GEOPHYSICAL DATA ANALYSIS

Robert L. Parker

CHAPTER 1: STOCHASTIC PROCESSES

1. Introducing Ordered Data

In the statistical portion of this class the observations is far have been almost entirely of the kind that has no special order associated with it. From now on we will study data in which the order is important. Usually we will call the sequence $x_1, x_2, x_3 \cdots\ x_n$, a **time series**; the index n denotes the time $t = n\,\Delta t$ at which the observation was made with respect to some initial time; obviously we are sampling here uniformly in time at an interval Δt. Obviously we could equally well be treating a data set sampled in space or, much more rarely, some quite different independent variable. For theoretical purposes it is useful to be able to treat time as continuous sometimes, and to work with $x(t)$, but here when we introduce the concept of a random variable for the observation, things can become mathematically extremely difficult without some further quite severe simplifications. For ideal models we will often want to consider infinite sequences, maybe infinite in both directions .

So a **stochastic process** $\{X_n\}$ or $\{X(t)\}$ is a family of random variables indexed by the integer n, when it is a discrete process, or by the real number t, when it is a continuous process. (Random variables or functions will normally be denoted by upper case letters, but the converse is not true.) As with ordinary random variables, X_1 for example, has no definite value, but is to be thought of as an infinite collection of values from which a particular experiment will extract a value. A given data series is conceptually the result of an experiment that could be repeated as many times as we like. We usually have just one **realization,** drawn from an **ensemble** of alternative series, all generated by the underlying process fully described by its **probability distribution function** (PDF). When we think of operations such as taking the average or **expected value** at a particular n or t, in general this is an average over different realizations for the same n or t. See Figure 1 for a picture.

Even in discrete time the general stochastic process is a horribly complex affair. To specify it completely requires the joint PDF of every element X_n with every other element. The simplest case is the Gaussian model, which you already have met. Given N data we have the joint PDF is in the form

$$\phi(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{1}{2}N}\det(C)} \exp\left[-\tfrac{1}{2}(\mathbf{X}-\bar{\mathbf{x}})^T C^{-1}(\mathbf{X}-\bar{\mathbf{x}})\right] \tag{1.1}$$

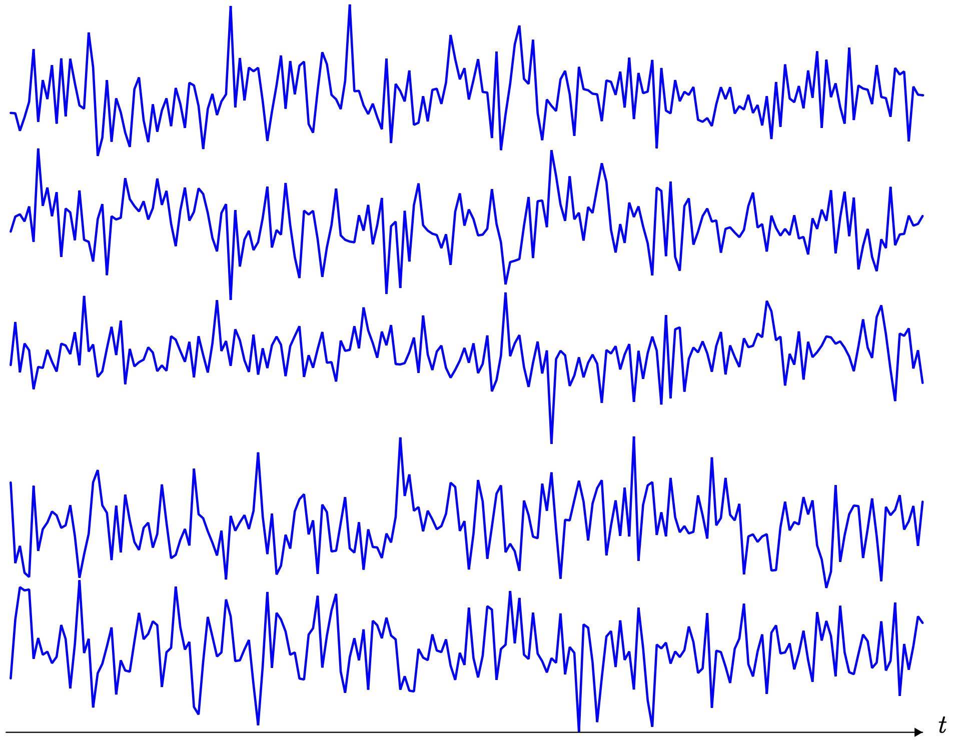
where $\mathbf{X} = (X_1, X_2, \cdots X_N)^T$, $\bar{\mathbf{x}} \in \mathrm{I\!R}^N$ is vector of mean values, and $C \in \mathrm{I\!R}^{N\times N}$ is the covariance matrix, a symmetric, positive definite matrix that describes the correlations among the random variables X_n. The reason this is complicated is that to describe completely these N random variables, we need a total of ½$N(N+3)$ parameters, far more than we are likely to have observations: if we have even a short time series of say 50 numbers we would need 1325 parameters; for a reasonably long data series of 2000 members, more than 2 million parameters. This is for the Gaussian distribution, the simplest kind of PDF. Now imagine what this means if we consider continuous time instead; the problem is almost completely intractable, and such a general treatment has no practical value because it would be impossible to make estimates of the necessary parameters, even if the mathematics were possible, which turns out to be very hard for the general case. See *Priestley,* Chapter 3 for a further discussion.

2. Stationary Processes and Autocovariance

The perfectly general stochastic process is too general to be useful, and so the conventional wisdom is to focus on a much more restrictive class of random processes called **stationary processes.** The idea here is that, while the actual observables vary in time, the *underlying statistical description is time invariant.* This can be weakened a bit, but we will consider only these so-called **completely stationary** processes. This has the gratifying effect of reducing the number of parameters needed to a manageable number. For example, the mean value of a stationary process:

$$\mathcal{E}\,[X_n] = \bar{x} \tag{2.1}$$

Figure 1: Five realizations of the same stochastic process.



is a constant independent of n, or of time t if the process is continuous. It is often assumed the mean is zero, since it is a trivial operation to add the mean value back into the data series, if necessary. Of course, many observational series do not look as if the mean is constant – there may be a **secular trend.** Stationarity is such a powerful and useful property that one often attempts to convert an evidently nonstationary series into a stationary process, for example, by fitting a straight line trend, or forming a new stationary sequence by differencing:

$$Y_n = X_{n+1} - X_n \tag{2.2}$$

Recall from the variance of single random variable the definition

$$\sigma_X^2 = \mathrm{var}\,[X_n] = \mathcal{E}\,[(X_n - \bar{x})^2]\,. \tag{2.3}$$

With stationarity this number must also be independent of n (or t). And the covariance between any two random variables in the sequence cannot depend on where we are in the series and therefore

$$\mathrm{cov}\,[X_m,\, X_n] = \mathcal{E}\,[(X_m - \bar{x})(X_n - \bar{x})] = R_X(m-n) \tag{2.4}$$

that is, a function R_X of the interval between the two points. The function R_X is called the **autocovariance.** For continuous processes this is usually written

$$\mathrm{cov}\,[X(t),\, X(t+\tau)] = R_X(\tau)\,. \tag{2.5}$$

Then τ is called the **lag**. Observe that by definition

$$R_X(0) = \sigma_X^2\,. \tag{2.6}$$

Also notice that, because of stationarity, one can set $s = t - \tau$ in (2.5) and the result will be the same, since the answer is independent of which time was selected. This yields:

$$R_X(-\tau) = R_X(\tau) \tag{2.7}$$

which shows that the autocovariance function is an even function of the lag. We see then that a stationary process does not contain information on which way time is flowing – it is the same process if time is reversed.

Returning to a stochastic process with a Gaussian PDF as in (1.1), we see that in place of the vector $\bar{\mathbf{x}}$ of mean values we have a single number. How about the covariance matrix? You may recall that the j-kth entry of C is

$$C_{jk} = \mathrm{cov}\,[X_j,\, X_k] = R_X(j-k)\,. \tag{2.8}$$

Hence the covariance matrix has the same values on all its diagonals

$$C = \begin{bmatrix} R_X(0) & R_X(1) & R_X(2) & R_X(3) & \cdots \\ R_X(1) & R_X(0) & R_X(1) & R_X(2) & \cdots \\ R_X(2) & R_X(1) & R_X(0) & R_X(1) & \cdots \\ R_X(3) & R_X(2) & R_X(1) & R_X(0) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{2.9}$$

This type of matrix is called a **Toeplitz** matrix. Now instead of ½$N(N+3)$, there are only $N+1$ parameters in the Gaussian stationary PDF.

Just because the stochastic process is composed of random variables does not mean it is completely unpredictable. Recall the correlation coefficient of two random variables:

$$\rho_{XY} = \frac{\mathrm{cov}\,[X, Y]}{\sqrt{\mathrm{var}\,[X]\,\mathrm{var}\,[Y]}}\,. \tag{2.10}$$

Then from (2.5) we have the correlation coefficient between any two points in a sequence in continuous time is

$$\rho(\tau) = \frac{R_X(\tau)}{\sigma_X^2} \tag{2.11}$$

and a similar result for discrete processes. This function is called the **autocorrelation function.** Thus, unless the autocovariance function is exactly zero for the lag τ, one can predict something about the value further along in the sequence at $X(t+\tau)$ from value at t, because they are correlated.

So far we have considered the average of the process, and averages of the products $(X(t_1)-\bar{x})(X(t_2)-\bar{x})$. Such averages are examples of **moments** of the process. The second order moments concern the variance and autocovariance function, and nothing else. Higher order moments, those involving the product of three or more Xs can obviously be defined, but if the process is based on a Gaussian PDF, all the information about the PDF is contained in the second order moments and the higher moments can be predicted. Even if the PDF is not Gaussian a great deal can be learned about a process from its second order moments, and so the higher order moments are rarely investigated. We will follow this path.

3. White Noises and their Relatives

Let us look a few concrete examples of stationary stochastic processes. The simplest (but artificial) example is that of **white noise**. This is defined as a stationary process in which the random variable at any point is independent of every other variable. It is also common to assume the mean value of white noise is zero. In the discrete case we have

$$R_W(n) = \sigma_0^2\, \delta_{n0} \tag{3.1}$$

where δ_{jk} is the Kroenecker delta symbol. Hence in this case the random process is unpredictable to the extent that we learn nothing about the next or subsequent values from the current one. The results are still not completely unpredictable, because we can say something about the range of values to be expected on account of the known variance, σ_0^2.

For continuous processes, it turns out that a white noise is singular because the variance at any point must be infinite:

$$R_W(\tau) = s^2\, \delta(\tau) \tag{3.2}$$

But these definitions do **not** completely specify the stochastic process. At any particular time t (or index n) X is a random variable with a PDF; that PDF will be the same for every t, but so far we have no specified it. Of course, the most common choice is the Gaussian, so that

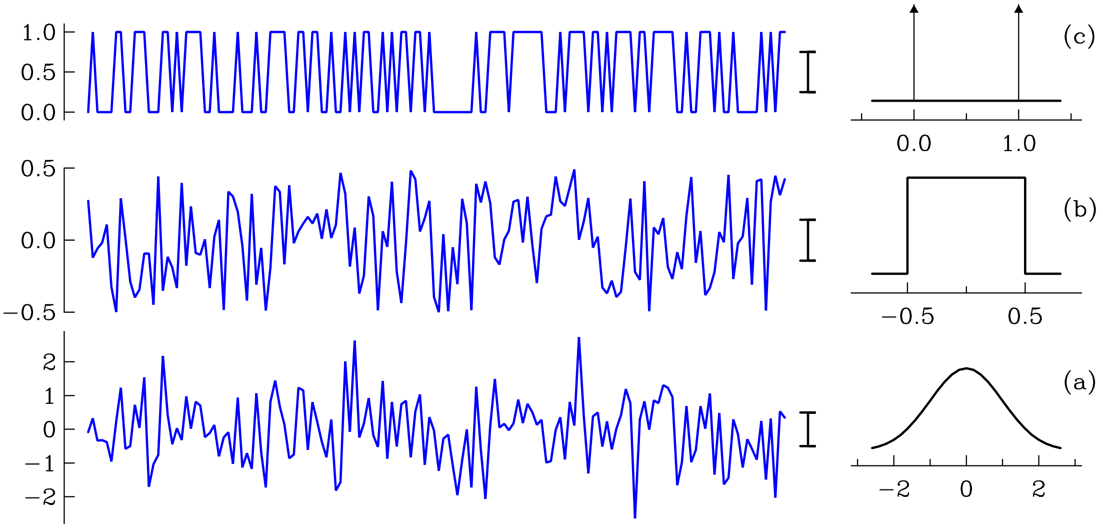
$$\phi(X) = \frac{1}{\sigma_0\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(X/\sigma_0)^2}\,. \tag{3.3}$$

Because of the statistical independence, the joint PDF is

$$\psi(X_1, X_2, X_3, \cdots) = \phi(X_1)\cdot\phi(X_2)\cdot\phi(X_3)\cdot\cdots. \tag{3.4}$$

This is of course **Gaussian white noise;** shown by in Figure 2a.

Figure 2: Three white noises; the vertical bar is 1 σ long.



