1(\Xi)$. Then

$$< \xi >= m = nm_1, < (\xi - m)^2 >= \sigma^2 = n\sigma_1^2. \tag{10.49}$$

Define the normalized variable

$$\tilde{\xi} = \frac{\xi - m}{\sigma} = \tilde{\xi}_1 + ... + \tilde{\xi}_n,\ \tilde{\xi}_i = \frac{\xi_i - m_1}{\sqrt{n}\sigma_1}. \tag{10.50}$$

Suppose that the characteristic function of ξ_i is $\phi_1(t)$. Then by the shift and scaling theorems of Fourier transforms, the characteristic function of $\tilde{\xi}_i$ is $\tilde{\phi}_1(t) = \phi_1(t/(\sigma_1\sqrt{n}))\exp(-im_1 t/(\sigma_1 n))$. Hence the characteristic function for $\tilde{\xi}$ must be

$$\tilde{\phi}(t) = \tilde{\phi}_1(t)^n = \left[e^{-im_1 t/\left(\sigma_1\sqrt{n}\right)}\phi_1\left(t/\left(\sigma_1\sqrt{n}\right)\right)\right]^n. \tag{10.51}$$

Now $\tilde{\phi}_1(t)$ is evidently the characteristic function of a random variable with zero mean and variance $1/n$. Thus it must be true that (expanding in a Taylor Series), and using (10.46, 10.47),

$$e^{-im_1 t/\left(\sigma_1\sqrt{n}\right)}\phi_1\left(t/\left(\sigma_1\sqrt{n}\right)\right) = 1 - \frac{t^2}{2n} + O\left(\frac{t^3}{\sigma_1 n^{3/2}}\right). \tag{10.52}$$

Thus to lowest order, (10.51) is

$$\tilde{\phi}(t) = \left(1 - \frac{t^2}{2n}\right)^n. \tag{10.53}$$

Taking the limit as n goes to infinity, and invoking L'Hôpital's rule on the log of $\tilde{\phi}(t)$, we have

$$\tilde{\phi}(t) \to e^{-t^2/2} \tag{10.54}$$

Thus in this limit, the probability density for $\tilde{\xi}$ is the inverse tranform of (10.54) and is

$$\tilde{p}(\Xi) = \frac{1}{\sqrt{2\pi}}e^{-\Xi^2/2} \tag{10.55}$$

that is, $G(0,1)$. Or, using the scaling and shift theorems,

$$p(\Xi) = \frac{1}{\sqrt{n}\sigma_1\sqrt{2\pi}}e^{-(\Xi - nm_1)^2/\left(2\sqrt{n}\sigma_1\right)}. \tag{10.56}$$

That is to say, we have shown that sums of large numbers of random variates have a tendency to become Gaussian, whether or not the underlying probability densities are themselves Gaussian. Result (10.56) is a special case of the so-called Central Limit Theorem (CLT), which can be proved under much more general circumstances. There are clearly restrictions that could prevent the limit (10.54) from being reached, but the CLT is often valid. Note the special case of a sample average,

$$\tilde{m} = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \tag{10.57}$$

for which it follows immediately that $\tilde{m}$ will have a probability density $G\left(m, \sigma^2/n\right)$.

Exercise. Suppose that $p_1(\Xi) = 1/2\left\{\delta(\Xi - 1) + \delta(\Xi + 1)\right\}$. Study the behavior of (10.51) as $n \to \infty$. What is the limiting probability density of the sum? Hint: use the binomial expansion on $(1/2^n)\left(e^{it} + e^{-it}\right)^n$.

Stationarity

Consider a $G\left(0, \sigma^2\right)$ time series x_m such that $R_{nm} =< x_n x_m >$. Given the zero mean, and specified $\mathbf{R}$, we infer that the joint probability density $p_{x_1 x_2 \ldots x_k}$ is (10.23) with $\mathbf{m} = \mathbf{0}$. Let us suppose that R_{nm} depends only upon the time difference $\tau = n - m$, so that $R_{nm} =< x_n x_{n+\tau} >= R_\tau$, independent of n. Such a time series is said to be "stationary in the wide sense". If it is also true that all statistics, including all higher moments such as $< x_n x_m x_p >$, etc. only depend upon the time interval among the

time series elements, the time series is said to be "stationary" or "stationary in the strict sense". It is another nice property of normal variates that if they are stationary in the wide-sense, they are stationary in the strict sense (most readily seen by observing that in (10.23) **R** has all of its parameters dependent solely upon τ, and not upon the absolute time. Hence, the probability density depends only upon τ, as does any statistical property derived from it. The theory for stationary time series is highly developed, and it is commonly assumed that one is dealing in nature with stationary processes, but one must be alert to the potential failure of the assumption.

Sample Estimates

In working with random variables, it is very important to distinguish between a theoretical value, e.g., the true average, written $< y >$, or m, and the *sample* value, such as the sample average, written as either $< y >_N$ or $\tilde{m}$, where the N subscript indicates that it was based upon N observations. We use a tilde, $\tilde{}$, to mark an estimate of a variable; estimates are themselves always random variables (e.g. $\tilde{m}$) where the parameter itself, m is not.

The usual sample average is

$$\tilde{m} =< y >_N = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{10.58}$$

A useful property of a sample mean (or *estimator* of the mean) is that it's own expected value should be equal to the true value (this is not always true, and as will be seen, it is not always completely desirable.) For example,

$$< \tilde{m} >= \frac{1}{N}\sum_{i=1}^{N} < y_i >= \frac{1}{N}\sum_{i=1}^{N} m = m \tag{10.59}$$

Such an estimator is said to be unbiased. (If we could average N experimental values in M separate experiments, the average of the sample averages would be expected to be the true average.)

It helps to know what is the scatter of an estimate about its true value, so consider the variance:

$$< (\tilde{m} - m)^2 >=< \left[\frac{1}{N}\sum_{i=1}^{N} y_i - m\right]^2 >=< \left[\frac{1}{N}\sum_{i=1}^{N} (y_i - m)\right]^2 > \tag{10.60}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} < (y_i - m)^2 >= \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \tag{10.61}$$

so that the standard deviation of a sample mean about the true mean is $\sigma/\sqrt{N}$, which is the famous "square-root of N" rule.

Now consider the sample variance:

$$\tilde{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{m})^2 \tag{10.62}$$

computed as the sample variance about the sample mean. A little algebra (left as an exercise), shows that,

$$< \tilde{\sigma}^2 >= \frac{N-1}{N}\sigma^2 \neq \sigma^2 \tag{10.63}$$

that is to say, the sample variance is *biassed* (although it is asymptotically unbiased as $N \to \infty$). The bias is readily removed by re-defining

$$\tilde{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \tilde{m})^2. \tag{10.64}$$

The origin of the bias in this case is that in (10.62) the N terms being averaged are not independent of each other; rather since the sum is taken over the values of $y_i - \tilde{m}$, the last term could be predicted from the preceding $N-1$ of them by the requirement that the sample average actually is $\tilde{m}$. Consequently there are only $N-1$ independent terms ($N-1$ "degrees-of-freedom") in the sum.

Consistent Estimators

Data are used to estimate various properties characterizing the statistical population from which the observation come. For example, the sample mean (10.58), is intended to be a good estimate of the true average m, from which the data were drawn. Many other properties are estimated, including the variance and the power density spectrum (defined below). Any estimate of a property has to be examined for at least two desirable properties: (1) that the average value of the estimate should be the true value (for an unbiased estimate), at least as the number of samples becomes large, and (2) that as the number of samples becomes large, the variance of the sample about its mean value (which one might hope is the true value) ought to go to zero. Estimators with these two properties are said to be "consistent". Sometimes however, demanding an unbiased estimator greatly increases the variance of the estimate, and one may deliberately permit a bias if the variance can thereby be reduced. The sample mean was shown above to be unbiased, and by (10.60) its variance about the true value diminishes with N; hence it is a consistent estimator. One can also show that $\tilde{\sigma}^2$ in the form (10.64) is also a consistent estimator as is the original definition, in the asymptotic limit.

A useful idea in this context is the Chebyschev inequality. Let ξ be any random variable with true mean m, variance σ^2 and second moment $<\xi^2>= m_2$ (that is not taken about the mean). Then

$$m_2 = \int_{-\infty}^{\infty} \Xi^2 p(\Xi)\, d\Xi \geq \left(\int_{-\infty}^{-\delta} + \int_{\delta}^{\infty}\right) \Xi^2 p(\Xi)\, d\Xi \tag{10.65}$$

$$\geq \delta^2 \left(\int_{-\infty}^{-\delta} + \int_{\delta}^{\infty}\right) p(\Xi)\, d\Xi \tag{10.66}$$

since ξ^2 is always greater than or equal to δ^2 in the integral. So we have the weak inequality

$$\int_{|\xi|>\delta} p(\Xi)\, d\Xi \leq \frac{m_2}{\delta^2}, \tag{10.67}$$

which we can read as "the probability that $|\xi| > \delta$ is less than or equal to m_2/δ^2." If we replace ξ by $\xi - m$, and then $m_2 = \sigma^2$, we have

$$\text{prob}\{|\xi - m| \geq \delta\} \leq \frac{\sigma^2}{\delta^2} \tag{10.68}$$

where "prob" denotes "the probability that". The sense of the inequality can be inverted to

$$\text{prob}\left\{|\xi - m| \leq \delta\right\} \geq 1 - \frac{\sigma^2}{\delta^2}, \tag{10.69}$$

which is the Chebyschev inequality. These inequalities in turn lead to the important idea of "convergence in probability". ξ_k (where k is a parameter) is said to converge in probability to c, as $k \to \infty$, if $\text{prob}\{|\xi_k - c| \geq \delta\} \to 0$. This is convergence that is "almost surely", written a.s., and differs from ordinary mathematical convergence in being a statement that deviation from the asymptotic value becomes extremely improbable, but not impossible, in the limit.

Let us apply the Chebyschev inequality to the sample mean (10.58), whose variance is σ^2/N. Then by (10.69),

$$\text{prob}\left\{|\tilde{m} - m| \geq \delta\right\} \leq \frac{\sigma^2}{N\delta^2} \tag{10.70}$$

Thus as $N \to \infty$ (corresponding to the parameter k above), the probability that the sample mean differs from the true mean by any amount δ can be made arbitrarily small. It is thus a consistent estimator. For the spectral estimators, etc. used below, one needs formally to show convergence in probability to the true values, but this demonstration is normally left to the reader.

Confidence Intervals

Consider the sample mean $\tilde{m}$, whose own mean is m and whose variance is σ^2/N. If it is derived from x_i which are normal independent, identically distributed (i.i.d) variates, then $\tilde{m}$ is $G\left(m, \sigma^2/N\right)$. Let us make a table of the canonical $G\left(0,1\right)$ variable:

$\eta\left(\alpha/2\right)$	prob$[-\eta \leq X \leq \eta]$	α
1.0	0.683	0.317
1.96	0.950	0.050
2.00	0.954	0.046
2.58	0.990	0.010
3.00	0.997	0.003

Table Caption. (Taken from Jenkins and Watts, 1968, p. 71). Here η is the value, symmetric about the mean of 0, between which the probability is $1 - \alpha$ that a random sample X would lie. α is the fraction of the value which would lie outside the range $\pm\eta$. Thus a random sample X would have a probability of $0.95 = 1 - \alpha$ of lying in the range ± 1.96. In many trials, one would expect 5% of the values to lie, by chance, outside this range. The normal density is symmetric about the mean, and which is not true for more general densities.

Then

$$\text{prob}\left\{-\eta\left(\alpha/2\right) \leq \frac{\tilde{m} - m}{\sigma/\sqrt{N}} \leq \eta\left(\alpha/2\right)\right\} = 1 - \alpha \tag{10.71}$$

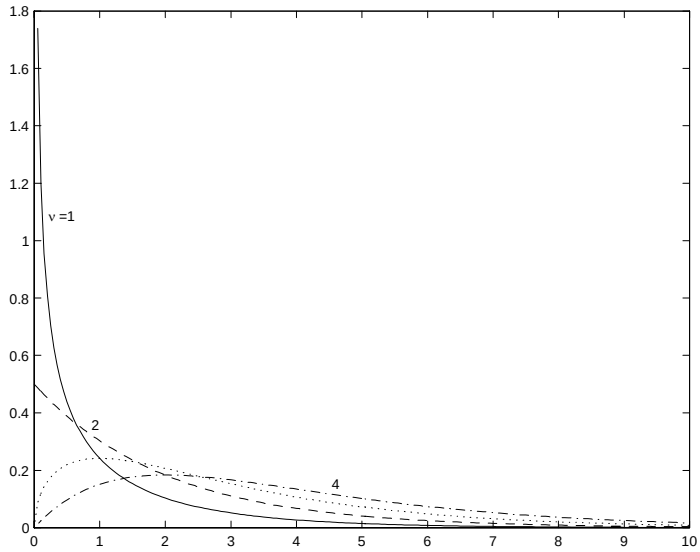

FIGURE 9. Probability density for $\chi^2_\nu,\ \nu = 1, ..., 4$. Note growth of long positive tail as ν increases. $m = \nu$, and $\sigma^2 = 2\nu$. For large ν, χ^2_ν approaches a normal probability density. The cases $\nu = 1, 2$ are exceptional in having their maximum at 0.

can be read from the table for any given α. We can re-arrange (10.71) to:

$$\text{prob}\left\{m - \frac{\eta(\alpha/2)\,\sigma}{\sqrt{N}} \leq \tilde{m} \leq m + \frac{\eta(\alpha/2)\,\sigma}{\sqrt{N}}\right\} = 1 - \alpha \tag{10.72}$$

Alternatively,

$$\text{prob}\left\{\tilde{m} - \left(\sigma/\sqrt{N}\right)\eta(\alpha/2) \leq m \leq \tilde{m} + \left(\sigma/\sqrt{N}\right)\eta(\alpha/2)\right\} = 1 - \alpha. \tag{10.73}$$

This last form is the most useful one: suppose to be specific that $\alpha = .05$. Then for a sample mean $\tilde{m}$ obtained from a random variate having variance σ^2, the probability is 0.95 that the true value, m, lies in the interval $\tilde{m} \pm \sigma\eta(.05/2)$. This interval is said to be a $1-\alpha\%$ confidence interval (here a 95% confidence interval). It clearly diminishes to zero as $N \to \infty$.