We allow any other PDF for X_n; suppose we choose a uniform distribution:

$$\phi(X) = b^{-1}\,\mathrm{box}\,(X/b) \tag{3.5}$$

where now $\mathrm{var}\,[X] = b^2/12$. This is a different kind of white noise, easily seen in Figure 2b. The joint PDF is given by (3.4) again, with the appropriate choice of ϕ. This kind of white noise turns up quite frequently in real measurements: **round-off noise.** Suppose a discrete time series is recorded and the number is rounded, say to 1 decimal. If the series varies by much more than ± 0.1 over each Δt, the value recorded is the true value of the measurement plus an unpredictable (ie random) amount that lies in $(-0.05, +0.05)$ with a uniform distribution. Thus the recorded series appears to be the true signal with a white noise added, since the consecutive values of the rounding error are uncorrelated. The noise is zero mean and its variance is in this case $0.1^2/12$. If the last significant digit doesn't change very often in the series, then the round-off noise added is no longer uncorrelated; but then it seems likely the signal is not being recorded with enough accuracy.

Another random white series with a limited range is the **random telegraph signal:** this one switches discontinuously between only two values, say, 0 and 1:

$$\phi(X) = \tfrac{1}{2}[\delta(X) + \delta(X-1)]\,. \tag{3.6}$$

Here the mean value is not zero but one half, and the variance is a quarter. The random telegraph signal is used as a calibration signal for seismometers or other instruments since it is easy to generate electronically. A zero mean version is sometimes suggested as a model for the Earth's dipole moment over time scales of 10^5 to 10^6 years, but this is actually very implausible because the moment is far from constant between reversals. See Figure 2c for a picture.

These three examples of white noise are clearly different to the eye. Somewhat remarkably to me at least, if they are normalized to the same variance and converted into a sound track, they sound identical – the ear doesn't have a very good density function discriminator.

These three sequences were obviously not observational; they were made with a random number generator in MATLAB. We can obtain other kinds of stochastic sequences by filtering white noise in various ways; the series in Figure 1 were generated that way. If one filters white noise, the result tends to be Gaussian in distribution as a consequence of the Central Limit Theorem. Many physical processes can be thought of as resulting from this kind of process, so Gaussian distributions are often assumed in a signal, and are quite often observed too, but not always. A common exception appears to be traces of marine magnetic anomalies, which are usually heavy in the tails compared with a Gaussian distribution.

Let us briefly consider the simplest kind of filter, a **FIR:** this stands for a **Finite Impulse Response filter** (aka MA or **Moving Average**). For a FIR filter we convolve the white noise sequence W_j with a finite number of weights w_k:

$$Y_n = \sum_{k=1}^{K} w_k W_{n-k} \,. \tag{3.7}$$

Then we can calculate the autocovariance of the new sequence using the definition (2.4); let's assume for simplicity the mean is zero. Then

$$R_Y(l) = \mathrm{cov}\,[Y_{n+l}, Y_n] = \mathcal{E}\,[Y_{n+l} Y_n] \tag{3.8}$$

$$= \mathcal{E}\,[\sum_{j=1}^{K} w_j W_{n+l-j} \sum_{k=1}^{K} w_k W_{n-k}] \tag{3.9}$$

$$= \sum_{j=1}^{K} \sum_{k=1}^{K} w_j w_k \, \mathcal{E}\,[W_{n+l-j}\, W_{n-k}] \tag{3.10}$$

$$= \sum_{j=1}^{K} \sum_{k=1}^{K} w_j w_k \, \sigma_0^2 \delta_{l-j+k,0} \quad . \tag{3.11}$$

The delta symbol vanishes except when $l - j + k = 0$, namely when $j = l + k$; so

$$R_Y(l) = \sigma_0^2 \sum_{k=1}^{K} w_k \, w_{k+l} = \sigma_0^2 \, w_k * w_{-k} \,. \tag{3.12}$$

Of course we see that the previously uncorrelated (zero covariance) white noise W_k has become correlated. Notice that R_Y does go to zero once $| l | > K$.

The student is invited to verify that the same result is obtained for the continuous time version: if

$$Y = w * W \tag{3.13}$$

$$R_Y = s^2 \, w(t) * w(-t) \tag{3.14}$$

A lot of space is wasted (in my opinion) in books (eg, Priestley) applying various recursive filters (AR, ARMA filters) to white noise and looking at the consequences.

Let us move on to some examples of actual observations that might be realizations of stationary processes.

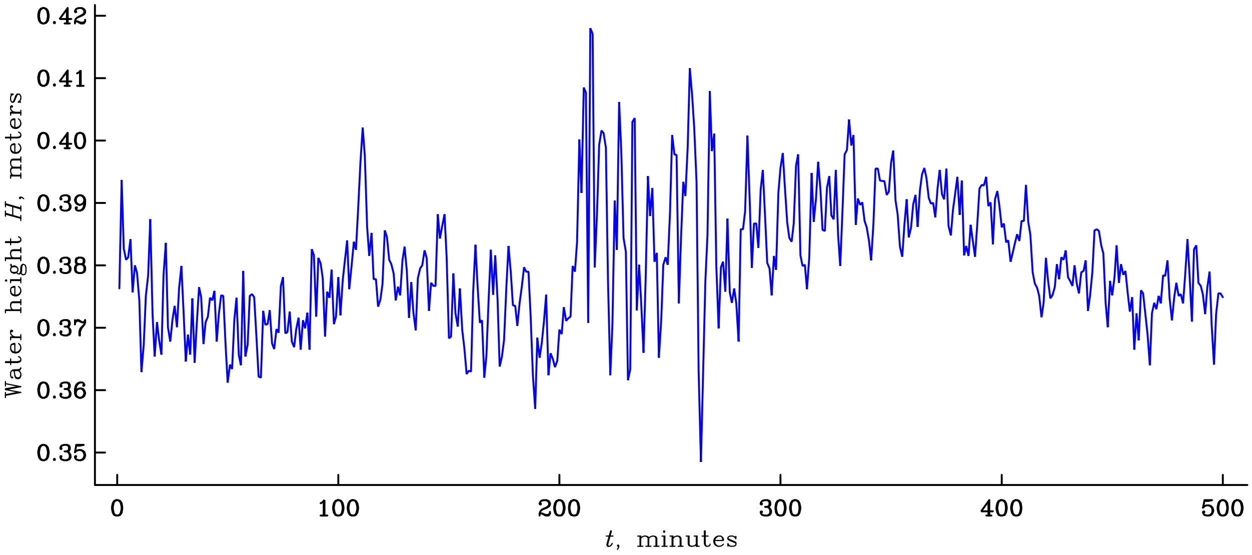
4. Examples from the Real World

Our first example is a recording of the water height at the edge of a lake over a period of several hours. The water is rising and falling in response to the wind, but because of standing waves in the basin, there are well-defined oscillations in time, known as a **seiche.** Below in Figure 3 we see a record of nearly 10 hours; I have data for the best part of a full day, 1301 minutes. First observe, the mean is clearly not zero. Next notice the oscillations, which are not completely regular, but none-the-less there is a suggestion of periodicity. This definitely looks like a stochastic process, but one might be skeptical that it is stationary, given the amplitude increase at about 210 minutes. None-the-less we will stick with that model because it is so useful.

We can ask if this looks like a good approximation to a white noise. Then values at one time would not be related, even on average, to values at other times. That looks improbable to the eye. We can also draw a scatter plot, for example, as shown in Figure 4 where the observed value at one time is plotted against the height 8 minutes later. A very clear correlation is visible. I estimate the correlation coefficient by (2.10) to be 0.626. In the sample there were 1293 data, and we can calculate the probability that ρ would be this big in an uncorrelated Gaussian sample by the t test (see Rice's book for details): we find $t_{N-2} = 28.8$, and the chances of random data exceeding this value are less than 10^{-180}.

How good is the Gaussian model? Ignoring the fact this is a sequence, and just treating the values as an ordered set, we can find a histogram, or better, the cumulative distribution function. This can be tested against the Gaussian model with the Kolmogorov-Smirnov test. Figure 5 shows the quantile-quantile plot on the Gaussian hypothesis: the fit is remarkably good. From the value if the D_N statistic we calculate

Figure 3: Water height on the shore of the Salton Sea; a seiche.



that the probability of a random sample exceeding the observed value is 0.83, so the fit to the Gaussian model is excellent, maybe even slightly too good.

Figure 4: Scatter plot of data in Figure 3 with a lag of 8 minutes.

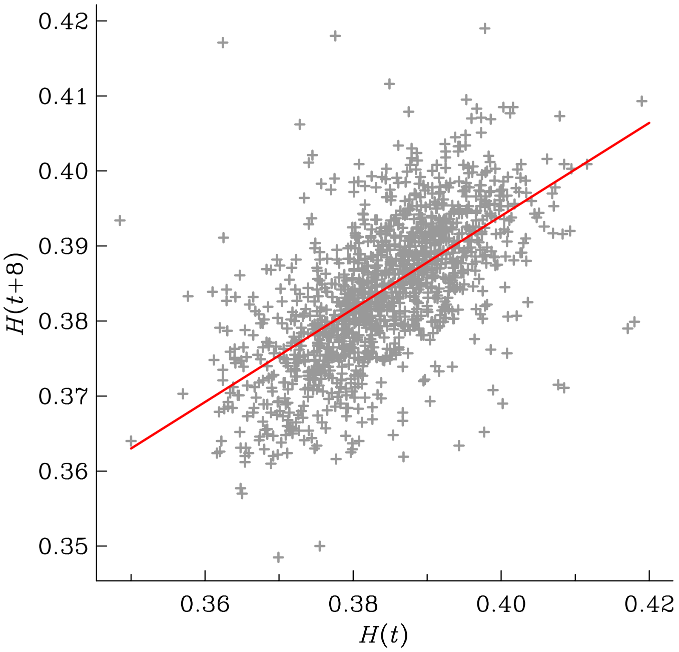
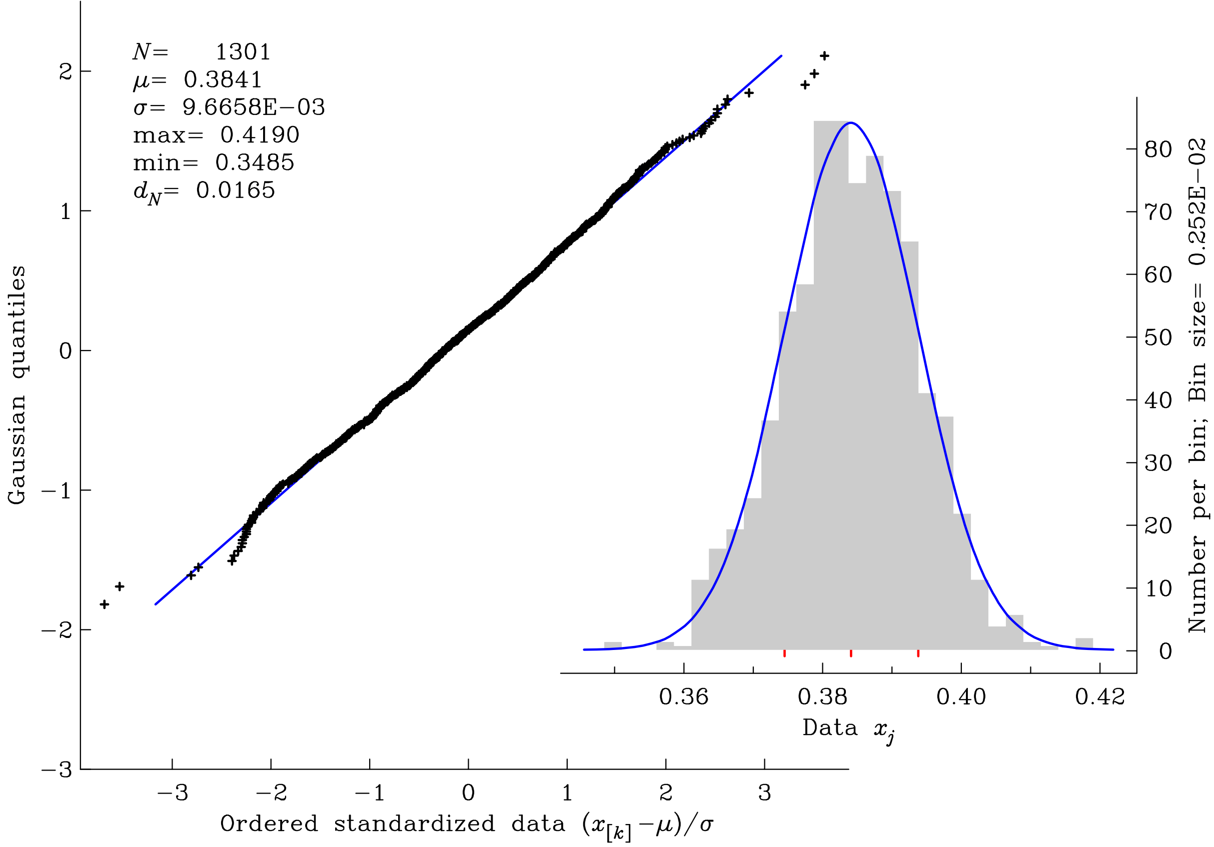


Figure 5: Q-Q plot and density function of seiche data.



The second geophysical data sequence is a series of values in space, the magnetic field measured on by a high flying aircraft (7,000 meters altitude) over the south-eastern Pacific Ocean. In Figure 6 I show only the vertical component Z, and the horizontal, along flight-path component X. Both components are plotted after removal of a standard geomagnetic model, so that their mean values should be nearly zero. The spacing of the sampling is uniform at 3.5 km. Again, as with the lake level data, we see an irregular line with a certain amount of order. A stochastic process seems like a good model, but here there seems to be little evidence of a regular oscillation, or even an irregular one. A feature to notice here is that the two components appear to be related, varying together in some not very obvious fashion: there is a phase lag and perhaps a suggestion of differentiation of Z to obtain X. We will discuss later how to look at pairs of time series for evidence of common variability of this kind.

Concentrating for the moment on the Z component, let us look at an estimate of the autocorrelation function, shown in Figure 10. I will not describe yet how that estimate was made because that will be the subject of a later lecture. For now notice how the R_Z dies away monotonically from one. This means that neighboring values are likely to be very similar indeed, but as one separates two samples in space, their correlation fades away so that by a lag of 30 samples, they are uncorrelated. It is easy to believe this series could be generated by smoothing white noise, that is, by applying a suitable FIR filter to white noise.

The Q-Q plot shows something I mentioned earlier. The magnetic anomaly cumulative plot does not follow the Gaussian model very well: it is rather asymmetric with a large positive tail and a compressed lower tail. The K-S test says that random Gaussian variables would generate such a large value for d_N only 18 percent of the time. This is not a resounding rejection of the Gaussian model, but tells us we should be suspicious of it. The reason for this commonly observed behavior is not understood.

Figure 6: Magnetic anomaly profile over the eastern Pacific Ocean.

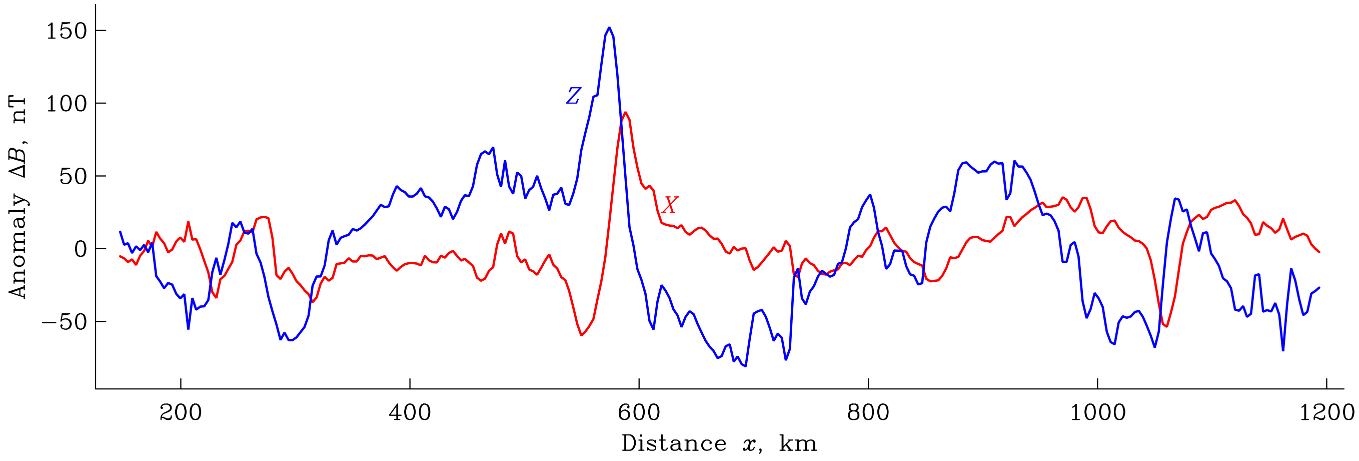


Figure 7: Estimated autocorrelation function for magnetic component Z.

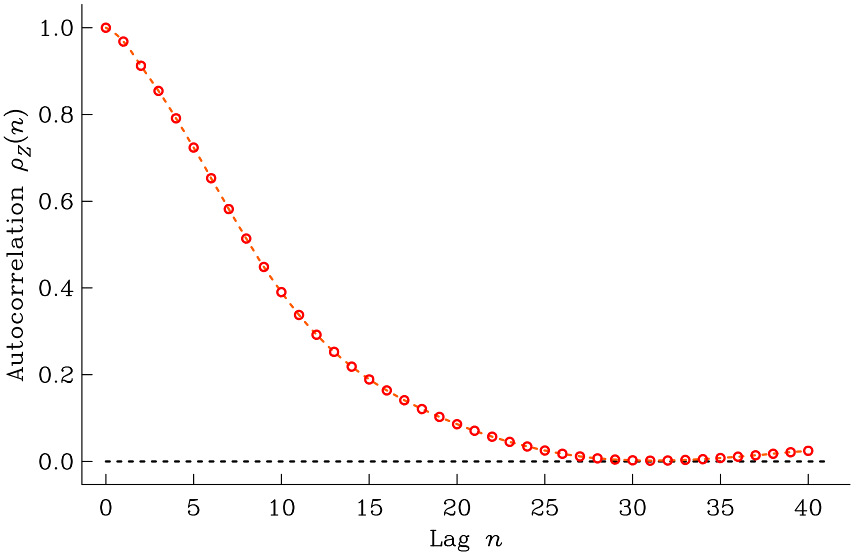
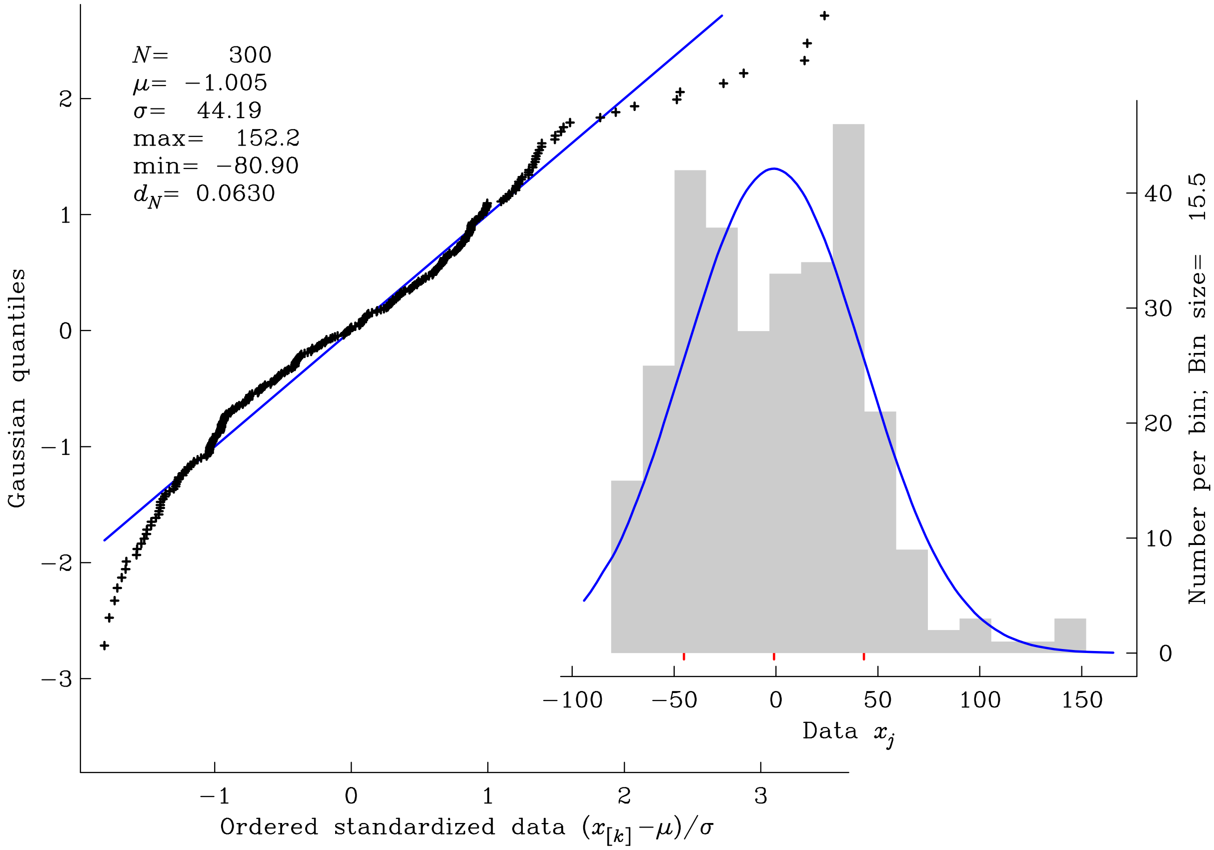
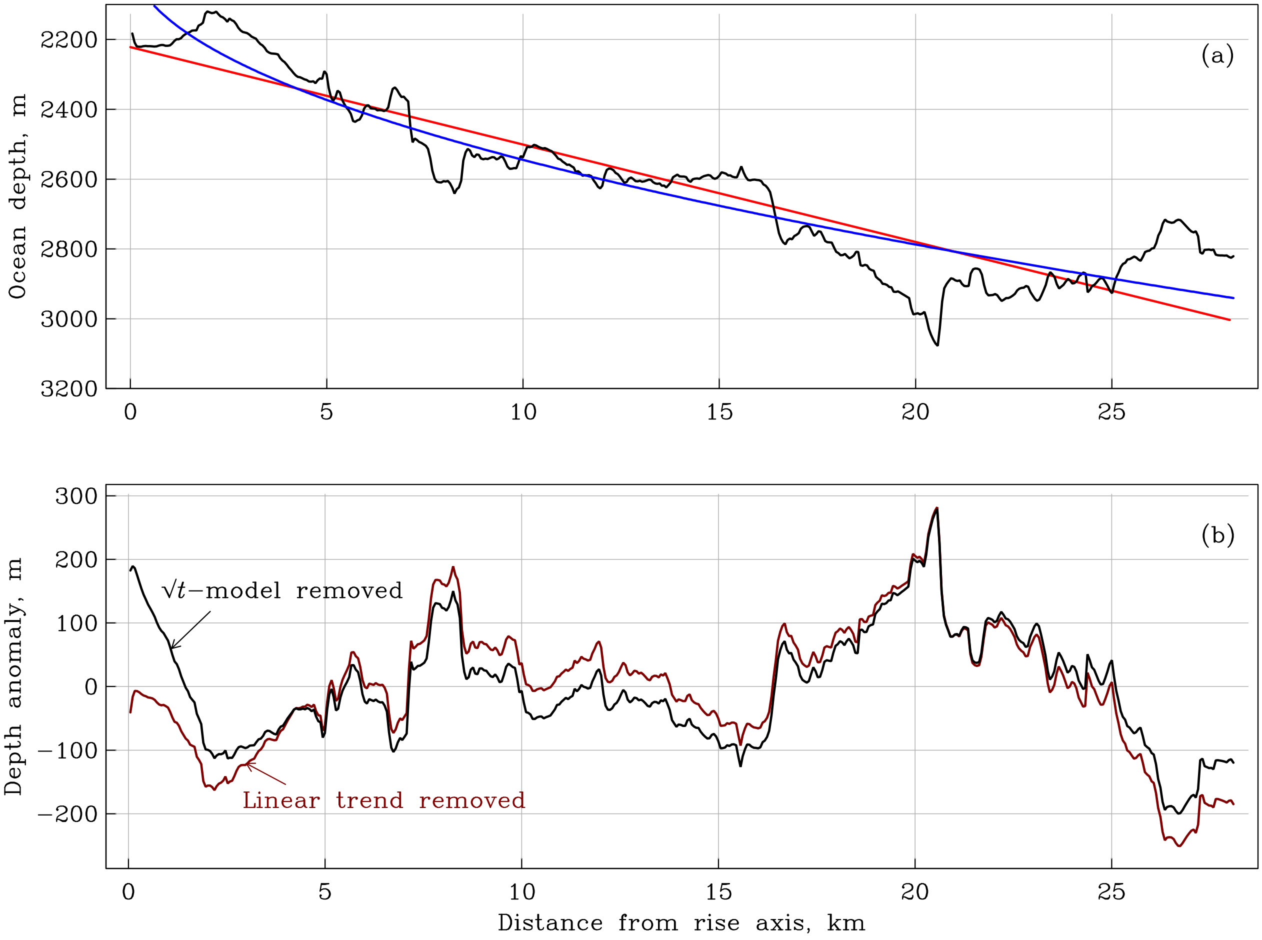


Figure 8: Q-Q plot and density function of Z data.



As a final example I show in Figure 9(a) a bathymetry profile across the East Pacific Rise, an example of a spatial series, which could be a true time series if I had calculated the age of seafloor from the spreading rate. Here the depth signal is very obviously not stationary, because as we expect from marine geology, the seafloor becomes deeper with age. As a realization of a stationary process, the depth curve is a miserable failure because the mean value is not independent of t. But if we remove an average trend, (the line shown), we get a much more satisfactory-looking approximation to a stationary series, as we see from the gray line in Figure 9(b). Now the geologists among the group will, know I should not have removed a straight line but a parabola, because of the famous $\sqrt{t}$ approximation. The least-squares best-fitting age curve is shown also, and it fits the observations slightly better than the straight line. The residual is the black curve in Figure 9(b).

Figure 9: EPR bathymetry, and a depth anomaly.



What we have done is to model the observations as a steadily evolving part, plus a random, stationary part. Unusually in this example, we have a good model for the evolving piece of the model; normally we would just take a straight-line trend to represent that, and as you can see in this case the difference is not very large.

The statistics here are not quite Gaussian again – the distribution is heavy tailed, but not at a high level of significance according to K-S.