Consider now a non-symmetric probability density. One that proves very useful to us is the so-called χ^2_ν (chi-square with ν degrees-of-freedom),

$$\begin{aligned} p_\xi(X) &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} X^{\nu/2-1}\exp(-X/2), \quad X > 0 \\ &= 0, \quad X \leq 0 \end{aligned} \tag{10.74}$$

whose mean is ν, and variance is 2ν. (This probability density describes a variable $x = \sum_1^\nu \xi_i^2$, where the ξ_i are independent, $G(0,1)$.) It is plotted in fig. 9 for $1 \leq \nu \leq 4$. For present purposes, we note only that the probability density is non-symmetric about its mean, having a long tail toward high positive values. Consider the tail at the low end containing a fraction $a/2$ of the values from a set of random

trials, and denote the value X below which this tail falls as $\eta_-(\alpha/2)$. Correspondingly, the tail at the high end occupying $\alpha/2$ of the trial results is to lie to the right of $X = \eta_+(\alpha/2)$. Suppose now that we have a variable ξ which has been estimated as $\tilde{\xi}$ and which is thought to be distributed in χ^2_ν with mean ν and variance 2ν. Then

$$\text{prob}\left\{\eta_-(\alpha/2) \leq \tilde{\xi} \leq \eta_+(\alpha/2)\right\} = 1-\alpha. \tag{10.75}$$

To employ this in a practical example, consider the sample variance $\tilde{\sigma}^2$ (10.64) constructed from N idenditically distributed variables, which are $G\left(m,\sigma^2\right)$. It is easy to show (sketched out below) that $(N-1)\tilde{\sigma}^2/\sigma^2$ will be a χ^2_ν variable with $\nu = N-1$ (it's expected value is $N-1$, and its variance is $2(N-1)$). Then setting $\tilde{\xi} = (N-1)\tilde{\sigma}^2/\sigma^2$ in (10.75) and rearranging, we obtain

$$\text{prob}\left\{\frac{\tilde{\sigma}^2(N-1)}{\eta_+(\alpha/2)} \leq \sigma^2 \leq \frac{\tilde{\sigma}^2(N-1)}{\eta_-(\alpha/2)}\right\} = 1-\alpha \tag{10.76}$$

and the true variance would like between these two (unequal) bounds about $\tilde{\sigma}^2$. Because $\eta_- < \eta_+$, the upper limit of the $\alpha\%$ confidence limit will be further above σ^2 than the lower limit will be below (for $\nu > 2$). $\eta_\pm$ are tabulated in various statistics books or can be calculated from various software packages. (It is sometimes useful to have the χ^2_ν distribution for the sum, ξ, of squared zero-mean Gaussian variates of variance σ^2, or:

$$\begin{aligned} p_\xi(X) &= \frac{1}{\sigma^2 2^{\nu/2}\Gamma(\nu/2)}\left(\frac{X}{\sigma^2}\right)^{\nu/2-1}\exp\left(-X/\left(2\sigma^2\right)\right), \quad X \geq 0 \\ &= 0, \quad X < 0.) \end{aligned} \tag{10.77}$$

In examining the form (10.73), one might object that σ^2 is not likely to be known. This led Gosset (writing under the famous pseudonym of "Student") to show that he could find the probability density of the variable $T_{N-1} = \sqrt{N}(\tilde{m}-m)/\tilde{\sigma}$ and which is not dependent upon σ. The resulting probability density is called Student's $t-$distribution. We leave its discussion to the references (see, e.g., Cramér, 1946, Section 18.2).

White Noise

A white noise process is a stationary time series (sequence) with a uniform in time variance, zero mean, and in which knowledge of the value at one time θ_m carries no information about its value at any other time, including the immediately preceding and following times. That is, $< \theta_m\theta_{m'} >= \sigma^2_\theta\delta_{mm'}$. A sequence of coin flips is such a process. The general terminology is that these are independent identically distributed variables (or an i.i.d). Often, we will assume that the values are normally distributed, as in (10.2) with $m = 0$.

White noise is the simplest possible stochastic process. Let us therefore consider its Fourier transform or series using the real form (6.29-6.31). The coefficients are

$$a_k = \frac{2}{N}\sum_{p=0}^{N-1}\theta_p \cos\left(\frac{2\pi kp\Delta t}{T}\right), k = 0, ..., N/2, \tag{10.78}$$

$$b_k = \frac{2}{N}\sum_{p=0}^{N-1}\theta_p \sin\left(\frac{2\pi kp\Delta t}{T}\right), k = 1, ..., N/2-1. \tag{10.79}$$

It follows immediately that

$$\begin{aligned}
&< a_k >=< b_k >= 0 \\
&< a_k^2 >= \frac{4}{N^2} < \sum_{p=0}^{N-1}\theta_p \cos\left(\frac{2\pi kp\Delta t}{T}\right)\sum_{r=0}^{N-1}\theta_r \cos\left(\frac{2\pi kp\Delta t}{T}\right) > \\
&= \frac{4}{N^2}\sum\sum < \theta_p\theta_r > \cos\left(\frac{2\pi kp\Delta t}{T}\right)\cos\left(\frac{2\pi kp\Delta t}{T}\right) \\
&= \frac{4}{N^2}\sum\sum \delta_{pr}\sigma_\theta^2 \cos\left(\frac{2\pi kp\Delta t}{T}\right)\cos\left(\frac{2\pi kp\Delta t}{T}\right) \\
&= \frac{2}{N}\sigma_\theta^2
\end{aligned} \tag{10.80}$$

by (6.30). Similarly,

$$< b_k^2 >= \frac{2}{N}\sigma_\theta^2 \tag{10.81}$$

$$< a_n b_k >= 0 \tag{10.82}$$

$$< a_k a_n >=< b_k b_n >= \frac{2}{N}\delta_{kn}\sigma_\theta^2, \tag{10.83}$$

omitting the zero and Nyquist frequencies. For these frequencies,

$$< a_0^2 >=< a_{N/2}^2 >= \frac{4}{N}. \tag{10.84}$$

To say this in words: the Fourier transform of a white noise process has zero mean and is uncorrelated from one frequency to another; the sine and cosine amplitudes are uncorrelated with each other at all frequencies, and the variance of the sine and cosine components is uniform with frequency. If the θ_m are normally distributed, $G\left(0, \sigma_\theta^2\right)$ then it follows immediately that a_k, b_k are also normally distributed $G\left(0, 2\sigma_\theta^2/N\right)$. The Parseval relationship requires

$$\frac{1}{N}\sum_{m=0}^{N-1}\theta_m^2 = \frac{a_0^2}{4} + \frac{1}{2}\sum_{m=1}^{N/2-1}\left(a_m^2 + b_n^2\right) + \frac{a_{N/2}^2}{4}, \tag{10.85}$$

here including the mean and Nyquist frequencies. (The true mean is zero, but the actual sample mean will not be, although it is often set to zero for numerical reasons. Doing so, means that only $N-1$ of

the terms on the left in (10.85) would then be independent.) To check this last equation, we can see if it holds on the average:

$$\frac{1}{N}\sum_{m=0}^{N-1} < \theta_m^2 > \stackrel{?}{=} \frac{< a_0^2 >}{4} + \frac{1}{2}\sum_{n=1}^{N/2-1} < \left(a_n^2 + b_n^2\right) > + \frac{< a_{N/2}^{2/2} >}{4}, \text{ or,}$$

$$\sigma_\theta^2 \stackrel{?}{=} \frac{\sigma_\theta^2}{N} + \frac{1}{2}\sum_{n=1}^{N/2-1}\left(\frac{2}{N}+\frac{2}{N}\right)\sigma_\theta^2 + \frac{\sigma_\theta^2}{N} = \frac{\sigma_\theta^2}{N} + \frac{1}{2}\left(\frac{N}{2}-1\right)\frac{4}{N}\sigma_\theta^2 + \frac{\sigma_\theta^2}{N} = \sigma_\theta^2, \tag{10.86}$$

as required.

As the record length grows, the number N, of Fourier coefficients grows linearly, but the mean square power $(1/N)\sum_{m=0}^{N-1}\theta_m^2$ in (10.85) remains fixed, independent of N. Thus the expected value of any $a_n^2+b_n^2$ is reduced by the factor $1/N$ to compensate for their growing population.

If one computes the phase of the Fourier series as

$$\phi_n = \tan^{-1}\left(b_n/a_n\right) \tag{10.87}$$

it is readily seen that ϕ_n has a uniform probability density

$$p_\phi\left(\Phi\right) = \frac{1}{2\pi}. \tag{10.88}$$

(To see this algebraically, we recognize that if θ_m are $G\left(0,\sigma_\theta^2\right)$, a_n, b_n are uncorrelated Gaussian variables distributed $G\left(0,\sigma_\theta^2/2N\right)$ and hence they are also independent. Because they are independent, their joint probability density is the product of their identical probability densities:

$$p_{a_n,b_n}\left(\Xi_1,\Xi_2\right) = p_{a_n}\left(\Xi_1\right)p_{b_n}\left(\Xi_2\right) = \tag{10.89}$$
$$\frac{1}{2\pi\left(\sigma_\theta^2/2N\right)}\exp\left(-\Xi_1^2/\left(\sigma_\theta^2/2N\right)\right)\exp\left(-\Xi_2^2/\left(\sigma_\theta^2/2N\right)\right)$$

We can thus conclude that there is no information content in the phase of a white noise process. What then is the probability density for a_n^2, b_n^2 and $a_n^2+b_n^2$? The probability densities for the first two must be identical, and are thus the densities for the square of a normal variate with 0 mean and variance $\sigma_\theta^2/2N$. Let us use normalized variables so that they have unit variance, i.e., consider $a_n' = a_n/\left(\sigma_\theta/\sqrt{2N}\right), b_n' = b_n/\left(\sigma_\theta/\sqrt{2N}\right)$ each of which will be $G(0,1)$. Using the rule for change of variable we find that,

$$p_{a_n'^2}\left(X\right) = p_{b_n'^2}\left(X\right) = \frac{X^{-1/2}}{\sqrt{2\pi}}\exp\left(-X/2\right) \tag{10.90}$$

which is again χ_1^2 (recall Eq. (10.15)). Invoking the machinery for sums of independent variables, we find that the probability density for $r_n'^2 = a_n'^2 + b_n'^2$ is

$$p_{r_n'^2}\left(X\right) = \frac{1}{2}\exp\left(-X/2\right) \tag{10.91}$$

which is called χ_2^2 (chi-square with two-degrees-of-freedom) and whose mean is 2 and variance is 4 (one has to do the integral; see Eq. 10.74).

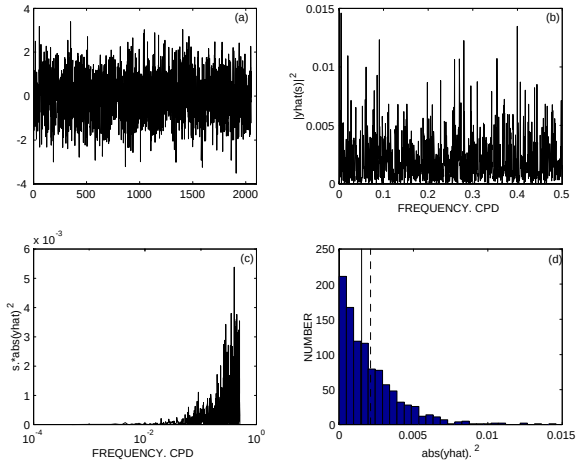

FIGURE 10. (a) 2048 numbers from a pseudo-random generator. These approximate a white noise process. (b) Linear scale periodogram of the values in (a). (c) Semi-logarithmic scale plot of $s\,|\alpha_n|^2$ in which the multiplication by s compensates for the logarithmic squeezing of the frequencies at the high end. Notice the visual impression that energy increases with increasing frequency—an artifact of the plotting method. (d) 30 bin histogram of the values $|\alpha_n|^2$ which should and do approximate a χ^2_2 variate. The mean (dashed) and median (solid) values are shown. There is evidently a much greater chance of a value far larger than the mean to occur than there is for a much smaller one to occur, owing to the limiting value at zero. But more smaller values below the mean occur than above the mean of χ^2_2.

Exercise. Using the results of the exercise on P. 37, find the probability density for the amplitude and phase of the Fourier coefficients.

Fig. 10 shows the Fourier coefficients of a pseudo-random θ_n, their sum of squares, $r_n^2 = a_n^2 + b_n^2$, and the histogram of occurrences of r_n^2 along with the theoretical χ^2_2 (scaling the variables by $\sigma_\theta/2N$ to restore the original variance). The true mean would be

$$< r_n^2 >= 2\frac{\sigma_\theta^2}{2N} = \frac{\sigma_\theta^2}{N} \tag{10.92}$$

and the variance of r_n^2 would be

$$< \left(r_n^2 - < r_n^2 >\right)^2 >= 2\left(\frac{\sigma_\theta^2}{N}\right)^2 = \left(\frac{\sigma_\theta^2}{N}\right)^2. \tag{10.93}$$

That is, the variance is proportional to the square of the of the power density.

Definition. $r_n^2 = a_n^2 + b_n^2 = |\alpha_n|^2$ is called the "periodogram" of a random process.

REMARK 1. *There is a common, naive, assertion about the results of Fourier transforming a stochastic process. Basically, it says that "obviously, the low frequency Fourier components are less-well determined than are the high frequency ones, because many more oscillations of the high frequency components exist in the record, whereas for the first record harmonic, $s_1 = 1/T$, only one cycle is covered by the data." Unfortunately this seemingly obvious conclusion is false. Consider that every Fourier harmonic, s_n, differs from its two neighboring harmonics, s_{n-1}, s_{n+1} by exactly one cycle per record length. All of the data are required to determine the Fourier amplitudes of these neighbors, separating them by consideration of their difference by one cycle in the record. s_1 differs exactly in the same way from s_0 (the record mean), and s_2. Thus precisely the same amount of information is available about s_1 as about any other s_n. This statement is consistent with (10.93): the variance of the periodogram values is independent of n.*

It appears from (10.93) that the variance of the estimates diminishes with N. But the ratio of the variance to the estimate itself, is independent of N. That is, because the number of Fourier coefficients increases with N, each has a diminishing proportional contribution to the constant record variance (power), and the variability of the estimate as a fraction of the true expected value remains unchanged even as the record length goes to infinity. This behavior confused scientists for years.

11. Spectral Estimation

We have seen that the Fourier coefficients (periodogram) of a normal white noise process are uncorrelated (and thus independent) random variables, whose phase has no information except that its underlying probability density is uniform. The squares of the absolute values are distributed in χ_2^2 such that the mean values $< a_n^2 + b_n^2 >= \sigma_\theta^2/N$ (if one prefers to use the complex Fourier series, $< |\alpha_n|^2 >= \sigma_\theta^2/2N$, with the understanding that $< |\alpha_n|^2 >=< |\alpha_{-n}|^2 >$, that is, half the variance is at negative frequencies, but one could decide to double the power for positive n and ignore the negative values). Its variance about the mean is given by (10.93).

Suppose we wished to test the hypothesis that a particular time series is in fact white noise. Figure 10b is so noisy, that one should hesitate in concluding that the Fourier coefficients have no structure inconsistent with white noise. To develop a quantitative test of the hypothesis, we can attempt to use the result that $< \left|\alpha_n^2\right| >$ should be a constant independent of n. A useful, straightforward, approach is to exploit the independence of the neighboring Fourier coefficients. Let us define a power spectral estimate as

$$\tilde{\Psi}^\nu (s_n) = \frac{1}{[\nu/2] + 1} \sum_{p=n-[\nu/4]}^{n+[\nu/4]} |\alpha_n|^2 \tag{11.1}$$

where for the moment, ν is taken to be an odd number and $[\nu/4]$ means "largest integer in $\nu/4$". That is the power spectral estimate is a local average over $[\nu/2]$ of the squared Fourier Series coefficients surrounding frequency s_n. Fig. 11a shows the result for a white noise process, where the average was over 2 local values surrounding s_n (4 degrees of freedom). Fig. 11b shows an average over 6 neighboring

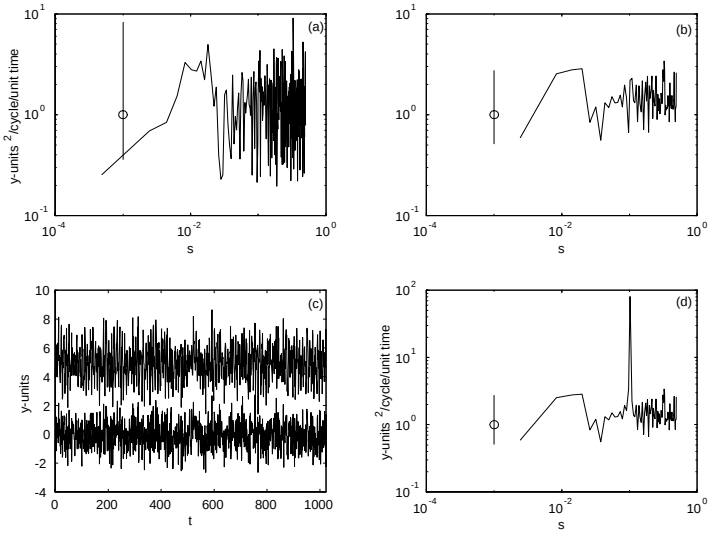

FIGURE 11. (a) Power density spectral estimate of a white noise process averaged over 2 frequency bands (4 degrees-of-freedom), and (b) over 6 frequency bands (12 degrees of freedom). An approximate 95% confidence interval is shown as obtained from (10.76). (c) Lower curve is the white noise process whose spectra are shown in (a), (b), and the upper curve is the same white noise ($\sigma_\theta^2 = 1$) plus a unit amplitude sine wave, displaced upwards by 5 units. Visually, the presence of the sine wave is difficult to detect. (d) Power density of upper curve in (c), making the spectral peak quite conspicuous, and much larger, relative to the background continuum than the 95% confidence limit.

frequency estimates (12 degrees-of-freedom). The local averages are obviously a good deal smoother than the periodogram is, as one expects from the averaging process. The probability density for the local average $\Psi^\nu(s_n)$ is evidently that for the sum of $[\nu/2]$ variables, each of which is a χ_2^2 variable. That is, $\tilde{\Psi}^\nu(s_n)$ is a χ_ν^2 random variable with ν degrees-of-freedom (2 degrees-of-freedom come from each periodogram value, the sine and cosine part, or equivalent real and imaginary part, being uncorrelated variables). The mean $< \tilde{\Psi}^\nu(s_n) >= \sigma_\theta^2/N$ and the variance is

$$< \left(\tilde{\Psi}^\nu(s_n) - \Psi^\nu(s_n)\right)^2 >= \frac{\sigma_\theta^4}{2N^2\nu}, \tag{11.2}$$

which goes to zero as $\nu \to \infty$ for fixed N. Visually, it is much more apparent from Fig. 11 that the underlying Fourier coefficients have a constant value, albeit, some degree of variability cannot be ruled out as long as ν is finite.

This construct is the basic idea behind spectral *estimation.* The periodogram is numerically and visually extremely noisy. To the extent that we can obtain an estimate of its mean value by local frequency band averaging, we obtain a better-behaved statistical quantity. From the probability density

of the average, we can construct expected variances and confidence limits that can help us determine if the variances of the Fourier coefficients are actually independent of frequency.

The estimate (11.1) has one obvious drawback: the expected value depends upon N, the number of data points. It is often more convenient to obtain a quantity which would be independent of N, so that for example, if we obtain more data, the estimated value would not change; or if we were comparing the energy levels of two different records of different length, it would be tidier to have a value independent of N. Such an estimate is easily obtained by dividing $\tilde{\Psi}^{\nu}(s_n)$ by $1/N$ (multiplying by N), to give

$$\tilde{\Phi}^{\nu}(s_n) = \frac{1}{([\nu/2]+1)/N} \sum_{p=n-[\nu/4]}^{n+[\nu/4]} |\alpha_p|^2 . \tag{11.3}$$

This quantity is the called the estimated "power spectral density", and it has a simple interpretation. The distance in frequency space between the individual periodogram values is just $1/N$ (or $1/N\Delta t$ if one puts in the sampling interval explicitly). When averaging, we sum the values over a frequency interval $([\nu/2]/N\Delta t)$. Because $\sum_{p=n-[\nu/4]}^{n+[\nu/4]} |\alpha_n|^2$ is the fractional power in the range of averaging, $\tilde{\Phi}^{\nu}(s_n)$, is just the power/unit frequency width, and hence the *power density.* (If one works with the Fourier transform the normalization factors change in the obvious way.) For a stochastic process, the power density is independent of the data length.

Exercise. Show analytically that the power density for a pure sinusoid is *not* independent of the data length.

A very large number of variations can be played upon this theme of stabilizing the periodogram by averaging. We content ourselves by mainly listing some of them, leaving the textbooks to provide details.

The average over the frequency band need not be uniform. One may prefer to give more weight to frequencies towards the center of the frequency band. Let W_n be any set of averaging weights, then (11.3) can be written very generally as

$$\tilde{\Phi}^{\nu}(s_n) = \frac{1}{([\nu/2]+1)/N} \frac{\sum_{p=n-[\nu/4]}^{n+[\nu/4]} W_p |\alpha_p|^2}{\sum_{p=n-[\nu/4]}^{n+[\nu/4]} W_p}, \tag{11.4}$$

and the bands are normally permitted to overlap. When the weights are uniform, and no overlap is permitted, they are called the "Daniell" window. When the averaging is in non-overlapping bands the values of $\tilde{\Phi}^{\nu}(s_n)$ in neighboring bands are uncorrelated with each other, asymptotically, as $N \to \infty$. With overlapping bands, one must estimate the expected level of correlation. The main issue is the determination then of ν—the number of degrees-of-freedom. There are no restrictions on ν being even or odd. Dozens of different weighting schemes have been proposed, but rationale for the different choices is best understood when we look at the Blackman-Tukey method, a method now mainly of historical interest.

An alternative, but completely equivalent estimation method is to exploit explicitly the wide-sense stationarity of the time series. Divide the record up into $[\nu/2]$ non-overlapping pieces of length M (Fig.

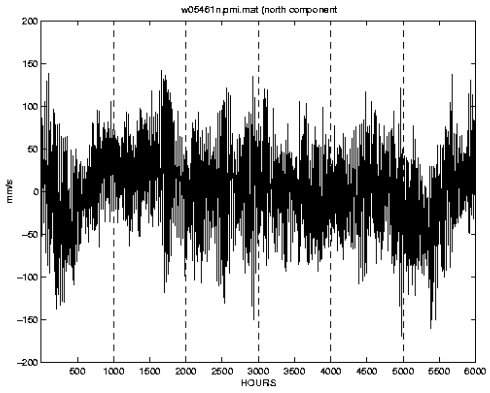

FIGURE 12. North component of a current meter record divided into 6 non-overlapping segments. One computes the periodogram of each segment and averages them, producing approximately 2 degrees of freedom from each segment (an approximation dependent upon the spectrum not being too far from white).

12) and form the periodogram of each piece $\left|\alpha_n^{(p)}\right|^2$ (The frequency separation in each periodogram is clearly smaller by a factor $[\nu/2]$ than the one computed for the entire record.) One then forms

$$\tilde{\Phi}^{\nu}(s_n) = \frac{1}{([\nu/2]+1)/M} \sum_{p=1}^{[\nu/2]} \left|\alpha_n^{(p)}\right|^2 \tag{11.5}$$

For white noise, it is possible to prove that the estimates in (11.3, 11.5) are identical. One can elaborate these ideas, and for example, allow the sections to be overlapping, and also to do frequency band averaging of the periodograms from each piece prior to averaging those from the pieces. The advantages and disadvantages of the different methods lie in the trade-offs between estimator variance and bias, but the intent should be reasonably clear. The "method of faded overlapping segments" has become a common standard practice, in which one permits the segments to overlap, but multiplies them by a "taper", W_n prior to computing the periodogram. (See Percival and Walden).

12. The Blackman-Tukey Method

Prior to the advent of the FFT and fast computers, power density spectral estimation was almost never done as described in the last section. Rather the onerous computational load led scientists, as far as possible, to reduce the number of calculations required. The so-called Blackman-Tukey method, which became the de facto standard, begins with a purely theoretical idea. Let $< x_n >= 0$. Define the "sample autocovariance",

$$\tilde{R}(\tau) = \frac{1}{N} \sum_{n=0}^{N-1-|\tau|} x_n x_{n+\tau}, \tau = 0, \pm 1, \pm 2, ..., \pm N - 1, \tag{12.1}$$

where as τ grows, the number of terms in the sum necessarily diminishes. From the discrete convolution theorem, it follows that,

$$\mathcal{F}\left(\tilde{R}(\tau)\right) = \sum_{\tau=-(N-1)}^{N-1} \tilde{R}(\tau) \exp(-2\pi i s\tau) = \frac{1}{N} |\hat{x}(s)|^2 = N |\alpha_n|^2 \tag{12.2}$$

Then the desired power density is,

$$< N |\alpha_n|^2 >= \Phi(s) = \sum_{\tau=-(N-1)}^{N-1} < \tilde{R}(\tau) > \exp(-2\pi i s\tau). \tag{12.3}$$

Consider

$$\begin{aligned} < \tilde{R}(\tau) >&= \frac{1}{N} \sum_{m=0}^{N-1-|\tau|} < x_m x_{m+\tau} >, \tau = 0, \pm 1, \pm 2, ..., \pm N - 1 \\ &= \frac{N - |\tau|}{N} R(\tau), \end{aligned} \tag{12.4}$$

by definition of $R(\tau)$. First letting $N \to \infty$, and then $\tau \to \infty$, we have the Wiener-Khinchin theorem:

$$\Phi(s) = \sum_{\tau=-\infty}^{\infty} R(\tau) \exp(-2\pi i s\tau) = \sum_{\tau=-\infty}^{\infty} R(\tau) \cos(2\pi s\tau) \tag{12.5}$$

the power density spectrum of a stochastic process is the Fourier transform of the autocovariance. This relationship is an extremely important *theoretical* one. (One of the main mathematical issues of time series analysis is that the limit as $T = N\Delta t \to \infty$ of the Fourier transform or series,

$$\alpha_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) \exp\left(\frac{2\pi i n t}{T}\right) dt, \tag{12.6}$$

whether an integral or sum does not exist (does not converge) because of the stochastic behavior of $x(t)$, but the Fourier transform of the autocovariance (which is *not random*) does exist.). It is important to recognize that unlike the definition of "power spectrum" used above for non-random (deterministic) functions, an expected value operation is an essential ingredient when discussing stochastic processes.

It is very tempting (and many people succumbed) to assert that $\tilde{R}(\tau) \to R(\tau)$ as N becomes very large. The idea is plausible because (12.1) looks just like an average. The problem is that no matter how large N becomes, (12.2) requires the Fourier transform of the *entire* sample autocovariance. As the

lag $\tau \to N$, the number of terms in the average (12.1) diminishes until the last lag has only a single value in it—a very poor average. While the lag 0 term may have thousands of terms in the average, the last term has only one. The Fourier transform of the sample autocovariance includes these very poorly determined sample covariances; indeed we know from (12.2) that the statistical behavior of the result must be exactly that of the periodogram—it is unstable (inconsistent) as an estimator because its variance does not diminish with N.