GEOPHYSICAL DATA ANALYSIS

Robert L. Parker

CHAPTER 2: SPECTRAL ANALYSIS OF STOCHASTIC PROCESSES

1. Spectral Analysis

At this point in the class it should not be a surprise that we will introduce a decomposition based on frequencies, a **spectral analysis.** Many things are simpler when looked at through these glasses – any time-invariant system, and solutions of differential equations with constant coefficients, have already been beaten to death with filters and Fourier transforms. As I mentioned elsewhere, most physical processes have behavior that can be characterized by the frequency of the variation, because which set of physical laws provide the dominant approximation depends on frequency. Recall my example about the magnetic field at a point on the Earth's surface: at frequencies of one over 1 million years (3×10^{-14} Hz) the physics is that of the slow moving fluids of the core dynamo; at 1 over a year (3×10^{-8} Hz) the field is governed by the behavior of the solar wind; at 1×10^{8} Hz the magnetic field most likely originates from a local TV station. Given a natural process with a random appearance, one naturally would like to decompose it into components of various frequencies, in effect, to take its Fourier transform.

To create a viable theoretical basis we need to consider stationary stochastic processes, which are random functions extending throughout all time with time-invariant properties. For functions of a real variable (continuous-time signals) we have the classical Fourier transform:

$$\hat{x}(f) = \int_{-\infty}^{\infty} x(t)\, \mathrm{e}^{-2\pi \mathrm{i} f t}\, dt\,. \qquad (1.1)$$

For sequences (discrete-time signals) we have the spectrum on the finite interval (–½, ½):

$$\hat{x}(f) = \sum_{n=-\infty}^{\infty} x_n\, \mathrm{e}^{-2\pi \mathrm{i} f n},\ -½ \le f \le ½ \qquad (1.2)$$

which allows reconstruction by the spectral representation:

$$x_n = \int_{-½}^{½} \hat{x}(f)\, \mathrm{e}^{2\pi \mathrm{i} f n}\, df\,. \qquad (1.3)$$

Can we apply either of these to their corresponding stochastic processes? The answer is no, because to make sense, the integral in (1.1) or the sum in (1.2) must converge, which requires some kind of decrease in amplitude, or energy, as t or n gets large. For a stationary process, that does not happen.

Also, if we put a stochastic process in $f(t)$ in (1.1) we would obtain another random function. Our goal is to characterize $X(t)$ with an ordinary function describing its properties in frequency (as the autocorrelation function does in time) not generate another random process. We can see in an intuitive way what we require, as follows. Suppose we wanted to know how much variability the stationary process exhibits at a frequency f_0. We could build (or design) a narrow band-pass filter ϕ_{f_0} that only allowed signals through in the frequency band $(f_0 - ½df,\ f_0 + ½df)$ with unit gain. Now send the signal X through that filter, and out would come a stochastic process which would have a very limited frequency content. We measure its amplitude by the variance; we would expect the variance to proportional to the width df of the band-pass filter. Then define

$$S_X(f_0)\, df = \mathrm{var}\,[\phi_{f_0} * X] \tag{1.4}$$

that is the variance of the filtered process will be some positive value times df; it will vary as the center frequency f_0 is varied, and be proportional to the variance in X at that frequency. The variance of X in a frequency band is called the **power** in that band and so S_X is the **Power Spectrum** of X, or more grandly its **Power Spectral Density.** Equation (1.4) is our informal definition of $S_X(f_0)$. Notice this definition works equally well for continuous or discrete processes. In the days before computers, analog spectral analysers were built based on this principle: a large number of narrow band-pass filters followed by rectifiers to measure the variance in each band.

2. Two Definitions of the PSD

We begin with a strict definition of the **Power Spectral Density** (PSD) of a stationary process as a kind of Fourier Transform. Let us study the definition for a continuous stationary process. In what follows we will always assume the process $X(t)$ has **zero mean.** Two problems with the ordinary FT were noted in Section 1: (a) the FT of X would not be defined on the infinite interval, and (b) the answer would be a random process, not a statistic of X. We fix these two things as follows: first define $X_T(t)$ as the process $X(t)$ on the finite interval $(-T, T)$:

$$X_T(t) = \begin{cases} X(t), & -T \le t \le T \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Any particular realization of this process (which is not stationary) has bounded 2-norm and thus has an ordinary FT:

$$\hat{X}_T(f) = \mathcal{F}\,[X_T](f) = \int_{-\infty}^{\infty} X_T(t)\, \mathrm{e}^{-2\pi i f t}\, dt = \int_{-T}^{T} X_T(t)\, \mathrm{e}^{-2\pi i f t}\, dt\,. \tag{2.2}$$

Note that $\hat{X}_T$ is still a random function of f, however. So we find its magnitude, square it, and take the expected value: $\mathcal{E}\,[\,|\hat{X}_T(f)|^2]$. Our plan is to let T tend to infinity, but we can easily see that this number would grow to finity, and so we divide by the interval length $2T$ to tame the growth: we define the function of frequency

$$S_X(f) = \lim_{T\to\infty} \frac{1}{2T} \mathcal{E}\,[\,|\,\hat{X}_T(f)\,|^2] \tag{2.3}$$

$$= \lim_{T\to\infty} \mathcal{E}\left[\frac{1}{2T}\left|\int_{-T}^{T} X_T(t)\,\mathrm{e}^{-2\pi i f t}\,dt\right|^2\right]. \tag{2.4}$$

This is the definition of the PSD; it can be shown to exist for all stationary processes X with zero mean and a bounded variance. It is obviously real and non-negative.

Equation (2.4) defines what is often called the **two-sided** PSD, because we allow f to run from $-\infty$ to ∞. When X is real, the usual case, it is easily seen from the well-known properties of the FT that $S_X(f)$ is an even function of f and therefore only values for $f \geq 0$ need be specified. Quite commonly a **one-sided PSD** is used, given by $2S_X(f)$ for $f \geq 0$; we will see in a moment why it is convenient to put in the factor of two.

Looking at (2.4) we can observe that S_X at any particular f is obtained as some kind of second order moment of X – only products of X with itself are needed, no third order moments enter. We have already introduced another second-order moment of X, the autocovariance. Does S_X provide independent information about X, or is there a connection between R_X and S_X? Somewhat surprisingly, to me anyhow, is the following answer: the functions $R_X(t)$ and $S_X(f)$ *contain exactly the same information.* In fact, S_X is the FT of R_X

$$S_X(f) = \mathcal{F}\,[R_X] = \int_{-\infty}^{\infty} R_X(t)\,\mathrm{e}^{-2\pi i f t}\,dt \tag{2.5}$$

Equation (2.5) is sometimes used as an alternative the definition of the PSD.

Before we establish the truth of (2.5) we observe a few consequences. Since R_X is a real even function of t, (2.5) implies that S_X is a real and even in f. But the fact that S_X must be non-negative puts severe restrictions on what functions R_X are allowed to be autocovariances; clearly not every even R_X with an FT is going to have a positive FT. Now take the inverse transform of (2.5):

$$R_X(t) = \int_{-\infty}^{\infty} S_X(f)\,\mathrm{e}^{2\pi i f t}\,df\,. \tag{2.6}$$

Now recall from the definition of R_X that

$$R_X(0) = \mathcal{E}\,[X(t)\,X(t)] = \mathrm{var}\,[X] = \sigma_X^2 \tag{2.7}$$

remembering that X is a zero-mean process. Setting $t = 0$ in (2.5) gives the **important result:**

$$\sigma_X^2 = \int_{-\infty}^{\infty} S_X(f)\, df\ . \tag{2.8}$$

In words, *the area under the power spectrum is the process variance.* That is why we double S_X if we use the one-sided PSD, to preserve this property.

Now we verify (2.5). We start with the squared magnitude of the FT of X_T:

$$|\,\hat{X}_T\,|^2 = \hat{X}_T\, \hat{X}_T^{\,*} \tag{2.9}$$

Now recall that the FT of a convolution is the product of the FTs; further notice that, since X_T is real

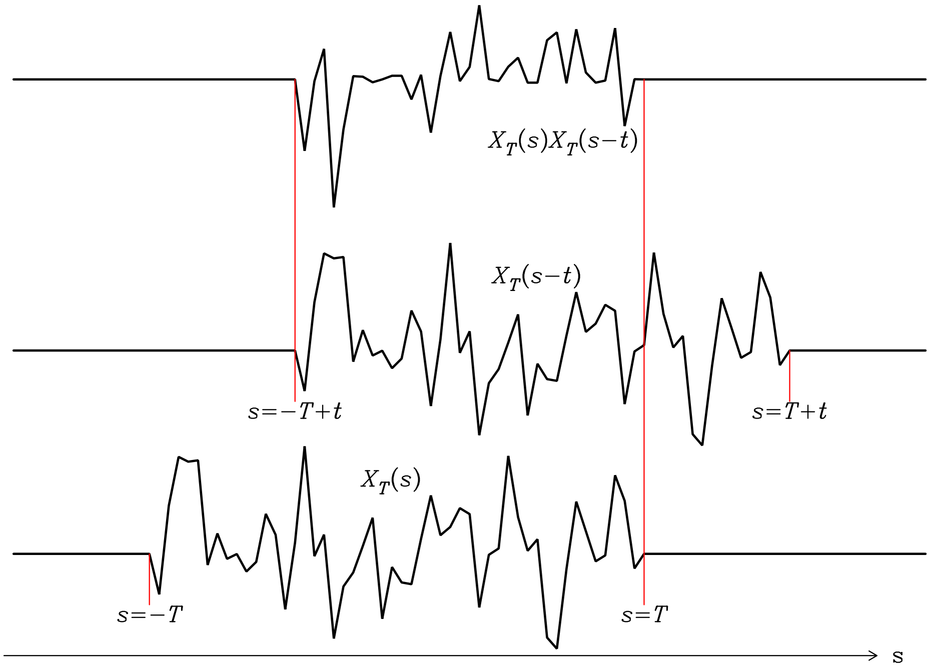
$$\hat{X}_T(f)^* = \int_{-\infty}^{\infty} X_T(t)\, e^{2\pi ift}\, dt = \int_{-\infty}^{\infty} X_T(-t) e^{-2\pi ift}\, dt = \mathcal{F}\ [X_T(-t)] \tag{2.10}$$

Combining the Convolution Theorem with (2.10) and (2.9), we have

$$|\,\hat{X}_T\,|^2 = \mathcal{F}\ [X_T(t) * X_T(-t)] = \mathcal{F}\ [\int_{-\infty}^{\infty} X_T(s)\, X_T(s-t)\, ds] \tag{2.11}$$

In (2.3) we have normalized by the interval $2T$. So let us put that into the

Figure 1: The integrand of (2.12).



definition of another function:

$$R_T(t) = \frac{1}{2T} X_T(t) * X_T(-t) = \frac{1}{2T} \int_{-\infty}^{\infty} X_T(s)\, X_T(s-t)\, ds \tag{2.12}$$

Then by (2.11)

$$\frac{|\hat{X}_T|^2}{2T} = \mathcal{F}\,[R_T] = \int_{-\infty}^{\infty} e^{-2\pi i f t}\, R_T(t)\, dt \tag{2.13}$$

Our definition of the PSD is (2.3); let us plug (2.13) into that

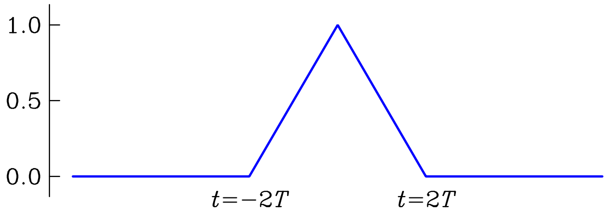
$$S_X(f) = \lim_{T\to\infty} \mathcal{E}\left[\frac{|\hat{X}_T|^2}{2T}\right] = \lim_{T\to\infty} \int_{-\infty}^{\infty} e^{-2\pi i f t} \mathcal{E}\,[R_T(t)]\, dt \tag{2.14}$$

From (2.12) we see that $R_T(t)$ is even in t and so we can always write $R_T(t) = R_T(|t|)$; in the following we will assume $t \geq 0$ and then replace t by $|t|$ at the end. We know $X_T(s)$ vanishes outside the interval $(-T, T)$ and therefore the integrand of (2.12) must vanish when $s > T$ or when $|s-t| > T$; see Figure 1. Therefore we can reduce the interval of integration in (2.12) to be on $(-T+t,\, T)$ instead of the whole real line. Also observe that once $t > 2T$ the nonzero sections cease to overlap, and the integrand is identically zero. These considerations lead to

$$R_T(t) = \begin{cases} \dfrac{1}{2T} \displaystyle\int_{-T+t}^{T} X_T(s)\, X_T(s-t)\, ds, & 0 \leq t < 2T \\ 0, & t \geq 2T\,. \end{cases} \tag{2.15}$$

Further simplifications ensue when we take the expected value, as dictated by (2.14); for the segment $0 \leq t < 2T$

Figure 2: The function $\Lambda_T(t)$.



$$\mathcal{E}\,[R_T(t)] = \frac{1}{2T}\int_{-T+t}^{T} \mathcal{E}\,[X_T(s)\,X_T(s-t)]\,ds = \frac{1}{2T}\int_{-T+t}^{T} R_X(-t)\,ds \tag{2.16}$$

where we have introduced R_X, the autocovariance of the process. Since $R_X(-t) = R_X(t)$, which is independent of s, we can evaluate s the integral explicitly:

$$\mathcal{E}\,[R_T(t)] = \frac{R_X(t)}{2T}\int_{-T+t}^{T} 1\cdot ds = R_X(t)\left[1 - \frac{t}{2T}\right],\quad 0 \le t \le 2T \tag{2.17}$$

From (2.15) $\mathcal{E}\,[R_T(t)] = 0$ when $t \ge 2T$. Recalling that R_T is even, we can the negative t behavior from $R_T(t) = R_T(-t)$, and obtain the following complete description for the expected value of R_T:

$$\mathcal{E}\,[R_T(t)] = R_X(t)\,\Lambda_T(t) \tag{2.18}$$

where

$$\Lambda_T(t) = \begin{cases} 1 - |t|/2T, & |t| \le 2T \\ 0, & |t| > 2T\,. \end{cases} \tag{2.19}$$

See Figure 2 for a sketch of $\Lambda_T(t)$. Substituting (2.18) into (2.14) gives us this very plausible expression for the PSD:

$$S_X(f) = \lim_{T\to\infty} \mathcal{F}\,[R_X(t)\Lambda_T(t)] = \lim_{T\to\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-2\pi i f t} R_X(t)\,\Lambda_T(t)\,dt \tag{2.20}$$

If it is permitted to put the limit in (2.20) inside the integral we have the result we predicted, equation (2.5), since $\Lambda_T(s) \to 1$ as $T \to \infty$. Priestley (*Spectral Analysis and Time Series*, p 213-4) uses the Lebesgue Dominated Convergence Theorem, and the further condition that

$$\int_{-\infty}^{\infty} |R_X(t)|\;dt < \infty \tag{2.21}$$

to prove that it is permitted to reverse the order of the limit and the integral. For those interested, I give in the Appendix a proof of my own that makes a different set of assumptions about R_X.

3. Some Properties of the PSD

We continue to consider only continuous processes. By establishing the key fact that the autocovariance function and PSD are Fourier transforms of each other, we have shown that they contain the same information. But as we will see later, the PSD is much more useful for the interpretation of actual data because of the intuitive idea that a process divides up naturally as a sum of processes with different frequencies. So while the PSD is the property of the stationary process that is the most informative, the definition via the limit, equation (2.3) is usually very awkward to handle, and the alternative relation (2.5) through autocovariance is the most useful for doing theory, as we will illustrate.

First the simplest example, take white noise. Recall for a continuous process, a white noise has a delta function autocovariance, equation (3.2) in Chapter 1. Then The PSD is

$$S_W(f) = \int_{-\infty}^{\infty} R_W(t)\, \mathrm{e}^{-2\pi \mathrm{i} ft}\; dt = \int_{-\infty}^{\infty} s^2 \delta(t)\, \mathrm{e}^{-2\pi \mathrm{i} ft}\; dt = s^2 \qquad (3.1)$$

which is a constant independent of frequency. White noise has the same power at every frequency, which is why the term is borrowed from physics, because white light is ideally composed of light with the same property, equal power at each frequency. Recall from (2.8) that the area under the PSD is the process variance. For white noise that is infinite, so continuous time white noise is not a physically realizable phenomenon. Notice that the different white noises of Chapter 1 all have exactly the same frequency content, a flat spectrum. The PSD is simply a second order property, and does not concern itself with other details of the distribution defining the stationary process.

Let us consider next what happens to the PSD of process if it has been filtered. In the continuous parameter case we will say

$$Y = g * X \qquad (3.2)$$

where g is a filter function, possibly infinite in extent. When we know S_X we can calculate S_Y. As advertised the simplest approach is to compute the autocovariance of Y. Assume X has mean zero, then so does Y, and

$$R_Y(t) = \operatorname{cov}[Y(s),\, Y(s+t)] = \mathcal{E}\,[Y(s)\, Y(s+t)] \qquad (3.3)$$

$$= \mathcal{E}\left[\int_{-\infty}^{\infty} du\;\, g(u)\, X(s-u) \int_{-\infty}^{\infty} dv\, g(v)\, X(s+t-v) \right] \qquad (3.4)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{E}\,[X(s-u)\, X(s+t-v)]\;\, g(u)\, g(v)\; du\; dv \qquad (3.5)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X(t+u-v)\, g(u)\, g(v)\, du\, dv\,. \tag{3.6}$$

Next we replace the autocovariance on the right with its representation in terms of S_X: we apply (2.6):

$$R_Y(t) = \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} df\; S_X(f)\, \mathrm{e}^{2\pi \mathrm{i} f(t+u-v)}\; g(u)\, g(v) \tag{3.7}$$

$$= \int_{-\infty}^{\infty} df\; \mathrm{e}^{2\pi \mathrm{i} ft}\, S_X(f) \int_{-\infty}^{\infty} du\; \mathrm{e}^{2\pi \mathrm{i} fu}\, g(u) \int_{-\infty}^{\infty} dv\; \mathrm{e}^{-2\pi \mathrm{i} fv}\, g(v) \tag{3.8}$$

$$= \int_{-\infty}^{\infty} df\; \mathrm{e}^{2\pi \mathrm{i} ft} [S_X(f)\, \hat{g}(f)^*\, \hat{g}(f)] \tag{3.9}$$

$$= \mathcal{F}^{-1}[S_X(f)\, \hat{g}(f)^*\, \hat{g}(f)] \tag{3.10}$$

If we take the FT of (3.10) and again recognize the FT of R_Y is the PSD S_Y we see

$$S_Y(f) = \hat{g}(f)^* \hat{g}(f)\, S_X(f) \tag{3.11}$$

$$= |\,\hat{g}(f)\,|^2\, S_X(f)\,. \tag{3.12}$$

This is an **important result.** It is the stochastic process version of the Convolution Theorem applied to filtering: when one filters a deterministic signal, the FT of the resultant function is multiplied by the frequency response of the filter. Here the new spectrum is found by multiplying the original by the *squared magnitude* of the filter function. Equation (3.12) also puts on a firm foundation the idea described in Section 1 that we can obtain a power spectrum by applying a series of ideal, narrow band-pass filters the stochastic process.

4. PSD of Discrete Processes

Every result for a continuous-time stationary stochastic process has an analog in the discrete theory, and there are no surprises. I will state the results here without derivation. Here is the definition of the PSD as a limit of finite Fourier transforms:

$$S_X(f) = \lim_{N\to\infty} \frac{1}{2N} \mathcal{E}\,[\,|\sum_{n=-N}^{N} X_n \mathrm{e}^{-2\pi inf}\,|^2], \quad -½ \le f \le ½. \tag{4.1}$$

The alternative definition through the autocovariance is

$$S_X(f) = \sum_{n=-\infty}^{\infty} R_X(n)\,\mathrm{e}^{-2\pi inf}, \quad -½ \le f \le ½ \tag{4.2}$$

and of course one can obtain the autocovariance from the PSD with the coefficients of the Fourier series expansion in (4.2)

$$R_X(n) = \int_{-½}^{½} S_X(f)\,\mathrm{e}^{2\pi inf}\,df\,. \tag{4.3}$$

Setting $n = 0$ in (4.3) gives

$$\sigma_X^2 = \int_{-½}^{½} S_X(f)\,df \tag{4.4}$$

so that the variance is again the integral of the PSD over frequency. If we restrict ourselves to convolution filters, then filtering a discrete sequence gives the power spectrum

$$S_{g*X}(f) = |\hat{g}(f)|^2\, S_X(f) \tag{4.5}$$

where

$$\hat{g}(f) = \sum_{n=-\infty}^{\infty} g_n\,\mathrm{e}^{-2\pi inf}\,. \tag{4.6}$$

5. Aliasing in the PSD

Although we have seen how aliasing affects sampled data earlier, we should see how sampling modifies the PSD for stochastic processes. This is an important issue in the real world because sampled data may be the only record we have of an underlying continuous physical signal whose power spectrum we would like to know. We call a continuous-time signal $Y(t)$, and we derive from it the discrete process:

$$X_n = Y(n\Delta t), \;\; n = 0, \pm 1, \pm 2, \cdots \tag{5.1}$$

where Δt is the sampling interval. As I asserted earlier it is almost always easier to derive properties of the spectrum from the autocovariance. So the autocovariance of the sampled series is

$$R_X(n) = \mathcal{E}\,[X_j\, X_{j+n}] \tag{5.2}$$

$$= \mathcal{E}\,[Y(j\Delta t)\, Y((j+n)\Delta t) = R_Y(j\Delta t)\,. \tag{5.3}$$

Thus the autocovariance of X_n is simply the sampled autocovariance of Y.

The PSD of the discrete process is given by (4.2), which we modify by including Δt in the exponent to scale the frequencies, replacing the Nyquist frequency of ½ in (4.2) by $1/2\Delta t$, and also scaling the expression by the same quantity:

$$S_X(f) = \Delta t \sum_{n=-\infty}^{\infty} R_X(n)\, \mathrm{e}^{-2\pi i n f \Delta t}, \quad -1/2\Delta t \le f \le 1/2\Delta t\,. \tag{5.4}$$

Now substitute (5.3)

$$S_X(f) = \Delta t \sum_{n=-\infty}^{\infty} R_Y(n\Delta t)\, \mathrm{e}^{-2\pi i n f \Delta t}\,. \tag{5.5}$$

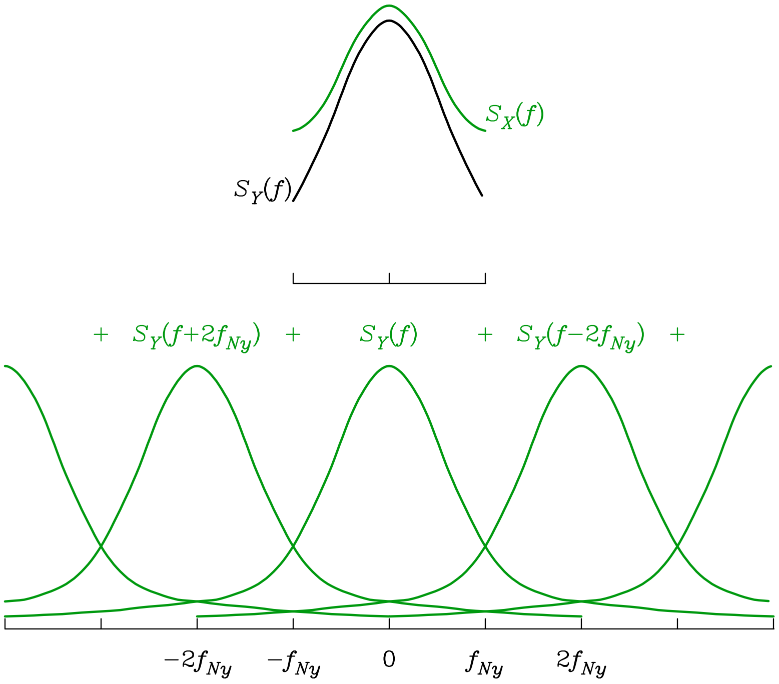
To sum this series we appeal to the **Poisson Sum Rule** given in our treatment of Fourier Theory

$$S_X(f) = \Delta t \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} R_Y(n\Delta t)\, \mathrm{e}^{-2\pi i n f \Delta t}\, \mathrm{e}^{-2\pi i n m}\, dn\,. \tag{5.6}$$

We change variables in the integral: set $t = n\Delta t$ and then recognize the definition of the continuous process PSD:

$$S_X(f) = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} R_Y(t)\, \mathrm{e}^{-2\pi i t(f + m/\Delta t)}\, dt \tag{5.7}$$

Figure 3: Aliasing in the PSD


$S_X(f)$
$S_Y(f)$
$+ \quad S_Y(f+2f_{Ny}) \quad + \quad S_Y(f) \quad + \quad S_Y(f-2f_{Ny}) \quad +$
$-2f_{Ny}$
$-f_{Ny}$
0
f_{Ny}
$2f_{Ny}$

$$= \sum_{m=-\infty}^{\infty} S_Y(f + m/\Delta t)\,. \tag{5.8}$$

Thus the discrete process PSD is a sum of spectra of the original continuous process, shifted by multiples of **twice the Nyquist frequency.** If the power in the continuous process has fallen off to low levels at the Nyquist frequency, the PSD of S_X will be a good approximation to S_Y, although in general S_X will be a factor of two or more above the "true" PSD at $f = 1/2\Delta t$. The moral is that to get a good PSD one must set the sampling rate high enough to avoid aliasing.