The origin of this instability is directly derived from the poorly estimated large-lag sample covariances.[4] The Blackman-Tukey method does two things at once: it reduces the variance of the periodogram, and minimizes the number of elements which must be Fourier transformed. This is a bit confusing because the two goals are quite different. Once one identifies the large lag τ values of $\tilde{R}(\tau)$ as the source of the statistical instability, the remedy is clear: get rid of them. One multiplies $\tilde{R}(\tau)$ by a "window" w_τ and Fourier transforms the result

$$\tilde{\Phi}^{\nu}(s) = \sum_{\tau=-(N-1)}^{\tau=N-1} \tilde{R}(\tau)\, w_\tau \exp(-2\pi i s\tau) \tag{12.8}$$

By the convolution theorem, this is just

$$\tilde{\Phi}^{\nu}(s) = \mathcal{F}\left(\tilde{R}(\tau)\right) * \mathcal{F}(w_\tau) \tag{12.9}$$

If w_τ is such that its Fourier transform is a local averaging operator, then (12.9) is exactly what we seek, a local average of the periodogram. If we can select w_τ so that it simultaneously has this property, and so that it actually vanishes for $|\tau| > M$, then the Fourier transform in (12.8) is reduced from being taken over N-terms to over $M << N$, that is,

$$\tilde{\Phi}^{\nu}(s) = \sum_{\tau=-(M-1)}^{\tau=M-1} \tilde{R}(\tau)\, w_\tau \exp(-2\pi i s\tau)\,. \tag{12.10}$$

The Blackman-Tukey estimate is based upon (12.9, and 12.10) and the choice of suitable window weights w_τ. A large literature grew up devoted to the window choice. Again, one trades bias against variance through the value M, which one prefers greatly to minimize. The method is now obsolete because the ability to generate the Fourier coefficients directly permits much greater control over the result. The bias discussion of the Blackman-Tukey method is particularly tricky, as is the determination of ν. Use of the method should be avoided except under those exceptional circumstances when for some reason only

[4]Some investigators made the situation much worse by the following plausible argument. For finite τ, the number of terms in (12.1) is actually not N, but $N - |\tau|$; .they argued therefore, that the proper way to calculate $R(\tau)$ was actually

$$\tilde{R}_1(\tau) = \frac{1}{N-|\tau|} \sum_{n=0}^{N-1-|\tau|} x_t x_{t+\tau} \tag{12.7}$$

which would be (correctly) an unbiassed estimator of $R(\tau)$. They then Fourier transformed $\tilde{R}_1(\tau)$ instead of $R(\tau)$. But this makes the situation much worse: by using (12.7) one gives greatest weight to the least well-determined components in the Fourier analysis. One has traded a reduction in bias for a vastly increased variance, in this case, a very poor choice indeed. (Bias does not enter if x_t is white noise, as all terms of both $< \tilde{R} >, < \tilde{R}_1 >$ vanish except for $\tau = 0$.)

$\tilde{R}(\tau)$ is known. (For large data sets, it is actually computationally more efficient to compute $\tilde{R}(\tau)$, should one wish it, by first forming $\tilde{\Phi}^{\nu}(s)$ using FFT methods, and then obtaining $\tilde{R}(\tau)$ as its inverse Fourier transform.)

13. Colored Processes

A complete statistical theory is readily available for white noise processes. But construction of the periodogram or spectral density estimate of real processes (e.g., Fig. 13) quickly shows that a white noise process is both uninteresting, and not generally applicable in nature.

For colored noise processes, many of the useful simplifications for white noise break down (e.g., the zero covariance/correlation between Fourier coefficients of different frequencies). Fortunately, many of these relations are still asymptotically (as $N \to \infty$) valid, and for finite N, usually remain a good zero-order approximation. Thus for example, it is commonplace to continue to use the confidence limits derived for the white noise spectrum even when the actual spectrum is a colored one. This degree of approximation is generally acceptable, as long as the statistical inference is buttressed with some physical reasoning, and/or the need for statistical rigor is not too great. Fortunately in most of the geophysical sciences, one needs rough rules of thumb, and not rigid demands for certainty that something is significant at 95% but not 99% confidence.

As an example, we display in Fig. 14 the power density spectral estimate for the current meter component whose periodogram is shown in Fig. 13. The 95% confidence interval shown is only approximate, as it is derived rigorously for the white noise case. Notice however, that the very high-frequency jitter is encompassed by the interval shown, suggesting that the calculated confidence interval is close to the true value. The inertial peak is evidently significant, relative to the background continuum spectrum at 95% confidence. The tidal peak is only marginally significant (and there are formal tests for peaks in white noise; see e.g., Chapter 8 of Priestley). Nonetheless, the existence of a theory which suggests that there should be a peak at exactly the observed period of 12.42 hours would convince most observers of the reality of the peak irrespective of the formal confidence limit. Note however, that should the peak appear at some unexpected place (e.g., 10 hours) where there is no reason to anticipate anything special, one would seek a rigorous demonstration of its reality before becoming too excited about it.

Another application of this type of analysis can be seen in Fig. 15, where the power density spectra of the tide gauge records at two Pacific islands are shown. For many years it had been known that there was a conspicuous, statistically significant, but unexplained, spectral peak near 4 days period at Canton Island, which is very near the equator.

Plotting of Power Density Spectra

Some words about the plotting of spectra (and periodograms) is helpful. For most geophysical purposes, the default form of plot is a log-log one where both estimated power density and the frequency scale are logarithmic. There are several reasons for this (1) many geophysical processes produce spectral densities which are power laws, s^{-q} over one or more ranges of frequencies. These appear as straight

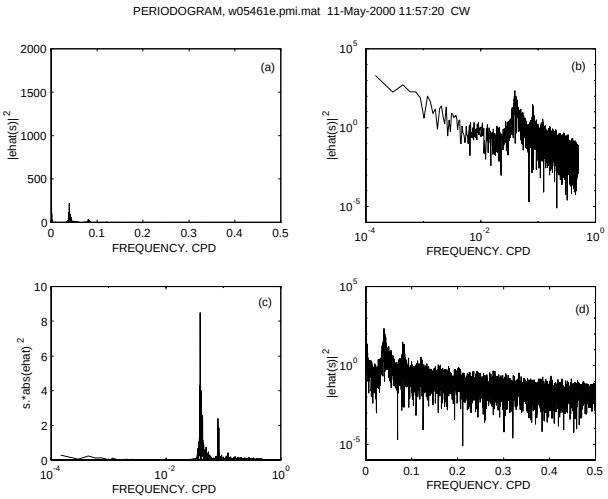

FIGURE 13. Periodogram of the zonal component of velocity at a depth of 498m, 27.9°N,54.9°W in the North Atlantic, plotted in four different ways. (a) Both scales are linear. The highest value is invisible, being lost against the $y-$axis. The rest of the values are dominated by a peak at the local inertial frequency ($f/2\pi$). In (b) a log-log scale is used. Now one sees the entirety of the values, and in particular the near power-law like behavior away from the inertial peak. But there are many more frequency points at the high frequency end than there are at the low frequency end, and in terms of relative power, the result is misleading.
(c) A so-called area-preserving form in which a linear-log plot is made of $s\left|\alpha_n\right|^2$ which compensates the abcissa for the crushing of the estimates at high frequencies. The relative power is proportional to the area under the curve (not so easy to judge here by eye), but suggesting that most of the power is in the inertial peak and internal waves ($s > f/2\pi$). (d) shows a logarithmic ordinate against linear frequency, which emphasizes the very large number of high frequency components relative to the small number at frequencies below $f/2\pi$.

lines on a log-log plot and so are easily recognized. A significant amount of theory (the most famous is Kolmogorov's wavenumber $k^{-5/3}$ rule, but many other theories for other power laws exist as well) leading to these laws has been constructed. (2) Long records are often the most precious and one is most interested in the behavior of the spectrum at the lowest frequencies. The logarithmic frequency scale compresses the high frequencies, rendering the low frequency portion of the spectrum more conspicuously. (3) The "red" character of many natural spectra produces such a large dynamic range in the spectrum that much of the result is invisible without a logarithmic scale. (4) The confidence interval is nearly constant on a log-power density scale; note that the confidence interval (and sample variance) are dependent upon $\tilde{\Phi}^{\nu}(s)$,

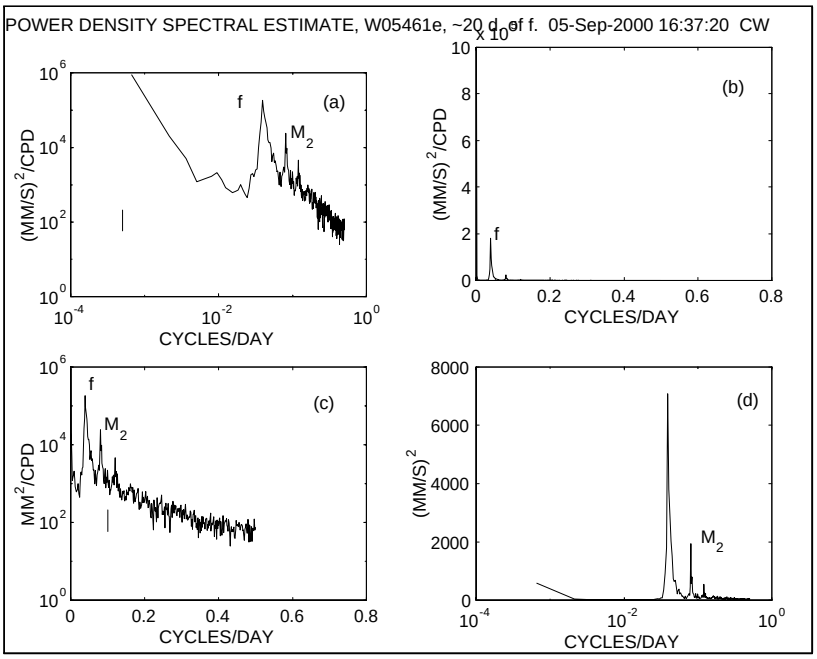

FIGURE 14. Four different plots of a power spectral density estimate for the east component of the current meter record whose periodogram is displayed in Fig. 13. There are approximately 20 degrees of freedom in the estimates. An approximate 95% confidence interval is shown for the two plots with a logarithmic power density scale; note that the high frequency estimates tend to oscillate within about this interval. (b) is a linear-linear plot, and (d) is a so-called area preserving plot, which is linear-log. The Coriolis frequency, denoted f, and the principal lunar tidal peaks (M_2) are marked. Other tidal overtones are apparently present.

with larger estimates having larger variance and confidence intervals. The fixed confidence interval on the logarithmic scale is a great convenience.

Several other plotting schemes are used. The logarithmic frequency scale emphasizes the low frequencies at the expense of the high frequencies. But the Parseval relationship says that all frequencies are on an equal footing, and one's eye is not good at compensating for the crowding of the high frequencies. To give a better pictorial representation of the energy distribution, many investigator's prefer to plot $s\tilde{\Phi}^{\nu}(s)$ on a linear scale, against the logarithm of s. This is sometimes known as an "area-preserving plot" because it compensates the squeezing of the frequency scale by the multiplication by s (simultaneously reducing the dynamic range by suppression of the low frequencies in red spectra). Consider how this behaves (Fig. 16) for white noise. The log-log plot is flat within the confidence limit, but it may not be immediately obvious that most of the energy is at the highest frequencies. The area-preserving plot looks "blue" (although we know this is a white noise spectrum). The area under the curve is proportional

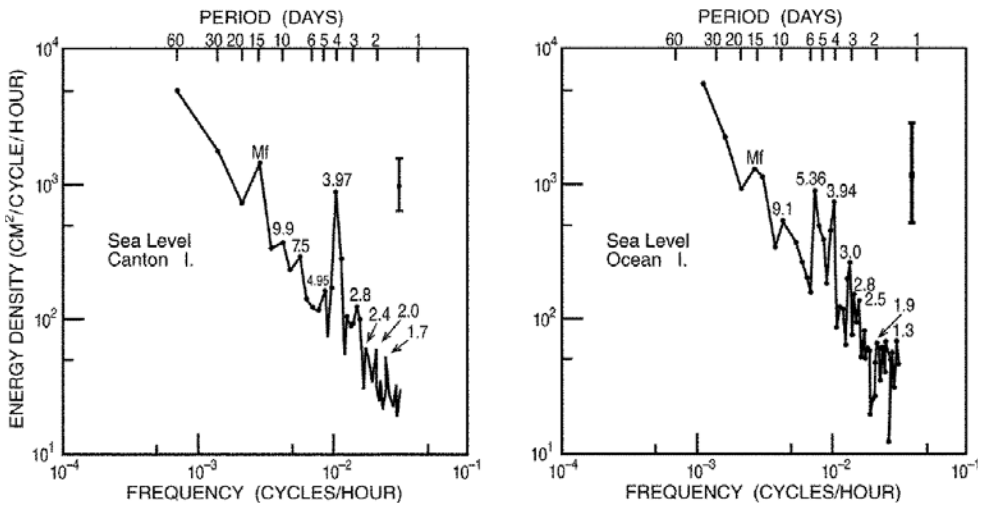

FIGURE 15. Power spectral density of sealevel variations at two near-equatorial islands in the tropical Pacific Ocean (Wunsch and Gill, 1976). An approximate 95% confidence limit is shown. Some of the peaks, not all of which are significant at 95% confidence, are labelled with periods in days (Mf is however, the fortnightly tide). Notice that in the vicinity of 4 days the Canton Island record shows a strong peak (although the actual sealevel variability is only about 1 cm rms), while that at Ocean I. shows not only the 4 day peak, but also one near 5 days. These results were explained by Wunsch and Gill (1976) as the surface manifestation of equatorially trapped baroclinic waves. To the extent that a theory predicts 4 and 5 day periods at these locations, the lack of rigor in the statistical arguments is less of a concern. (The confidence limit is rigorous only for a white noise spectrum, and these are quite "red". Note too that the high frequency part of the spectrum extending out to 0.5 cycles/hour has been omitted from the plots.)

to the fraction of the variance in any fixed logarithmic frequency interval, demonstrating that most of the record energy is in the high frequencies. One needs to be careful to recall that the underlying spectrum is constant. The confidence interval would differ in an obvious way at each frequency. A similar set of plots is shown in Fig. (14) for a real record.