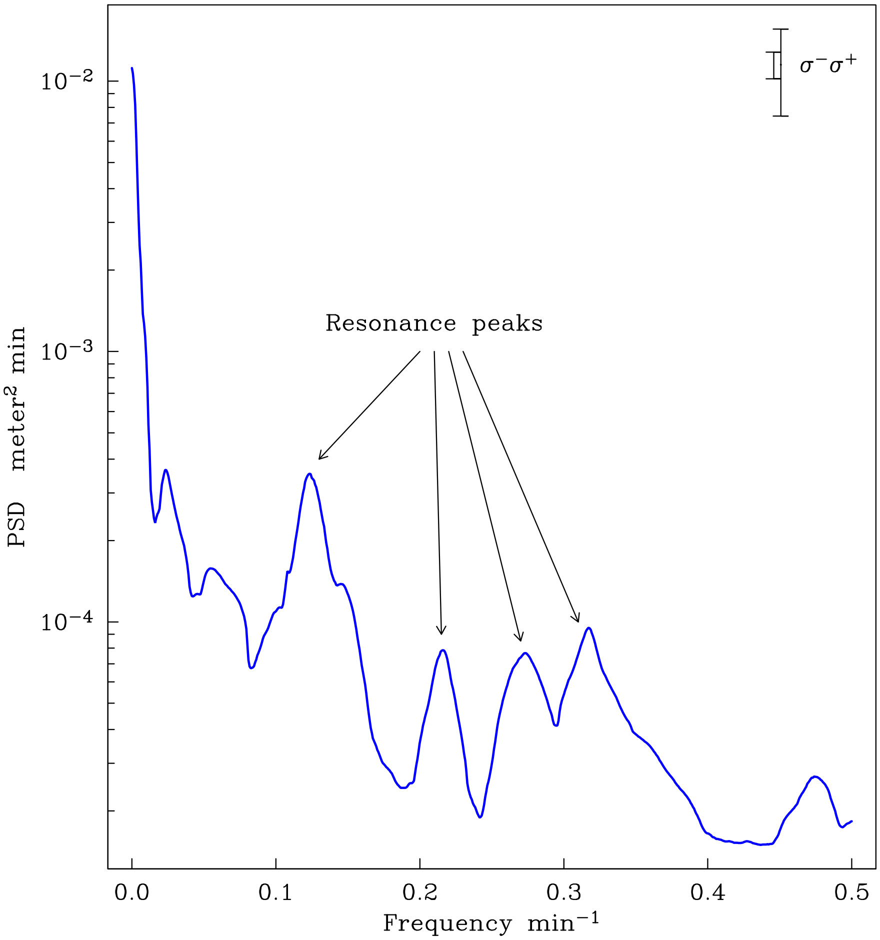
6. Illustrations

Let us look briefly at some PSDs estimated from real-world data. Exactly how the PSD is estimated is the subject of Chapter 3; it is a nontrivial topic which we must defer.

We begin with the Salton Sea seiche data. Below is a an estimate of the PSD of the 1301 data points. As I mentioned in Chapter 1 the data time series suggests a number of resonances excited by the wind. In the PSD we see a series of peaks, well resolved in frequency. The peaks are far from being delta-function lines, and this not a fault of the estimation process. The lake resonances are not perfect since we expect them to broad features due to frictional losses. The four peaks picked out are certainly significant, but other smaller ones are suggestive, though not delineated with confidence. I have plotted two bars, both representing the 1 standard deviation error estimates; the smaller one for places where the spectrum is smooth, the larger for regions of rapid variation.

You should notice several other important things about this graph. First it is plotted on linear-frequency, log-PSD axes. The spectrum covers a large dynamic range, which is almost universal in geophysical data. Plotted on a linear y axis nothing except the left side of the plot would be visible to the eye. **Always** plot your spectra with a log scale. A log

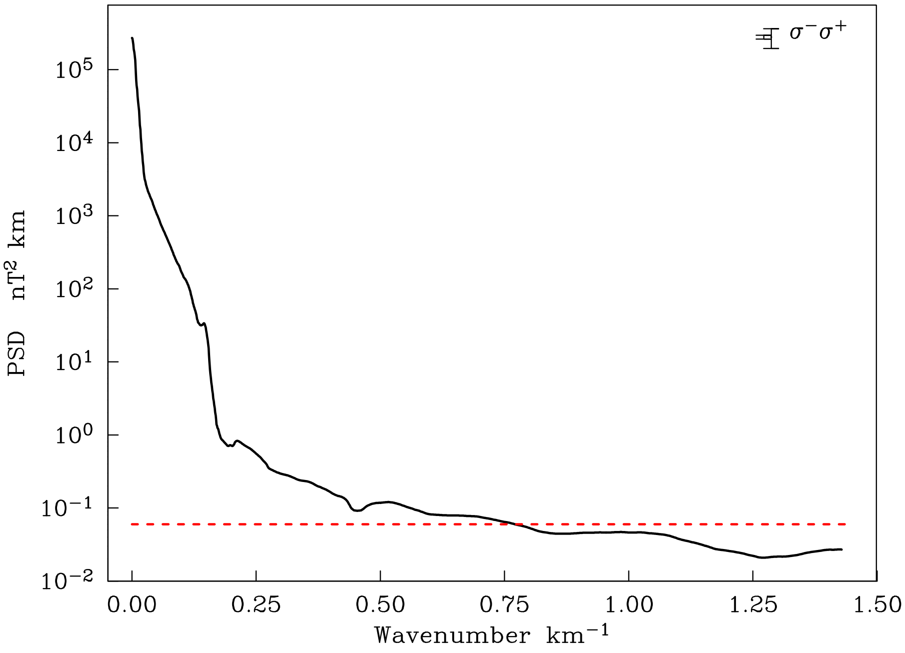
Figure 4: PSD of Salton Sea seiche data.



frequency axis is sometimes useful too, but not in this example. Next observe that rise in power at the low-frequency end. Such a rise is called a **red spectrum** because in optics red light has its power concentrated towards long wavelengths. Red spectra are the norm in geophysical work. This is because attenuation processes operate more efficiently at high frequencies or short wavelengths; most natural filters (like seismic losses, or upward continuation in potential fields) are low-pass filters. Finally, notice the units of the power spectrum: PSD is a measure of variance per unit frequency. Since the frequency here is one cycle per minute, and the series measures water depth, the units are those of squared-length multiplied by time. The Power Theorem for Fourier decomposition means that one could, for example, estimate how much kinetic energy is tied up in the largest resonance by finding the area under the PSD in an appropriate frequency range.

Although the autocorrelation function holds exactly the same information as the PSD, a plot of R_X is almost always useless. We plotted the autocovariance for one component of the magnetic data from the plane flying across the Pacific (Figure 7, Section 4): about the only thing one can deduce from it is that the data series is far from white noise. Let us look at the PSD. This will take two graphs. In Figure 5 I show the whole spectrum. Notice again the very large dynamic range in power. The field values are sampled at an interval of 350 m, so $f_{Ny} = 1/(2 \times 0.35) = 1.43$ km^{-1}. The magnetometer is reported as having an accuracy of ±1 nT. If we take this to mean there is uncorrelated noise with a uniform PDF, with $b = 1$ from Chapter 1 (3.5), we find the variance is $1/12 = 0.0833\,\mathrm{nT}^2$. The

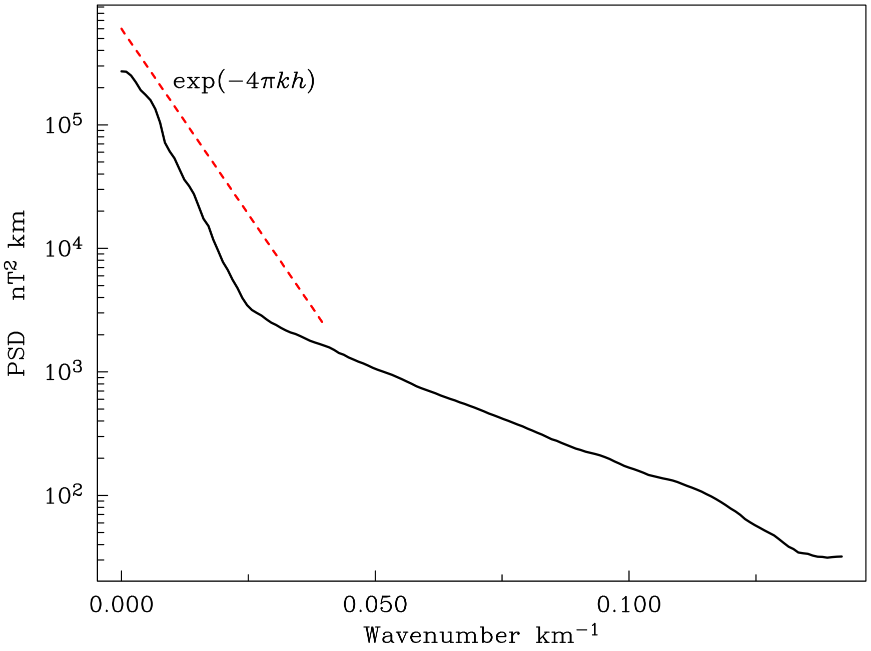
Figure 5: PSD of Z component in magnetic data.



dashed horizontal line corresponds to white noise with this variance: Figure 5 is a one-sided PSD plot, so the area under the dashed line is $\sigma_Z^2 = 1.43 \times 0.0582 = 0.0833\,\mathrm{nT}^2$. What this means is that two-thirds of the frequency content of the record is devoted to almost pure white noise, and is completely uninformative about geophysics — the true geophysical signal at wavenumbers greater than 0.5 km^{-1} (wavelengths shorter than 2 km) is evidently of such low amplitude, the noise of the magnetometer totally obscures it.

In Figure 6 I expand the frequency scale by a factor of ten to show the small wavenumber part of the spectrum more clearly. We see the PSD at this scale approximates two intersecting straight lines. When one looks into the theory of upward continuation of static magnetic fields the process is essentially that of low-pass filtering (as mentioned earlier) — roughly speaking, if $B_z(x)$ is the Z component on a profile at the level of the magnetic sources, then on a path of height h above the first line the field becomes $g * B_z$ where g is filter with response $\hat{g}(k) = \exp(-2\pi kh)$, an exponential fall with wavenumber k. Recall that the effect of the filter on the PSD is to square this response. I have plotted a line corresponding to $|\hat{g}|^2$ in Figure 6 for $h = 7 + 4 = 11\,\mathrm{km}$, the aircraft height plus the average ocean depth. It is plausible to assume that the PSD near the crustal sources of the field falls off fairly slowly, and after upward continuation the spectrum fits that prediction quite well. What then is the other straight-line segment? It can be shown that lack of stability in the gyro system orienting the coordinates for the measurements causes this effect

Figure 6: Lowest 10% in wavenumber of Figure 5.



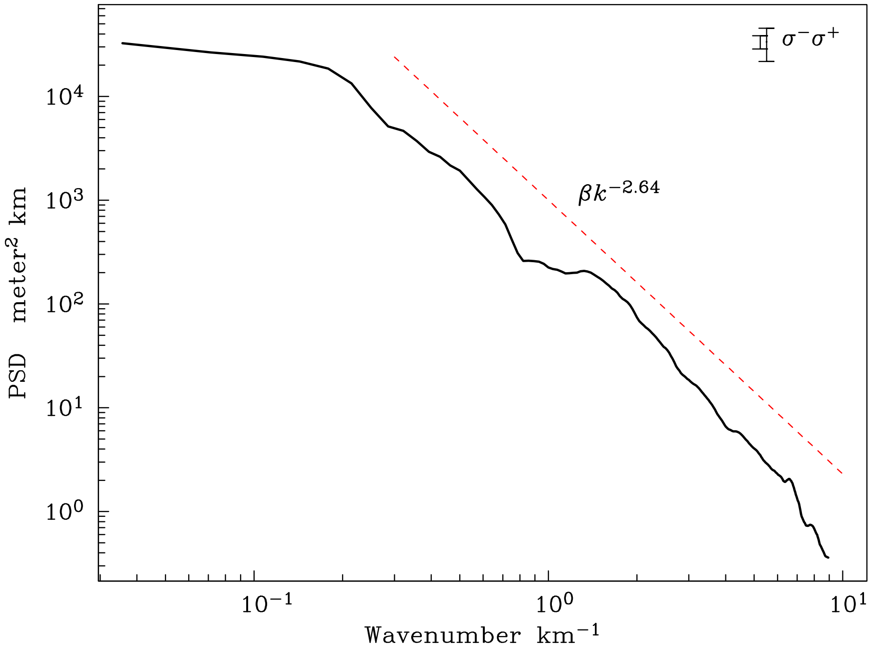
through nonlinear processes; for anyone interested, see Parker and O'Brien, *JGR,* v 102, pp 24815-24, 1997. So it turns out by looking at the power spectrum we have discovered that nothing above $k = 0.03\,\mathrm{km}^{-1}$ concerns geophysics in the observations; only the bottom one-thirtieth of the spectrum contains geophysical information, the rest is noise or one kind or another! It is worth remembering that a leveling off in the PSD at high frequency is often an indication that instrument noise has overwhelmed the signal under investigation at that point, and that the high frequency behavior in the record is probably not geophysical. Unfortunately, interpretation of high-frequency wiggles is a widespread occupation in data analysis.

We have seen in this example how the PSD has revealed a number of remarkable things about the original signal, properties that could not be deduced by visual inspection from the original sequence (shown in Chapter 1 Figure 6) nor from the autocovariance function in Figure 7. More still can be learned by combining the X, Y and Z components in their **cross spectra**, but we don't yet have the theory in hand for that.

Finally, we look at the PSD estimated from the stationary part of the marine bathymetry data set; this is shown in Figure 7. Notice here I have used a log wavenumber axis. The reason for this is to show that the power spectrum of the profile is quite well approximated by a straight line which on this graph means that the PSD has the form of a power law:

$$S_H(k) = \beta\, k^{-q} \tag{6.1}$$

Figure 7: PSD of ocean rise bathymetry.



A fit to the spectrum gives the estimate $q = 2.64$. Power-law behavior has the characteristic that there is no intrinsic length scale. This means that the stochastic process must look the same at any magnification — the hall-mark of a **fractal.** This idealization cannot be true for all scales, and clearly it breaks down in the graph at wavelengths longer than about 3 km. The geological process responsible for the terrain is repeated fracturing and faulting, and there are some theories (Fox and Hayes, *Reviews of Geophysics,* v 23, pp 1-48, 1985) predicting fractal behavior, although I do not believe it is possible to predict the exponent q very well.

Appendix: Proof of Equation (2.20)

In Section 2 I left the proof of the validity of the interchange of the integral and limit to a reference to Priestley. There is a reason why Priestley's proof is not entirely satisfactory, namely, the additional restriction (2.21) is too severe since it excludes autocovariance functions behaving like the sinc function, those associated with a discontinuity in S_X. Here is an alternative proof.

We examine the difference between the desired result and the function within the limit, and show that the difference vanishes under some mild restrictions, as T becomes large. We measure the discrepancy with the 2-norm:

$$\| f(t) \|^2 = \int_{-\infty}^{\infty} | f(t) |^2 \, dt \,. \tag{A1}$$

Define the number

$$\Delta_T = \| \mathcal{F}[\Lambda_T R_X - R_X] \| = \| \Lambda_T R_X - R_X \| \,. \tag{A2}$$

The second equality follows from the Power Theorem, the invariance of the 2-norm under the FT. Then, since Λ_T and R_X are both real and even

$$\Delta_T^2 = 2 \int_0^{\infty} (\Lambda_T(t) - 1)^2 R_X(t)^2 \, dt \tag{A3}$$

$$= 2 \int_0^{2T} (\Lambda_T(t) - 1)^2 R_X(t)^2 \, dt + 2 \int_{2T}^{\infty} R_X(t)^2 \, dt \tag{A4}$$

$$= 2 \int_0^{2T} \frac{t^2}{4T^2} R_X(t)^2 \, dt + 2 \int_{2T}^{\infty} R_X(t)^2 \, dt \,. \tag{A5}$$

To proceed we need to assume something more about the behavior of R_X. We know R_X is never greater than σ_X^2, so it is bounded; we will assume that it dies away for large t, but more rapidly than some power:

$$| R_X(t) | < \frac{c}{(1+t)^{\nu}} \tag{A6}$$

for some fixed values of c and ν. With this constraint we can see that

$$\Delta_T^2 < \frac{1}{2T^2} \int_0^{2T} \frac{c^2 t^2}{(1+t)^{2\nu}} \, dt + 2 \int_{2T}^{\infty} \frac{c^2}{(1+t)^{2\nu}} \, dt = \frac{A(T)}{T^2} + B(T) \,. \tag{A7}$$

Our interest lies in the behavior of Δ_T as T tends to infinity. By L'Hopital's rule of elementary calculus on the first term:

$$\lim_{T \to \infty} \frac{A(T)}{T^2} = \lim_{T \to \infty} \frac{A'(T)}{2T} \tag{A8}$$

$$= \lim_{T \to \infty} \frac{2Tc^2}{(1+2T)^{2\nu}} = 0, \quad \text{when } \nu > ½ \,. \tag{A9}$$

In the second term we find

$$B(T) < \int_{2T}^{\infty} \frac{2c^2}{t^{2\nu}}\, dt = \frac{2c^2}{2\nu - 1} \frac{1}{(2T)^{2\nu-1}} \tag{A10}$$

and, provided that $\nu > ½$, $B(T)$ tends to zero with large T. Thus if $\nu > ½$, both terms in (A7) tend to zero for large T and this means that Δ_T vanishes, and hence the discrepancy between the FT of R_X and the $\mathcal{F}[\Lambda_T R_X]$ also vanishes in the limit. In the context of bounded functions like $R_X(t)$ our class of functions in (A6) is much bigger than the one in Priestley's proof and includes the sinc function, for example.

GEOPHYSICAL DATA ANALYSIS

Robert L. Parker

CHAPTER 3: ESTIMATING THE POWER SPECTRAL DENSITY

1. Introduction

If we are willing to assume that given data set is (after suitable pre-processing as necessary) approximately modeled by a stationary stochastic process, we would like procedures that can make reasonably reliable estimates of the PSD, or the autocovariance function, of the underlying process. We confront a problem not previously encountered in statistical estimation: while we have of necessity only finitely many data values, the thing we need to know is a function, something with infinitely many unknowns. We will have to settle for a simplified version of the function, usually a smoothed version. In some estimation processes, the function is itself written in terms of a few parameters and modeled with a simple rational expression terms of sines and cosines of the frequency. These approaches (for example, **maximum entropy**) are considered by the experts to be unreliable and we will not cover them here. But we will make some simplifications, though not as drastic as the few-parameter model.

We will normally treat a discrete process $\{X_n\}$ often with sampling interval $\Delta t = 1$, and although the underlying stationary process is infinite, we will have at our disposal only N consecutive terms from a single realization. There will be no spectral lines, meaning no exactly periodic components. If these are suspected to be present they should be removed first by other means, just as the mean or a trend should be removed before spectral analysis. Finally, and importantly, we must assume the process is **ergodic.**

What does this last term mean? In the theory we have assumed that it is possible to generate as many realizations of the process as required, and then when averaging, such as the expectation $\mathcal{E}$ is needed, we take the average over the independent realizations. In most practical situation we are possession of exactly **one** realization, the data series under study. To reduce variance, it is absolutely essential to average something. In practice we must average over time. An *ergodic* stationary process is one in which averaging over infinite amounts in time gives the same answer as averaging over infinitely many repeated realizations. It is perfectly possible to invent stochastic processes for which this fails. We can guess that a process would be ergodic if the autocorrelation dies away fast enough, that pieces of the data series separated far enough are essentially independent. For discrete processes it is sufficient that:

$$\sum_{n=-\infty}^{\infty} R_X(n)^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f)^2 \, df < \infty \qquad (1.1)$$

for the process X_n to be ergodic. This is a pretty mild condition; any bounded spectrum satisfies it.

2. Several Bad Approaches

Spectral estimation is the subject of a huge literature. The reason for this is that the task is still something of an art, and there is no definitive best way. None-the-less, most of the methods in the literature are rather poor, and there are many people in the community who cling to decidedly inferior methods because of ignorance or reluctance to change. I will be recommending one specific approach based on the periodogram. Here I will mention briefly some of the inferior methods, so that when you see them in the literature you will know the author hasn't kept up, or is deluded. For a more complete catalog of inferior estimation methods see the introduction in Thomson's classic 1982 paper (Thomson, D. J., Spectrum estimation and harmonic analysis, *Proc. IEEE,* v 70, pp 1055-96, 1982).

Many estimation methods in statistics go as follows: look at the definition of the particular statistic, and if it involves the expectation operation over the process, replace the expectation by an average over the sample in hand. The mean is an obvious illustration: definition $\bar{x} = \mathcal{E}\,[X]$; estimator

$$\hat{\hat{X}} = \frac{1}{N} \sum_{n=1}^{N} x_n \, . \tag{2.1}$$

(We run into a notational problem, that the hat ˆ is used in statistics to denote an estimate, while in Fourier theory, it is the Fourier transform: we need both! We resolve the problem with this rule: if the hat is applied to an upper case letter, it is the estimator, not the FT.)

We can apply this approach to the two definitions of PSD in Chapter 2. The first, (2.2) Chapter 2, the limit over longer and longer sections of a discrete FT of the data, is called the **periodogram estimator** and ultimately will be the basis of the preferred method; much more on this estimator later. The second estimator would be a two stage affair: estimate the autocovariance function thus:

$$\hat{R}_X(n) = \frac{1}{N-n} \sum_{k=1}^{N-n} x_k \, x_{k+n} , \quad n = 0, 1, 2, \cdots N-1 \tag{2.2}$$

from the measured data series x_n; then take the discrete FT to find $\hat{S}_X$. This turns out to be a very poor estimator, both of R_X and S_X. One problem is that the variance of the autocovariance estimate in (2.2) gets worse and worse as n increases, and the higher variances are then spread across the whole spectrum. It is very difficult to find the uncertainties in these estimates and there is significant issue with bias too. This method is almost never seen today. (Except in some parts of the paleoclimate literature!)

Another class of estimators more widely advocated derives from the filter theory for stationary processes. Recall (3.11) Chapter 2: if $X = g * T$ then

$$S_X = |\hat{g}|^2 S_Y \, . \tag{2.3}$$

Suppose that you could somehow choose the filter g so that when Y is white noise, the filter output is the desired process X; then

$$S_X(f) = c \, |\hat{g}(f)|^2 \tag{2.4}$$

where c is a constant. The idea is usually implemented by selecting g to be a finite AR (autoregressive) filter with k weights:

$$X_n = Y_n + a_1 X_{n-1} + a_2 X_{n-2} + a_k X_{n-k} \tag{2.5}$$

where Y_n is a white process. By multiplying through these equations with X_j and taking the expectation, we can generate a set of equations (known as the **Yule-Walker equations**) for the unknown coefficients a_j in terms of estimates of the autocovariances, like (2.2); see Seion 9, and Priestley, pp 349-51. As with the earlier method based on estimated autocovariance, it is hard to find uncertainties. Also there is the question of the proper choice of k, the number of terms to be taken in the filter model. The answers vary wildly with different choices, and there is no good theory to decide what is the correct number. There are alternative methods for finding the coefficients, for example, **Burg's method,** and the ever popular but equally flawed **method of maximum entropy.**

The Yule-Walker approach does have a valuable application, however. As we will see, spectral estimates based on the periodogram method are biased if the PSD covers a wide range. Often a relatively short filter like (2.5) (with $k \le 5$ say) can be found that dramatically reduces the dynamic range, and then the filtered process can be safely treated, after which the effect of the filter is removed. The process is called **prewhitening.**

We return to the classic bad estimator, the **periodigram estimator**. As already noted the idea is to replace expectations with ordinary averages. Of course we have a limit in the definition, but we will simply ignore this inconvenience. Then (4.1) of the previous Chapter becomes the estimator:

$$\hat{S}_X(f) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x_n \mathrm{e}^{-2\pi i n f} \right|^2 . \tag{2.6}$$

Notice we have numbered the data sequence to start at zero, which is the usual convention for the DFT. Equation (2.6) is what one often encounters for an estimate of the PSD by unsophisticated people – it seems natural just to take the FFT and square the magnitudes of the coefficients. Methods based on modifications of (2.6) are called **direct spectral estimates.** Direct estimates are the kind I recommend, and since they start at the periodogram, we must study it in some detail to understand why it is bad estimator, and armed with that knowledge, fix the problems.

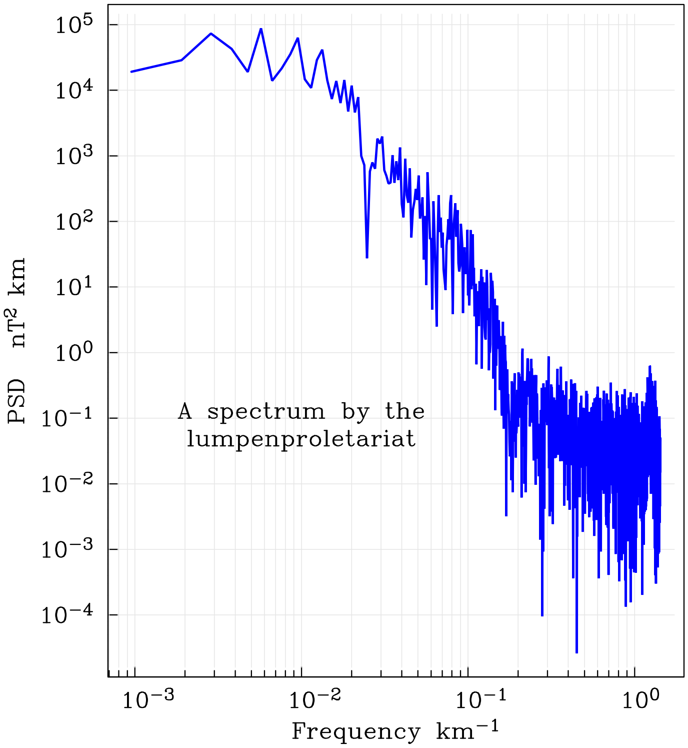


Figure 0: A primitive periodogram spectrum of the Project Magnetic *X* field shown in earlier Chapters

The spectrum shown on this page is typical of many still to be found in the literature. When you see the log-log scales and the dense wiggles at high frequency, you know immediately that the author needs to take a class in spectral estimation!

3. The Raw Periodogram: White Gaussian Noise

We begin with the simplest case: the periodogram for Gaussian white noise. Since all practical observations are discrete, we will study almost exclusively discrete stationary stochastic processes. For convenience we repeat a few well-known properties of the white noise, which we now call X_n. The X_n are iid Gaussian RVs with zero mean and variance σ^2. Hence $X_n \sim N(0, \sigma^2)$ and

$$R_X(n) = \operatorname{cov}[X_j, X_{j+n}] = \mathcal{E}[X_j X_{j+n}] = \sigma^2 \delta_{n0} . \tag{3.1}$$

We are attempting to estimate the true PSD of this process, which is

$$S_X(f) = \sigma^2, \quad -½ \le f \le ½ . \tag{3.2}$$

We will use (2.6) to estimate the PSD at $N+1$ evenly spaced frequencies: $f_m = m/N = m\Delta f$, with $m = 0, \pm 1, \pm 2, \cdots \pm N/2$, and we will take N to be an even number for convenience. Thus the frequencies sample the spectrum across the band, right up to the Nyquist frequency $f = \pm½$. These are easy frequencies to calculate with the FFT, but as we will see there are other reasons for this choice of frequencies. Define the real and imaginary parts of the DFT in (2.6):

$$A_m = \operatorname{Re} \sum_{n=0}^{N-1} x_n \mathrm{e}^{-2\pi i n m/N} ; \qquad B_m = \operatorname{Im} \sum_{n=0}^{N-1} x_n \mathrm{e}^{-2\pi i n m/N} \tag{3.3}$$

and then the periodogram estimate is

$$\hat{S}_X(m\Delta f) = \frac{A_m^2 + B_m^2}{N} . \tag{3.4}$$

We will now characterize the statistical distributions of A_m and B_m. Observe that by definition these are simply weighted sums of samples drawn from a Gaussian process. Therefore it follows that all the A_m and B_m must be Gaussian RVs too, and as such we can completely specify their joint distribution from a knowledge of the mean values and the covariances (Recall (1.1) and (2.8) from Chapters 1 and 2).