Beginners often use linear-linear plots, as it seems more natural. This form of plot is perfectly acceptable over limited frequency ranges; it becomes extremely problematic when used over the complete frequency range of a real record, and its use has clearly distorted much of climate science. Consider Fig. 17 taken from a tuned ocean core record (Tiedemann, et al., 1994) and whose spectral density estimate (Fig. 19) is plotted on linear-linear scale.

One has the impression that the record is dominated by the energy in the Milankovitch peaks (as marked; the reader is cautioned that this core was "tuned" to produce these peaks and a discussion of their reality or otherwise is beyond our present scope). But in fact they contain only a small fraction

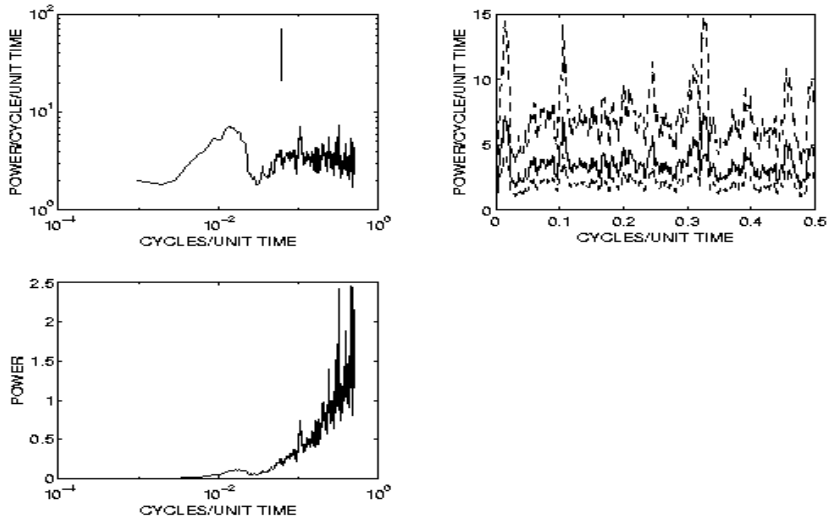

FIGURE 16. Power density spectral estimate for pseudorandom numbers (white noise) with about 12 degrees-of-freedom plotted in three different ways. Upper left figure is a log-log plot with a single confidence limit shown, equal to the average for all frequencies. Upper right is a linear plot showing the upper and lower confidence limits as a function of frequency, which is necessary on a linear scale. Lower left is the area-preserving form, showing that most of the energy on the logarithmic frequency scale lies at the high end, but giving the illusion of a peak at the very highest frequency.

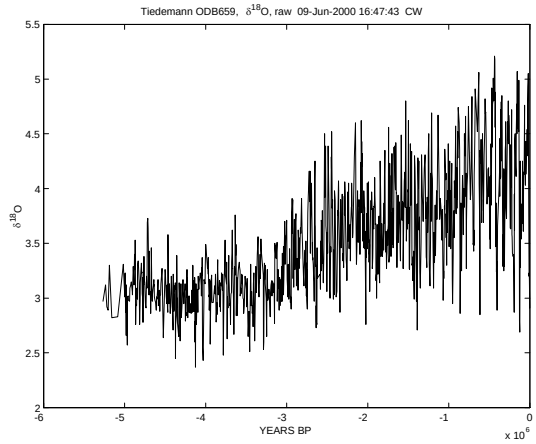

FIGURE 17. Time series from a deep-sea core (Tiedemann, et al., 1994). The time scale was tuned under the assumption that the Milankovitch frequencies should be prominent.

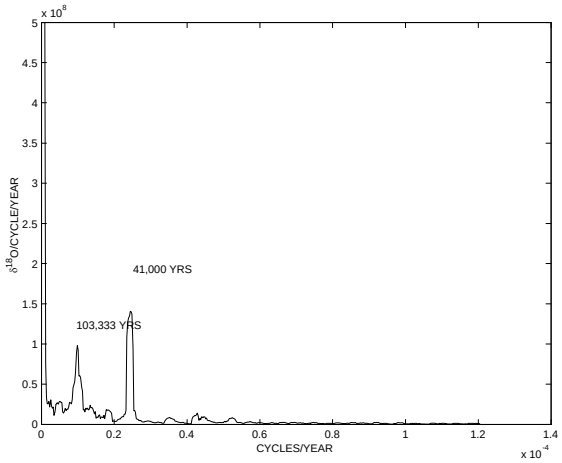

FIGURE 18. Linear-linear plot of the power density spectral estimate of the Tiedemann et al. record. This form may suggest that the Milankovitch periodicities dominate the record (although the tuning has assured they exist).

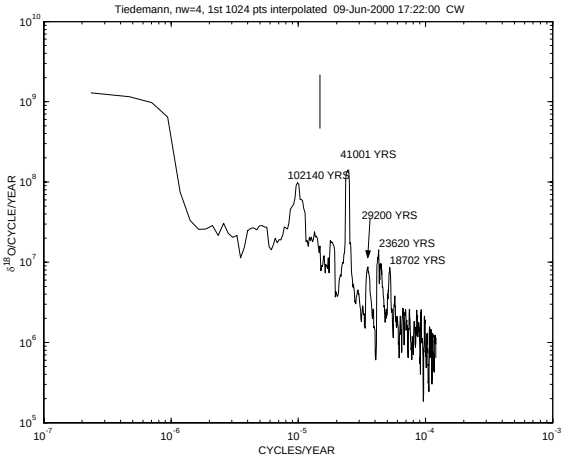

FIGURE 19. The estimated power density spectrum of the record from the core re-plotted on a log-log scale. Some of the significant peaks are indicated, but this form suggests that the peaks contain only a small fraction of the energy required to describe this record. An approximate 95% confidence limit is shown and which is a nearly uniform interval with frequency.

of the record variance, which is largely invisible solely because of the way the plot suppresses the lower values of the much larger number of continuum estimates. Again too, the confidence interval is different for every point on the plot.

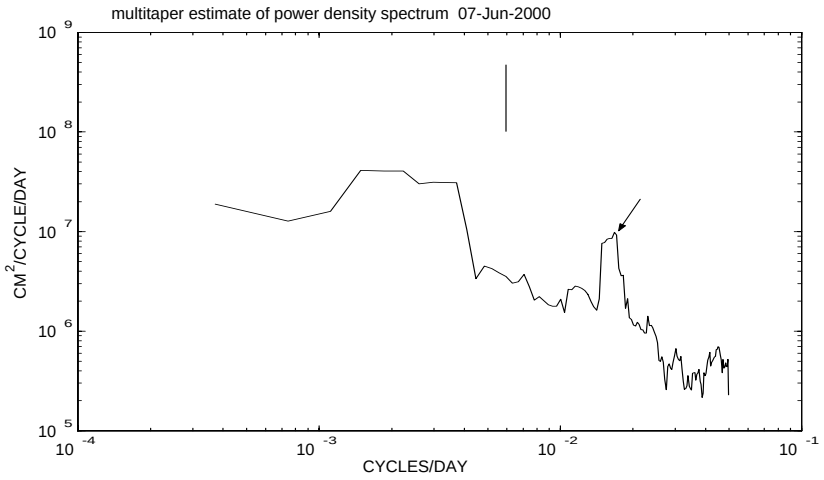

FIGURE 20. Power density spectra from altimetric data in the western Mascarene Basin of the Indian Ocean. The result shows (B. Warren, personal communication, 2000) that an observed 60 day peak (arrow) in current meter records also appears conspicuously in the altimetric data from the same region. These same data show no such peak further east. As described earlier, there is a residual alias from the M_2 tidal component at almost exactly the observed period. Identifying the peak as beng a resonant mode, rather than a tidal line, relies on the narrow band (pure sinusoidal) nature of a tide, in contrast to the broadband character of what is actually observed. (Some part of the energy seen in the peak, may however be a tidal residual.)

As one more example, Fig. 20 shows a power density spectrum from altimetric data in the Mascarene Basin of the western Indian Ocean. This and equivalent spectra were used to confirm the existence of a spectral peak near 60 days in the Mascarene Basin, and which is absent elsewhere in the Indian Ocean.

14. The Multitaper Idea

Spectral analysis has been used for well over 100 years, and its statistics are generally well understood. It is thus surprising that a new technique appeared not very long ago. The methodology, usually known as the *multitaper* method, is associated primarily with David Thompson of Bell Labs, and is discussed in detail by Percival and Walden (1993). It is probably the best default methodology. There are many details, but the basic idea is not hard to understand.

Thompson's approach can be thought of as a reaction to the normal tapering done to a time series before Fourier transforming. As we have seen, one often begins by tapering x_m before Fourier transforming it so as to suppress the leakage from one part of the spectrum to another. As Thompson has noted however, this method is equivalent to discarding the data far from the center of the time series (setting it to small values or zero), and any statistical estimation procedure which literally throws away data is

unlikely to be a very sensible one—real information is being discarded. Suppose instead we construct a series of tapers, call them $w_m^{(i)}, 1 \le i \le P$ in such as way that

$$\sum_{m=0}^{N-1} w_m^{(i)} w_m^{(j)} = \delta_{ij} \tag{14.1}$$

that is, they are orthogonal tapers. The first one $w_m^{(1)}$ may well look much like an ordinary taper going to zero at the ends. Suppose we generate P new time series

$$y_m^{(i)} = x_m w_m^{(i)} \tag{14.2}$$

Because of the orthogonality of the $w_m^{(i)}$, there will be a tendency for

$$\sum_{m=0}^{N-1} y_m^{(i)} y_m^{(j)} \approx 0, \tag{14.3}$$

that is, to be uncorrelated. The periodograms $\left|\alpha_k^{(i)}\right|^2$ will thus also tend to be nearly uncorrelated and if the underlying process is near-Gaussian, will therefore be nearly independent. We therefore estimate

$$\tilde{\Phi}^{2P}(s) = \frac{1}{P} \sum_{i}^{P} \left|\alpha_k^{(i)}\right|^2 \tag{14.4}$$

from these nearly independent periodograms.

Thompson showed that there was an optimal choice of the tapers $w_m^{(i)}$ and that it is the set of prolate spheroidal wavefunctions (Fig. 21). For the demonstration that this is the best choice, and for a discussion of how to compute them, see Percival and Walden (1993) and the references there to Thompson's papers. (Note that the prolate spheroidal wave functions are numerically tricky to calculate, and approximating sinusoids do nearly as well; see McCoy et al., 1998, who also discuss a special problem with the estimates near zero frequency)

15. Spectral Peaks

Pure sinusoids are very rare in nature, generally being associated with periodic astronomical phenomena such as tides, or the Milankovitch forcing in solar insolation. Thus apart from phenomena associated with these forcings, deterministic sinusoids are primarily of mathematical interest. To the extent that one suspects a pure "tone" in a record not associated with an astronomical forcing, it would be a most unusual, not-to-say startling, discovery. (These are called "line spectra".) If you encounter a paper claiming to see pure frequency lines at anything other than at the period of an astronomical phenomenon, it's a good bet that the author doesn't know what he is doing. (Many people like to believe that the world is periodic; mostly it isn't.)

But because both tides and Milankovitch responses are the subject of intense interest in many fields, it is worth a few comments about line spectra. We have already seen that unlike a stochastic process, the

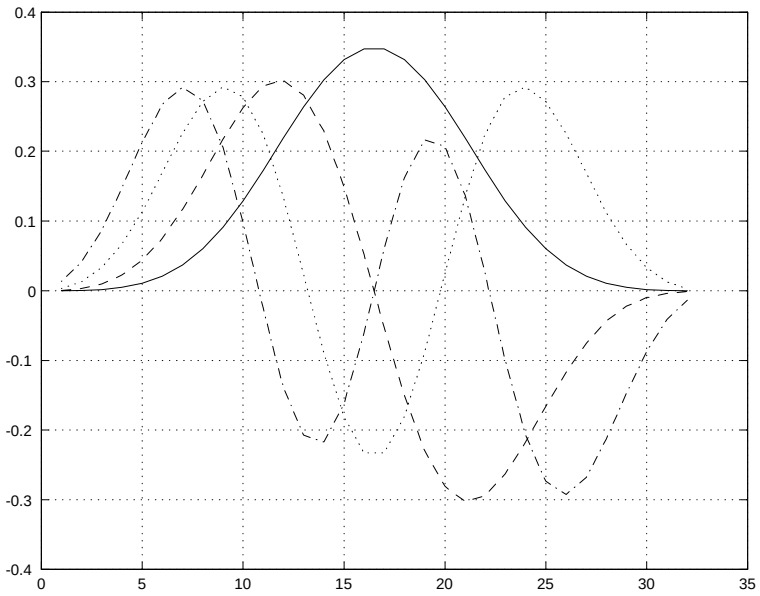

FIGURE 21. First 4 discretized prolate spheroidal wave functions (also known as Slepian sequences) used in the multitaper method for a data duration of 32. Methods for computing these functions are described in detail by Percival and Walden (1993) and they were found here using a MATLAB toolbox function. Notice that as the order increases, greater weight is given to data near the ends of the observations.

Fourier coefficients of a pure sinusoid do not diminish as the record length, N, increases (alternatively, the Fourier transform value increases with N, while for the stochastic process they remain fixed in rms amplitude). This behavior produces a simple test of the presence of a pure sinusoid: double (or halve) the record length, and determine whether the Fourier coefficient remains the same or changes.

Much more common in nature are narrow-band peaks, which represent a relative excess of energy, but which is stochastic, and not a deterministic sinusoid. A prominent example is the peak in the current meter spectra (Fig.14) associated with the Coriolis frequency). The ENSO peak in the Southern Oscillation Index (Wunsch, 1999) is another example, and many others exist. None of these phenomena are represented by line spectra. Rather they are what is sometimes called "narrow-band" random processes. It proves convenient to have a common measure of the sharpness of a peak, and this measure is provided by what electrical engineers call Q (for "quality factor" associated with a degree of resonance). A damped mass-spring oscillator, forced by white noise, and satisfying an equation like

$$m\frac{d^2x}{dt^2} + r\frac{dx}{dt} + kx = \theta(t) \tag{15.1}$$

will display an energy peak near frequency $s = (2\pi)^{-1}\sqrt{k/m,}$. as in Fig. 22. The sharpness of the peak depends upon the value of r. Exactly the same behavior is found in almost any linear oscillator (e.g., an

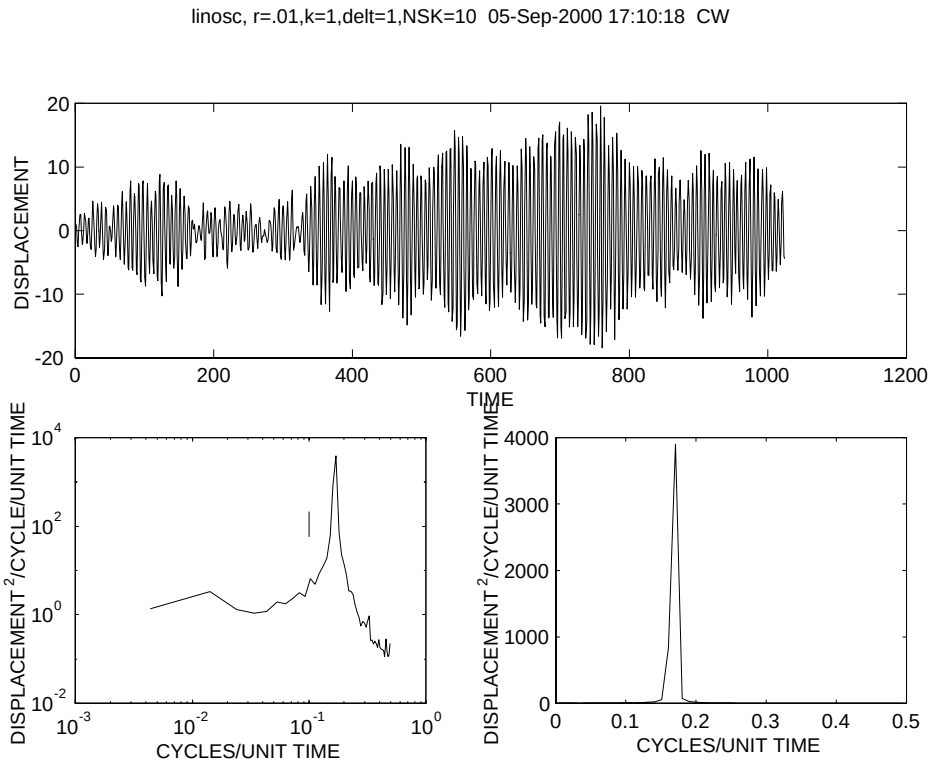

FIGURE 22. (Top) Time series of displacement of a simple mass spring oscillator, driven by white noise, and computed numerically such that $r/m = 0.1, k/m = 1$ with $\Delta t = 1$. Lower left and right panels are the estimated power density spectrum plotted differently. The Q here exceeds about 20.

organ pipe, or an $L - C$ electrical circuit). Peak width is measured in terms of

$$Q = \frac{s_0}{\Delta s}, \tag{15.2}$$

(e.g., *Jackson,* 1975). Here s_0 is the circular frequency of the peak center and Δs is defined as the bandwidth of the peak at its half-power points. For linear systems such as (15.1), it is an easy matter to show that an equivalent definition is,

$$Q = \frac{2\pi E}{< dE/dt >} \tag{15.3}$$

where here E is the peak energy stored in the system, and $< dE/dt >$ is the mean rate of energy dissipated over one cycle. It follows that for (15.1),

$$Q = \frac{2\pi s_0}{r} \tag{15.4}$$

(Jackson, 1975; Munk and Macdonald, 1960, p. 22). As r diminishes, the resonance is greater and greater, $\Delta s \to 0, < dE/dt > \to 0$, and $Q \to \infty$, the resonance becoming perfect.

Exercise. Write (15.1) in discrete form; calculate numerical x_m for white noise forcing, and show that the power density spectrum of the result is consistent with the three definitions of Q.

Exercise. For the parameters given in the caption of Fig. 22, calculate the value of Q.

Values of Q for the ocean tend to be in the range of $1-20$ (see, e.g., Luther, 1982). The lowest free elastic oscillation of the earth (first radial mode) has a Q approaching 10,000 (the earth rings like a bell for months after a large earthquake), but this response is extremely unusual in nature and such a mode may well be thought of as an astronomical one.

A Practical Point

The various normalizations employed for power densities and related estimates can be confusing if one, for example, wishes to compute the rms amplitude of the motion in some frequency range, e.g., that corresponding to a local peak. Much of the confusion can be evaded by employing the Parseval relationship (2.4). First compute the record variance, $\tilde{\sigma}^2$, and form the accumulating sum $\tilde{\Phi}^d(s_n) = \sum_{k=1}^{n}\left(a_k^2+b_k^2\right), n \le [N/2]$, assuming negligible energy in the mean. Then the fraction of the power lying between any two frequencies, n, n', must be $\left(\tilde{\Phi}^d(s_n)-\tilde{\Phi}^d(s'_n)\right)/\tilde{\Phi}^d\left(s_{[N/2]}\right)$; the root-mean-square amplitude corresponding to that energy is

$$\left(\sqrt{\left(\tilde{\Phi}^d(s_n)-\tilde{\Phi}^d(s'_n)\right)/\tilde{\Phi}^d\left(s_{[N/2]}\right)}\right)\tilde{\sigma}/\sqrt{2}, \tag{15.5}$$

and the normalizations used for $\tilde{\Phi}$ drop out.

16. Spectrograms

If one has a long enough record, it is possible to Fourier analyze it in pieces, often overlapping, so as to generate a spectrum which varies with time. Such a record, usually displayed as a contour plot of frequency and time is known as a "spectrogram". Such analyses are used to test the hypothesis that the frequency content of a record is varying with time, implying a failure of the stationarity hypothesis. The inference of a failure of stationarity has to be made very carefully: the χ^2_ν probability density of any spectral estimate implies that there is expected variability of the spectral estimates made at different times, even if the underlying time series is strictly stationary. Failure to appreciate this elementary fact often leads to unfortunate inferences (see Hurrell, 1995 and the comments in Wunsch, 1999).

The need to localize frequency structure in a time-or space-evolving record is addressed most generally by wavelet analysis. This comparatively recent development is described and discussed at length by Percival and Walden (2000) but often conventional spectrogram analysis is completely adequate (if seemingly less sophisticated).

17. Effects of Timing Errors

The problem of errors in the sampling times t_m, whether regularly spaced or not, is not a part of the conventional textbook discussion, because measurements are typically obtained from instruments for which clock errors are normally quite small. But for instruments whose clocks drift, but especially when analyzing data from ice or deep ocean cores, the inability to accurately date the samples is a major problem. A general treatment of clock error or "timing jitter" may be found in Moore and Thomson (1991) and Thomson and Robinson (1996), with a simplified version applied to ice core records in Wunsch (2000).

18. Cross-Spectra and Coherence

Definitions

"Coherence" is a measure of the degree of relationship, as a function of frequency, between two time series, x_m, y_m. The concept can be motivated in a number of ways. One quite general form is to postulate a convolution relationship,

$$y_m = \sum_{-\infty}^{\infty} a_k x_{m-k} + n_m \tag{18.1}$$

where the residual, or noise, n_m, is uncorrelated with x_m, $< n_m x_p >= 0$. The a_k are not random (they are "deterministic"). The infinite limits are again simply convenient. Taking the Fourier or $z-$transform of (18.1), we have

$$\hat{y} = \hat{a}\hat{x} + \hat{n}. \tag{18.2}$$

Multiply both sides of this last equation by $\hat{x}^*$ and take the expected values

$$< \hat{y}\hat{x}^* >= \hat{a} < \hat{x}\hat{x}^* > + < \hat{n}\hat{x}^* > \tag{18.3}$$

where $< \hat{n}\hat{x}^* >= 0$ by the assumption of no correlation between them. Thus,

$$\Phi_{yx}(s) = \hat{a}(s)\,\Phi_{xx}(s), \tag{18.4}$$

where we have defined the "cross-power" or "cross-power density," $\Phi_{yx}(s) =< \hat{y}\hat{x}^* >$. Eq. (18.4) can be solved for

$$\hat{a}(s) = \frac{\Phi_{yx}(s)}{\Phi_{xx}(s)} \tag{18.5}$$

and Fourier inverted for a_n.

Define

$$C_{yx}(s) = \frac{\Phi_{yx}(s)}{\sqrt{\Phi_{yy}(s)\,\Phi_{xx}(s)}}. \tag{18.6}$$

C_{yx} is called the "coherence"; it is a complex function, whose magnitude it is not hard to prove is $|C_{yx}(s)| \leq 1$.[5] Substituting into (18.5) we obtain,

$$\hat{a}(s) = C_{yx}(s) \sqrt{\frac{\Phi_{yy}(s)}{\Phi_{xx}(s)}}. \tag{18.7}$$

Thus the phase of the coherence is the phase of $\hat{a}(s)$. If the coherence vanishes in some frequency band, then so does $\hat{a}(s)$ and in that band of frequencies, there is no relationship between y_n, x_n. Should $C_{yx}(s) = 1$, then $y_n = x_n$ in that band of frequencies. (Beware that some authors use the term "coherence" for $|C_{yx}|^2$.)

Noting that

$$\Phi_{yy}(s) = |\hat{a}(s)|^2 \Phi_{xx}(s) + \Phi_{nn}(s), \tag{18.8}$$

and substituting for $\hat{a}(s)$,

$$\Phi_{yy}(s) = \left| C_{yx}(s) \sqrt{\frac{\Phi_{yy}(s)}{\Phi_{xx}(s)}} \right|^2 \Phi_{xx}(s) + \Phi_{nn}(s) = |C_{yx}(s)|^2 \Phi_{yy}(s) + \Phi_{nn}(s) \tag{18.9}$$

Or,

$$\Phi_{yy}(s) \left(1 - |C_{yx}(s)|^2\right) = \Phi_{nn}(s) \tag{18.10}$$

That is, the fraction of the power in y_n at frequency s, *not related to* x_m, is just $\left(1 - |C_{yx}(s)|^2\right)$, and is called the "incoherent" power. It obviously vanishes if $|C_{yx}| = 1$, meaning that in that band of frequencies, y_n would be perfectly calculable (predictable) from x_m. Alternatively,

$$\Phi_{yy}(s) |C_{yx}(s)|^2 = |\hat{a}(s)|^2 \Phi_{xx}(s) \tag{18.11}$$

which is the fraction of the power in y_n that is related to x_n. This is called the "coherent" power, so that the total power in y is the sum of the coherent and incoherent components. These should be compared to the corresponding results for ordinary correlation above.

Estimation

As with the power densities, the coherence has to be estimated from the data. The procedure is essentially the same as estimating e.g., $\Phi_{xx}(s)$. One has observations x_n, y_n. The $\tilde{\Phi}_{xx}(s), \tilde{\Phi}_{yy}(s)$ are estimated from the products $\hat{x}(s_n) \hat{x}(s_n)^*$, etc. as above. The cross-power is estimated from products $\hat{y}(s_n) \hat{x}(s_n)^*$; these are then averaged for statistical stability in the frequency domain (frequency band-averaging) or by prior multiplication of each by one of the multitapers, etc., and the sample coherence obtained from the ratio

$$\tilde{C}^{\nu}_{yx}(s) = \frac{\tilde{\Phi}^{\nu}_{yx}(s)}{\sqrt{\tilde{\Phi}^{\nu}_{yy}(s) \tilde{\Phi}^{\nu}_{xx}(s)}}. \tag{18.12}$$

Exercise. Let $y_n = x_n + (1/2) x_{n-1} + \theta_n$, where θ_n is unit variance white noise. Find, analytically, the coherence, the coherent and incoherent power as a function of frequency and plot them.

[5] Consider $< (\hat{x} + \lambda \hat{y})(\hat{x} + \lambda \hat{y})^* >= \langle |\hat{x}|^2 \rangle + \lambda \langle \hat{y} \hat{x}^* \rangle + \lambda^* \langle \hat{y}^* \hat{x} \rangle + |\lambda|^2 \langle |\hat{y}|^2 \rangle \geq 0$ for any choice of λ. Choose $\lambda = -\langle \hat{x}^* \hat{y} \rangle^2 / \langle |\hat{y}|^2 \rangle$ and substitute. One has $1 - \langle \hat{x} \hat{y}^* \rangle^2 / \langle |\hat{x}|^2 \rangle \langle |y|^2 \rangle \geq 0$.

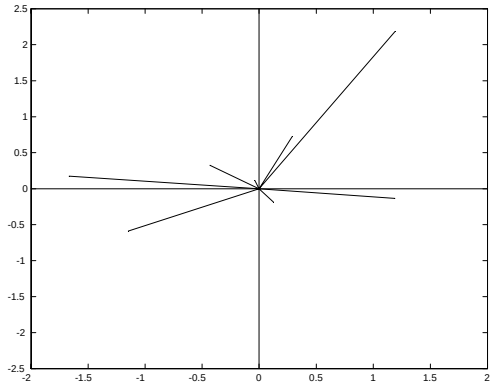

FIGURE 23. The coherence calculation involves summing vectors to produce a dominant direction (determined by the coherence phase) and amplitude determined by the degree of coherence. Here the true coherence $\gamma = 0$, and $\nu = 8$ random vectors are being summed. That they would sum to a zero-length vector is very improbable. As $\nu \to \infty$, the expected length would approach zero.

The Blackman-Tukey method would estimate $\tilde{\Phi}_{yx}^{\nu}(s)$ via the Fourier transform of the truncated (windowed) sample cross-covariance:

$$\tilde{\Phi}_{yx}^{\nu}(s) = \mathcal{F}\left(w_\tau \frac{1}{N} \sum_{n=0}^{N-1} x_n y_{n+\tau} \right),$$

in a complete analogy with the computation by this method of $\tilde{\Phi}_{xx}^{\nu}(s)$, etc. Again, the method should be regarded as primarily of historical importance, rather than something to be used routinely today.

$\tilde{C}_{yx}^{\nu}(s)$ is a somewhat complicated ratio of complex random variates and it is a problem to estimate its probability density. As it is a complex quantity, and as the magnitude, $\left|\tilde{C}(s)\right|$, and phase, $\tilde{\phi}(s)$, have different physical characteristics, the probability densities are sought for both. The probability density for the amplitude of the sample coherence was studied and tabulated by Amos and Koopmans (1962). As this reference is not so easily obtained, a summary of the results are stated here. One has to consider two time series for which the true coherence magnitude, at frequency s is denoted γ. Then note first, the estimate (18.12) is biassed. For example, if $\gamma = 0$ (no coherence), the expected value $\left|\left\langle \tilde{C}_{yx}^{\nu}(s) \right\rangle\right| > 0$. That is, the sample coherence for two truly incoherent time series is expected to be greater than zero with a value dependent upon ν. As $\nu \to \infty$, the bias goes to zero. More generally, if the coherence is finite, $\left| < \tilde{C}_{yx}^{\nu}(s) > \right| > \gamma$, there is a tendency for the sample coherence to be too large.

The reasons for the bias are easy to understand. Let us suppose that the cross-power and auto-power densities are being estimated by a local frequency band averaging method. Then consider the calculation of $\tilde{\Phi}_{yx}^{\nu}(s)$, as depicted in Fig. 23 under the assumption that $\gamma = 0$. One is averaging a group of vectors in the complex plane of varying magnitude and direction. Because the true coherence is zero, the theoretical average should be a zero magnitude vector. But for any finite number ν of such vectors, the probability

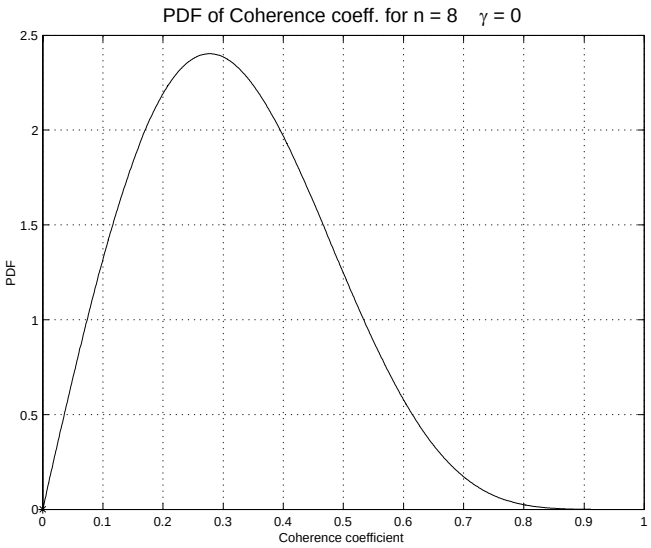

FIGURE 24. Probability density for the sample coherence magnitude, $|C_{yx}(s)|$, when the true value, $\gamma = 0$ for $\nu = 8$. The probability of actually obtaining a magnitude of zero is very small, and the expected mean value is evidently near 0.3.

that they will sum exactly to zero is vanishingly small. One expects the average vector to have a finite length, producing a sample coherence $\tilde{C}_{yx}^{\nu}(s)$ whose magnitude is finite and thus biassed. The division by $\sqrt{\tilde{\Phi}_{yy}^{\nu}(s)\,\tilde{\Phi}_{xx}^{\nu}(s)}$ is simply a normalization to produce a maximum possible vector length of 1.

In practice of course, one does not know γ, but must use the estimated value as the best available estimate. This difficulty has led to some standard procedures to reduce the possibility of inferential error. First, for any ν, one usually calculates the bias level, which is done by using the probability density

$$p_C(Z) = \frac{2\left(1-\gamma^2\right)^{\nu}}{\Gamma(\nu)\,\Gamma(\nu-1)} Z\left(1-Z^2\right)^{\nu-2} \sum_{k=0}^{\infty} \frac{\gamma^{2k}\Gamma^2(\nu+k)\,Z^{2k}}{\Gamma^2(k+1)} \tag{18.13}$$

for the sample coherence amplitude for $\gamma = 0$ and which is shown in Fig. 24 for $\nu = 8$. A conventional procedure is to determine the value C_0 below which $\left|\tilde{C}_{yx}^{\nu}(s)\right|$ will be confined 95% of the time. C_0 is the "level of no significance" at 95% confidence. What this means is that for two time series, which are completely uncorrelated, the sample coherence will lie below this value about 95% of the time, and one would expect, on average for about 5% of the values to lie (spuriously) above this line. See Fig. 25. Clearly this knowledge is vital to do anything with a sample coherence.

If a coherence value lies clearly above the level of no significance, one then wishes to place an error bound on it. This problem was examined by Groves and Hannan (1968; see also Hannan, 1970). As expected, the size of the confidence interval shrinks as $\gamma \to 1$ (see Fig. 26).

The probability density for the sample coherence phase depends upon γ as well. If $\gamma = 0$, the phase of the sample coherence is meaningless, and the residual vector in Fig. 23 can point in any direction at

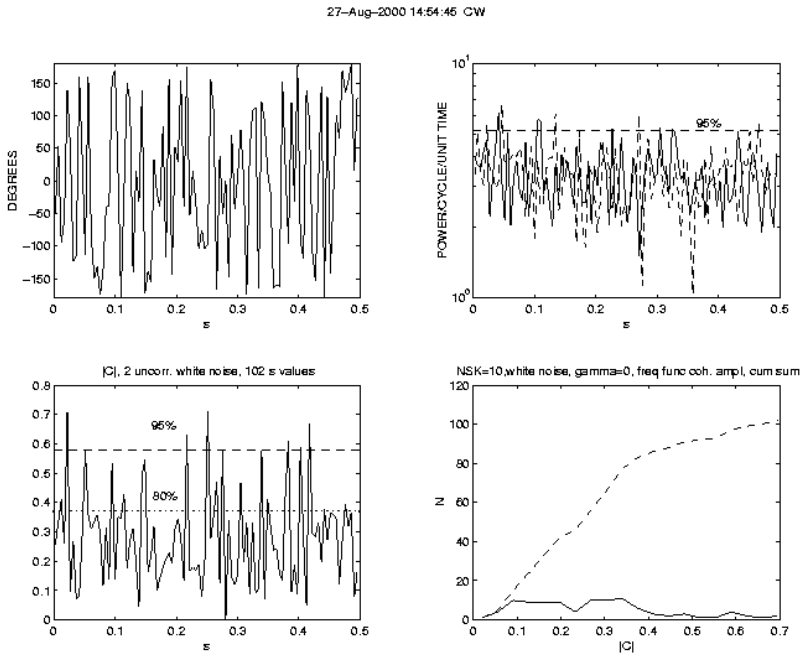

FIGURE 25. Power densities (upper right) of two incoherent (independent) white noise processes. The theoretical value is a constant, and the horizontal dashed line shows the value below which 95% of all values actually occur. The estimated coherence amplitude and phase are shown in the left two panels. Empirical 95% and 80% levels are shown for the amplitude (these are quite close to the levels which would be estimated from the probability density for sample coherence with $\gamma = 0$). Because the two time series are known to be incoherent, it is apparent that the 5% of the values above the 95% level of no significance are mere statistical fluctuations. The 80% level is so low that one might be tempted, unhappily, to conclude that there are bands of significant coherence—a completely false conclusion. For an example of a published paper relying on 80% levels of no significance, see Chapman and Shackleton (2000). Lower right panel shows the histogram of estimated coherence amplitude and its cumulative distribution. Again the true value is 0 at all frequencies. The phases in the upper left panel are indistinguishable from purely random, uniformly distributed, $-\pi \leq \tilde{\phi} \leq \pi$.

all—a random variable of range $\pm\pi$. If $\gamma = 1$, then all of the sample vectors point in identical directions, the calculated phase is then exactly found, and the uncertainty vanishes. In between these two limits, the uncertainty of the phase depends upon γ, diminishing as $\gamma \to 1$. Hannan (1970) gives an expression

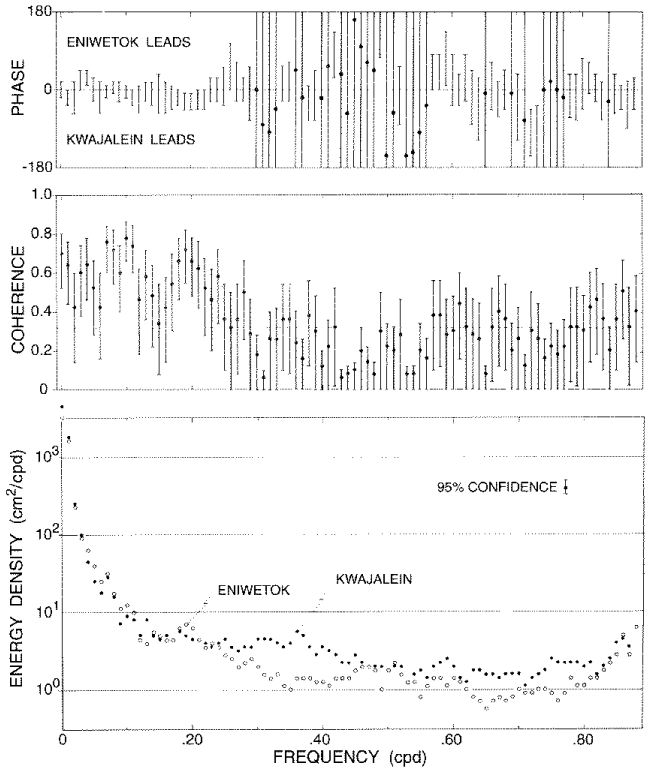

FIGURE 26. (from Groves and Hannan, 1968). Lower panel displays the power density spectral estimates from tide gauges at Kwajalein and Eniwetok Islands in the tropical Pacific Ocean. Note linear frequency scale. Upper two panels show the coherence amplitude and phase relationship between the two records. A 95% level-of-no-significance is shown for amplitude, and 95% confidence intervals are displayed for both amplitude and phase. Note in particular that the phase confidence limits are small where the coherence magnitude is large. Also note that the confidence interval for the amplitude can rise above the level-of-no-significance even when the estimated value is itself below the level.

for the confidence limits for phase in the form (his p. 257, Eq. 2.11):

$$\left|\sin\left[\tilde{\phi}(s)-\phi(s)\right]\right| \leq \left[\frac{1-\tilde{\gamma}^2}{(2\nu-2)\,\tilde{\gamma}^2}\right] t_{2\nu-2}(\alpha)\,. \tag{18.14}$$

Here $t_{2\nu-2}(\alpha)$ is the $\alpha\%$ point of Student's $t-$distribution with $2\nu-2$ degrees of freedom. An alternative approximate possibility is described by Jenkins and Watts (1968, p. 380-381).

Exercise. Generate two white noise processes so that $\gamma = 0.5$ at all frequencies. Calculate the coherence amplitude and phase, and plot their histograms. Compare these results to the expected probability densities for the amplitude and phase of the coherence when $\gamma = 0$. (You may wish to first read the section below on Simulation.)

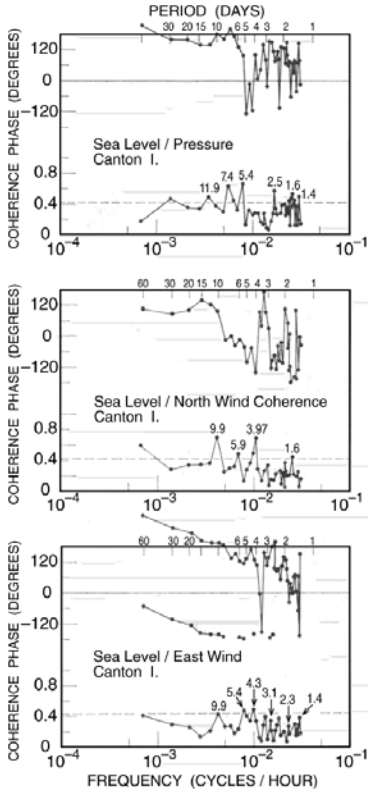

FIGURE 27. Coherence between sealevel fluctuations and atmospheric pressure, north wind and eastwind at Canton Island. An approximate 95% level of no significance for coherence magnitude is indicated. At 95% confidence, one would expect approximately 5% of the values to lie above the line, purely through statistical fluctuation. The high coherence at 4 days in the north component of the wind is employed by Wunsch and Gill (1976, from whom this figure is taken) in a discussion of the physics of the 4 day spectral peak.

Figure 27 shows the coherence between the sealevel record whose power density spectrum was depicted in Fig. 15 (left) and atmospheric pressure and wind components. Wunsch and Gill (1976) use the resulting coherence amplitudes and phases to support the theory of equatorially trapped waves driven by atmospheric winds.

The assumption of a convolution relationship between time series is unnecessary when employing coherence. One can use it as a general measure of phase stability between two time series. Consider for example, a two-dimensional wavefield in spatial coordinates $\mathbf{r} = (r_x, r_y)$ and represented by

$$\eta(\mathbf{r},t) = \sum_n \sum_m a_{nm} \cos(\mathbf{k}_{nm}\cdot\mathbf{r} - \boldsymbol{\sigma}_{nm} t - \phi_{nm}) \tag{18.15}$$

where the a_{nm} are random variables, uncorrelated with each other. Define $y(t) = \eta(\mathbf{r}_1, t)$, $x(t) = \eta(\mathbf{r}_2, t)$. Then it is readily confirmed that the coherence between x, y is a function of $\Delta\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$, as well as σ and the number of wavenumbers $\mathbf{k}_{nm}$ present at each frequnecy. The coherence is 1 when $\Delta\mathbf{r} = 0$, and with falling magnitude with growth in $|\Delta\mathbf{r}|$. The way in which the coherence declines with growing separation can be used to deduce the number and values of the wavenumbers present. Such estimated values are part of the basis for the deduction of the Garrett and Munk (1972) internal wave spectrum.

19. Simulation

In an age of fast computers, and as a powerful test of one's understanding, it is both useful and interesting to be able to generate example time series with known statistical properties. Such simulations can be done both in the frequency and time domains (next Chapter). Suppose we have a power spectral density, $\Phi(s)$—either a theoretical or an estimated one, and we would like to generate a time series having that spectral density.

Consider first a simpler problem. We wish to generate a time series of length N, having a given mean, m, and variance σ^2. There is a trap here. We could generate a time series having exactly this sample mean, and exactly this sample variance. But if our goal is to generate a time series which would be typical of an actual physical realization of a real process having this mean and variance, we must ask whether it is likely any such realization would have these precise sample values. A true coin will have a true (theoretical) mean of 0 (assigning heads as +1, and tails as -1). If we flip a true coin 10 times, the probability that there will be exactly 5 heads and 5 tails is finite. If we flip it 1000 times, the probability of 500 heads and tails is very small, and the probability of "break-even" (being at zero) diminishes with growing data length. A real simulation would have a sample mean which differs from the true mean according to the probability density for sample means for records of that duration. As we have seen above, sample means for Gaussian processes have a probability density which is normal $G\left(0, \sigma^2/N\right)$. If we select each element of our time series from a population which is normal $G(0, \sigma)$, the result will have a statistically sensible sample mean and variance. If we generated 1000 such time series, we would expect the sample means to scatter about the true mean with probability density, $G\left(0, \sigma^2/N\right)$.

So in generating a time series with a given spectral density, we should *not* give it a sample spectral density exactly equal to the one required. Again, if we generated 1000 such time series, and computed their estimated spectral densities, we could expect that their average spectral density would be very close to the required one, with a scattering in a χ^2_ν distribution. How might one do this? One way is to employ our results for the periodogram. Using

$$y_q = \sum_{n=1}^{[T/2]} a_n \cos\left(\frac{2\pi n q}{T}\right) + \sum_{n=1}^{[T/2]-1} b_n \sin\left(\frac{2\pi n q}{T}\right). \tag{19.1}$$

a_n, b_n are generated by a Gaussian

-random number generator $G(0, \Phi(s_n))$ such that $< a_n >=< b_n >= 0$, $< a_n^2 >=< b_n^2 >= \Phi(s = n/T)/2$, $< a_n a_m >=< b_n b_m >= 0, m \neq n$, $< a_n b_m >= 0$. The requirements on the a_n, b_n

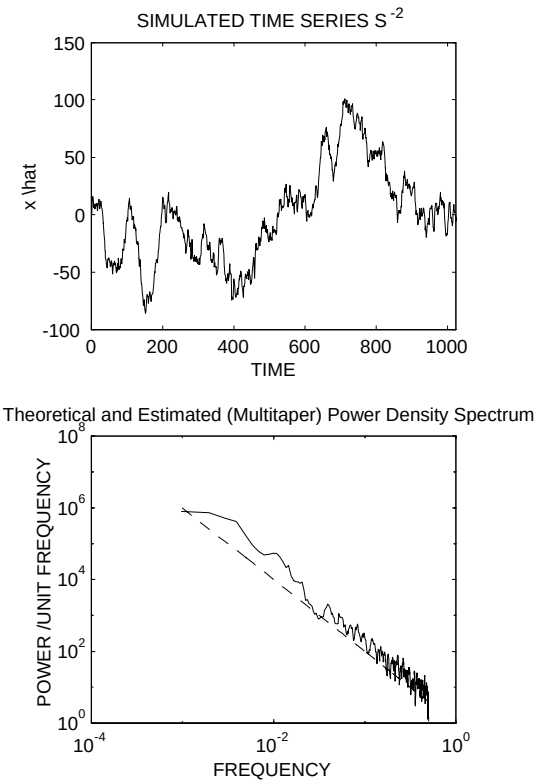

FIGURE 28. A simulated time series with power density proportional to s^{-2}. Although just a "red-noise" process, the eye seems to see oscillatory patterns that are however, ephemeral.

assure wide-sense stationarity. Confirmation of stationarity, and that the appropriate spectral density is reproduced can be simply obtained by considering the behavior of the autocovariance $< y_t y_q >$ (see Percival and Walden, 1993). A unit time step, $t = 0, 1, 2, ...$ is used, and the result is shown in Figure 28.

(This result is rigorously correct only asymptotically as $T \to \infty$, or for white noise. The reason is that for finite record lengths of strongly colored processes, the assumption that $\langle a_n a_{n'} \rangle = \langle b_n b_{n'} \rangle = 0$, etc., is correct only in the limit (e.g., Davenport and Root, 1958)).

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.

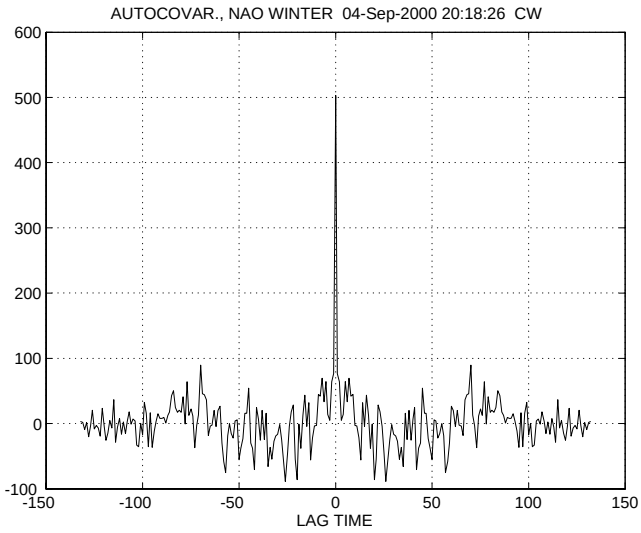

FIGURE 1. The estimated autocovariance of the North Atlantic Oscillation Index (NAO). Visually, and quantitatively, the autocovariance is dominated by the spike at the origin, and differs little from the autocovariance of a white noise process. By the Wiener filter theory, it is nearly unpredictable.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

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

Then we can make a forecast of x_m,

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

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

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

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, $\tilde{x}_m (-)$. A plausible idea is to *average the measurement with the forecast*, weighting the two inversely as their relative errors:

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

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

(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.

CHAPTER 3

Examples of Applications in Climate

A number of examples of the use of time series tools in the context of climate problems are found in the list below. These papers are available online as pdf files at *http://puddle.mit.edu/~cwunsch*

Wunsch, C., On sharp spectral lines in the climate record and the millennial peak. *Paleoceano.,* 15, 417-424, 2000.

Wunsch, C., The spectral description of climate change including the 100KY energy, *Clim. Dyn.,* DOI 10.1007/s00382-002-0279-z, 2002.

Wunsch, C. and D. E. Gunn, A densely sampled core and climate variable aliasing, *Geo-Mar. Letts.,* 23(1), DOI: 10.1007/s00367-003-0125-22003, 2003..

Huybers, P. and C. Wunsch, Rectification and precession signals in the climate system, *Geophys. Res. Letts,* 30, 19, 2011,doi:10.1029/2003GL017875, 2003

Wunsch, C., Greenland—Antarctic phase relations and millennial time-scale climate fluctuations in the Greenland cores, *Quaternary Sci. Revs.* 22, 1631-1646, 2003.

Wunsch, C. Quantitative estimate of the Milankovitch-forced contribution to observed climate change. *Quat. Sci. Revs.,* 23/9-10, 1001-1012, 2004.

1. References

Amos, D. E. and L. H. Koopmans 1962 *Tables of the Distribution of the Coefficient of Coherence for Stationary Bivariate Gaussian Processes.* Sandia Corp. Monograph, SCR-483, xx pp.

Bendat, J. S. and A. G. Piersol, 1986 *Random Data. Analysis and Measurement Procedures,* Second Edition, 566 pp., Wiley-Interscience, New York.

Bracewell, R. N. 1978 *The Fourier Transform and Its Applications.* 444 pp McGraw-Hill, New York.

Chapman, M. R. and N. J. Shackleton, 2000 Evidence of 550-year and 1000-year cyclicities in North Atlantic circulation patterns during the Holocene. *The Holocene,* 10, 287-291.

Claerbout. J. F. 1985 *Fundamentals of Geophysical Data Processing, with Applications to Petroleum Prospecting.* Blackwell, Palo Alto Ca., 274 pp.

Clemens, S. C. and R. Tiedemann 1997 Eccentricity forcing of Pliocene-Earth Pleistocene climate revealed in a marine oxygen-isotope record. *Nature*, 385, 801-804.

Cramér, H. 1946 *Mathematical Methods of Statistics.* 574 pp. Princeton U. Press, Princeton.

Davenport, W. B. Jr. and W. L. Root 1958 *An Introduction to the Theory of Random Signals and Noise.* 393 pp. McGraw-Hill, New York

Freeman, H. 1965 *Discrete-Time Systems. An Introduction to the Theory.* 241pp. John Wiley, New York.

Garrett, C. J. R. and W. H. Munk 1972 Space-time scales of internal waves. *Geophys. Fl. Dyn.*, 3, 225-264

Groves, G. W. and E. J. Hannan 1968 Time series regression of sea level on weather,. *Revs. Geophys.*, 6, 129-174.

Hamilton, J. D. 1994 *Time Series Analysis,* 793 pp. Princton Un. Press.

Hannan, E. J. 1970 *Multiple Time Series.* 536 pp. John Wiley, New York.

Hurrell, J. W., Decadal trends in the North Atlantic oscillation regional temperatures and precipitation. *Science,* 269, 676-679.

Jackson, D. D., 1975 *Classical Electrodynamics*, 2nd ed., 848 pp., John Wiley, New York.

Jenkins, G. M. and D. G. Watts 1968 *Spectral Analysis and Its Applications.* 525pp Holden-Day, San Francisco.

Jolliffe, I. T. 1986 *Principal Component Analysis.* 271 pp Springer-Verlag, New York.

Jury, E. I. 1964 *Theory and Application of the z-Transform Method.* 330.pp John Wiley, New York.

Körner, T. W. 1988 *Fourier Analysis.* 591pp Cambridge Un. Press, Cambridge.

Levinson, N. 1947 The Wiener RMS (root mean square) error criterion in filter design and prediction. *J. Math. Phys.*, 25, 261-278 (reprinted as an Appendix to Wiener, 1949).

Liebelt, P. B. 1967 *An Introduction to Optimal Estimation*, 273pp. Addison-Wesley, Reading, Mass.

Lighthill, M. J. 1958 *Fourier Analysis and Generalized Functions.* 79 pp Cambridge Un. Press.

McCoy, E. J., A. T. Walden, and D. B. Percival 1998. Multitaper spectral estimation of power law processes *IEEE Trans. Signal Processing*, 46, 655-668.

Moore, M. I. and P. J. Thomson 1991 Impact of jittered sampling on conventional spectral estimates. *J. Geophys. Res.*, 96, 18,519-18,526.

Munk, W. H. and G. J. F. MacDonald 1960 *The Rotation of the Earth: A Geophysical Discussion* 323pp., Cambridge University Press, Cambridge 323pp.

Parke, M. E., R. H. Stewart, D. L. Farless and D. E. Cartwright 1987 On the choice of orbits for an altimetric satellite to study ocean circulation and tides. *J. Geophys. Res.*, 92, 11,693-11,707.

Percival, D. B. and A. T. Walden 1993 *Spectral Analysis for Physical Applications. Multitaper and Conventional Univariate Techniques.* 583 pp Cambridge Un. Press, Cambridge.

Percival, D. B. and A. T. Walden, 2000 *Wavelet Methods for Time Series Analysis.* Cambriduge Un. Press, Cambridge 594 pp.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling 1992 *Numerical Recipes*, 2nd ed., Cambridge U. Press, Cambridge.

Priestley, M. B. 1982 *Spectral Analysis and Time Series. Volume 1: Univariate Series. Volume 2: Multivariate Series, Prediction and Control.* 890pp plus appendices (combined edition) Academic, London.

Stephenson, D. B., V. Pavan, and R. Bojariu 2000 Is the North Atlantic Oscillation a random walk? *Int. J. Climatol.*, **20**, 1-18.

von Storch, H. and F. W. Zwiers 1999. *Statistical Analysis in Climate Research.* Cambridge U. Press, Cambridge.

Thomson, P. J. and P. M. Robinson 1996 Estimation of second-order properties from jittered time series. *Annals Inst. Stat. Maths.*, 48, 29-48.

Tiedemann, R., M. Sarnthein, and N. J. Shackleton 1994. Astronomic time scale for the Pliocene Atlantic O1080 and dust flux records of Ocean Drilling Proram site 659. *Paleoceanog.*, 9, 619-638.

Tukey, J. W. 1984 Styles of spectrum analysis. in, *A Celebration in Geophysics and Oceanography - 1982. In Honor of Walter Munk*, SIO Reference Series 84-5, La Jolla Ca, 100-103.

Vautard, R. and M. Ghil 1989 Singular spectrum analysis in nonlinear dynamics with applications to paleoclimatic time-series, Physica D, 35, 395-424.

Wiener, N. 1949 *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, 163 pp The Technology Press of MIT and J. Wiley, New York.

Wunsch, C. and A. E. Gill 1976 Observations of equatorially trapped waves in Pacific sea level variations *Deep-Sea Res.*, 23,371-390.

Wunsch, C. and D. Stammer 1998 Satellite altimetry, the marine geoid and the oceanic general circulation. *Ann. Revs. Earth Plan. Scis.*, 26, 219-254.

Wunsch, C. 1991 Global-scale sea surface variability from combined altimetric and tide gauge measurements. *J. Geophys. Res.*, 96, 15,053-15,082.

Wunsch, C. 1996 *The Ocean Circulation Inverse Problem* . 437 pp. Cambridge University Press, Cambridge.

Wunsch, C. 1997 Wunsch, C. 1997 The vertical partition of oceanic horizontal kinetic energy *J. Phys. Oc.*, 27, 1770-1794.

Wunsch, C. 1999 The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Met. Soc.*, 80, 245-255.

Wunsch, C. 2000 On sharp spectral lines in the climate record and the millennial peak. *Paleoceanog.,* 5, 417-424.

Yaglom, A. M. 1962 *An Introduction to the Theory of Stationary Random Fields.* 235 pp. R. A. Silverman, translator Prentice-Hall, Englewood Cliffs, NJ.