First the mean values, which are easy:

$$\mathcal{E}[A_m] = \operatorname{Re} \sum_{n=0}^{N-1} \mathcal{E}[X_n] \mathrm{e}^{-2\pi i n m/N} = 0 \tag{3.5}$$

and similarly $\mathcal{E}[B_m] = 0$. To calculate the variances and covariances write

$$C_m = A_m + \mathrm{i} B_m . \tag{3.6}$$

Now consider

$$\mathcal{E}[C_j C_k^*] = \mathcal{E}[A_j A_k + B_j B_k] - \mathrm{i}\mathcal{E}[A_j B_k - A_k B_j] \tag{3.7}$$

$$= \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \mathcal{E}[X_m X_n] \, \mathrm{e}^{-2\pi i m j/N} \, \mathrm{e}^{2\pi \mathrm{i} n k/N} \tag{3.8}$$

$$= \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \sigma^2 \, \delta_{mn} \, \mathrm{e}^{-2\pi \mathrm{i}(mj - nk)/N} \tag{3.9}$$

$$= \sigma^2 \sum_{m=0}^{N-1} \mathrm{e}^{-2\pi \mathrm{i} m(j-k)/N} \, . \tag{3.10}$$

This is a sum we have seen several times before in this course. It is a geometrical series and was used to demonstrate the orthogonality of the complex vector basis $\mathrm{e}^{2\pi \mathrm{i} n/N}$. Recall that it vanishes unless $j = k$; in that case the sum is N. Therefore

$$\mathcal{E}\,[A_j A_k + B_j B_k] - \mathrm{i}\mathcal{E}\,[A_j B_k - A_k B_j] = N\,\sigma^2 \delta_{jk} \, . \tag{3.11}$$

In an exactly similar way we can calculate

$$\mathcal{E}\,[C_j\, C_k] = N\,\sigma^2 \delta_{j,-k} \, . \tag{3.12}$$

And so

$$\mathcal{E}\,[A_j A_k - B_j B_k] + \mathrm{i}\mathcal{E}\,[A_j B_k + A_k B_j] = N\,\sigma^2 \delta_{j,-k} \, . \tag{3.13}$$

Equations (3.11) and (3.13) give for each pair (j, k) two complex equations in $\mathcal{E}\,[A_j A_k]$, $\mathcal{E}\,[B_j B_k]$, and $\mathcal{E}\,[A_j B_k]$. We solve for these and, omitting the simple manipulations, obtain the following:

$$\begin{array}{rcll}
\mathrm{cov}\,[A_j, B_k] & = & 0, & \text{all } \; j, k \\
\mathrm{cov}\,[A_j, A_k] & = & 0, & j \neq k \\
\mathrm{cov}\,[B_j, B_k] & = & 0, & j \neq k \\
\mathrm{var}\,[A_j] & = & \tfrac{1}{2} N\sigma^2, & j \neq 0, \; j \neq \pm \tfrac{1}{2} N \\
\mathrm{var}\,[A_j] & = & N\sigma^2, & j = 0, \pm \tfrac{1}{2} N \\
\mathrm{var}\,[B_j] & = & \tfrac{1}{2} N\sigma^2, & j \neq 0, \; j \neq \pm \tfrac{1}{2} N \\
\mathrm{var}\,[B_j] & = & 0 & j = 0, \pm \tfrac{1}{2} N \, .
\end{array} \tag{3.14}$$

In summary, this reveals that the A_j and B_j are completely uncorrelated — the covariances all vanish. The variances of the individual variables are all $\tfrac{1}{2} N\sigma^2$, except for zero and the Nyquist frequency, where that of the real part A_j is doubled, and that of the imaginary part B_j vanishes. So if we exclude zero and the Nyquist frequency (both ±½ will be intended by this phrase), the DFT comprises iid Gaussian RVs $\sim N(0, \tfrac{1}{2} N\sigma^2)$; almost "Gaussian white noise in, Gaussian white noise out."

Such statistical independence is useful, but we would fail to get it if we tried to estimate S_X at frequencies other than those we have chosen; this is the main reason for the choice.

Now to the estimator of the PSD. Setting aside the zero frequency and the Nyquist for now, we must consider in (3.4) the weighted sum of two independent variables each one the square of a Gaussian RV with mean zero, and variance $\tfrac{1}{2} N\sigma^2$, which we will called ρ^2. This is usually treated by recalling that the sum of K iid squared Gaussian RVs is distributed as χ_K^2, the chi-squared distribution with K degrees of freedom; and here $K = 2$. For such a simple problem, that is too complicated. We need the expected value of $\hat{S}_X$:

$$\mathcal{E}\,[\hat{S}_X(m\Delta f)] = \frac{1}{N}\,\mathcal{E}\,[A_m^2 + B_m^2]\,. \tag{3.15}$$

But A_m and B_m are independent, zero-mean Gaussian variables, and so

$$\mathcal{E}\,[\hat{S}_X(m\Delta f)] = \frac{1}{N}\,(\mathcal{E}\,[A_m^2] + \mathcal{E}\,[B_m^2]) = \frac{1}{N}\,(\mathrm{var}\,[A_m] + \mathrm{var}\,[B_m]) \tag{3.16}$$

$$= \frac{1}{N}\,(\tfrac{1}{2}N\sigma^2 + \tfrac{1}{2}N\sigma^2) = \sigma^2\,. \tag{3.17}$$

This is the true value of the PSD, so the periodogram estimator is unbiased. The same holds at zero and the Nyquist frequency, as can readily be verified. That is the good news. Now for the variance.

Since the RVs are identical and independent, we can treat A_m alone then double the answer.

$$\mathrm{var}\,[\hat{S}_X(m\Delta f)] = \mathrm{var}\left[\frac{1}{N}\,(A_m^2 + B_m^2)\right] = \frac{2}{N^2}\,\mathrm{var}\,[A_m^2] \tag{3.18}$$

$$= \frac{2}{N^2}\,(\mathcal{E}\,[A_m^4] - \mathcal{E}\,[A_m^2]^2)\,. \tag{3.19}$$

The second expectation is again $\mathrm{var}\,[A_m] = \tfrac{1}{2}N\sigma^2$. The first is the fourth moment of a zero-mean Gaussian; with variance ρ^2 this is:

$$\int_{-\infty}^{\infty} x^4\, \mathrm{e}^{-x^2/2\rho^2}/\rho\sqrt{2\pi}\; dx = 3\rho^4\,. \tag{3.20}$$

Here $\rho^2 = \mathrm{var}\,[A_m] = \tfrac{1}{2}N\sigma^2$. Assembling the pieces in (3.18) we find

$$\mathrm{var}\,[\hat{S}_X(m\Delta f)] = \frac{2}{N^2}\,[3(\tfrac{1}{2}N\sigma^2)^2 - (\tfrac{1}{2}N\sigma^2)^2] = \sigma^4, \quad m \neq 0, \pm\tfrac{1}{2}N\,. \tag{3.21}$$

A short calculation of the same kind confirms that:

$$\mathrm{var}\,[\hat{S}_X(0)] = \mathrm{var}\,[\hat{S}_X(\pm\tfrac{1}{2})] = 2\sigma^4\,. \tag{3.22}$$

So the standard error of the estimate (square-root variance) is normally σ^2, which is the same size as the estimate itself. An uncertainty so large is bad enough, but what is worse is the fact that, as the number of data N grows towards infinity, the variance does not improve; hence the periodogram generates **inconsistent estimates.**

The reason for this behavior can be understood by a simple matter of counting. The periodogram gives $\tfrac{1}{2}N$ independent estimates from N data; the number of degrees of freedom per estimate is two. This does not improve with increasing N, so the variance doesn't get any smaller — we are not averaging over more data as we increase N.

4. The Raw Periodogram: Continuous Spectra

How do these results generalize to the spectra of processes other than white noise? For continuous power spectra, we will show that the periodogram estimator is **asymptotically unbiased,** which means as $N \to \infty$ the expected value of the estimate is the correct PSD. Also as N grows, the variance has the same kind of behavior as in the white-noise case:

$$\operatorname{var}[\hat{S}_X(f)] \to S_X(f)^2 . \tag{4.1}$$

Let us calculate the bias, when the number of samples is N, the practical situation. Sparing you some algebra, which goes exactly like the proof we gave showing the equivalence of the two definitions of PSD in the continuous time case, we find that the discrete DFT estimator (2.6) can also be written without approximation as

$$\hat{S}_X(f) = \sum_{n=-N+1}^{N-1} \left(1 - \frac{|n|}{N}\right) \hat{R}_X(n)\, \mathrm{e}^{-2\pi i n f} \tag{4.2}$$

where $\hat{R}_X$ is the estimator (2.2) of the autocovariance function:

$$\hat{R}_X(n) = \frac{1}{N - |n|} \sum_{k=0}^{N-|n|} X_k\, X_{k+n} . \tag{4.3}$$

This estimator is unbiased. (Show this.) To find the bias of the PSD estimator in (4.2) we take the expectation:

$$\mathcal{E}\,[\hat{S}_X(f)] = \sum_{n=-N+1}^{N-1} \left(1 - \frac{|n|}{N}\right) \mathcal{E}\,[\hat{R}_X(n)]\, \mathrm{e}^{-2\pi i n f} \tag{4.4}$$

$$= \sum_{n=-N+1}^{N-1} \left(1 - \frac{|n|}{N}\right) R_X(n)\, \mathrm{e}^{-2\pi i n f} . \tag{4.5}$$

Now recall that R_X is the integral over the true PSD, equation (4.3) of Chapter 2; substitute the integral:

$$\mathcal{E}\,[\hat{S}_X(f)] = \sum_{n=-N+1}^{N-1} \left(1 - \frac{|n|}{N}\right) \int_{-1/2}^{1/2} S_X(f')\, \mathrm{e}^{2\pi i f' n}\, df'\, \mathrm{e}^{-2\pi i f n} \tag{4.6}$$

$$= \int_{-1/2}^{1/2} S_X(f') \left\{ \sum_{n=-N+1}^{N-1} \left(1 - \frac{|n|}{N}\right) \mathrm{e}^{-2\pi i (f - f')n} \right\} df' \tag{4.7}$$

$$= \int_{-1/2}^{1/2} S_X(f')\, F_N(f' - f)\, df' \tag{4.8}$$

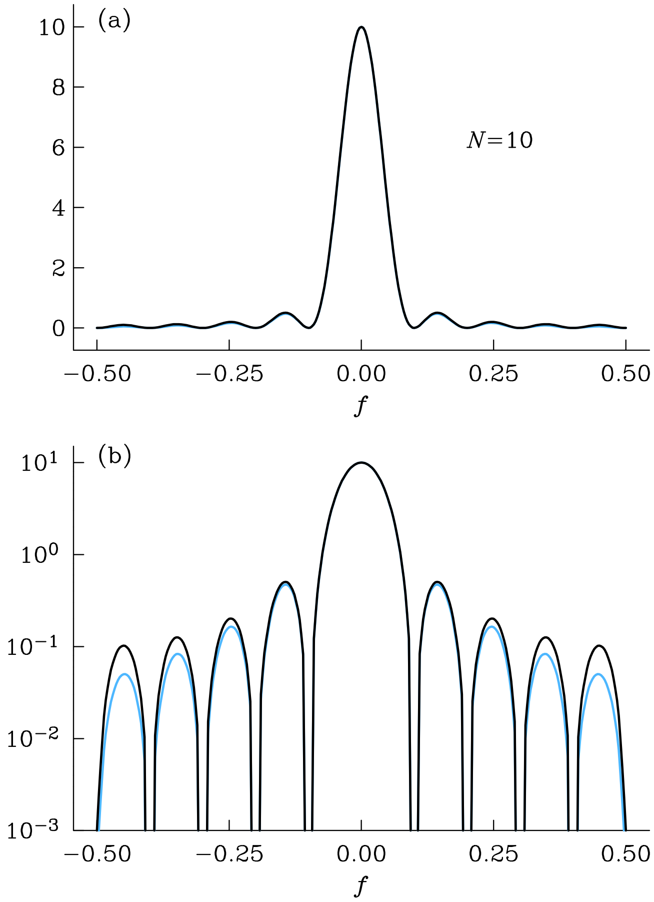
where the sum in (4.8) can be evaluated exactly, and is called the **Fejer kernel**:

$$F_N(f) = \frac{1}{N} \frac{\sin^2 \pi N f}{\sin^2 \pi f} . \tag{4.9}$$

This function is very well approximated by $N\,\mathrm{sinc}^2(Nf)$, even for modest values of N; see Figure 1.

Equation (4.8) demonstrates that the periodogram estimator *convolves the true spectrum with a function resembling sinc-squared*, a function with considerable amplitude away from the central peak. When the spectrum is flat the convolution has no effect, but when there are peaks or other variations, the effect can be serious, particularly in the most interesting cases, where the PSD has a large dynamic range. This bias is called **spectral leakage.** Notice that as N tends to infinity, the kernel gets taller and narrower ultimately yielding the correct expected value and so, as advertised, the estimate is asymptotically unbiased.

Figure 1: Fejer kernel: (a) linear scale; (b) log scale. Blue curve is sinc-squared approximation.



5. Simple Fixes for the Periodogram

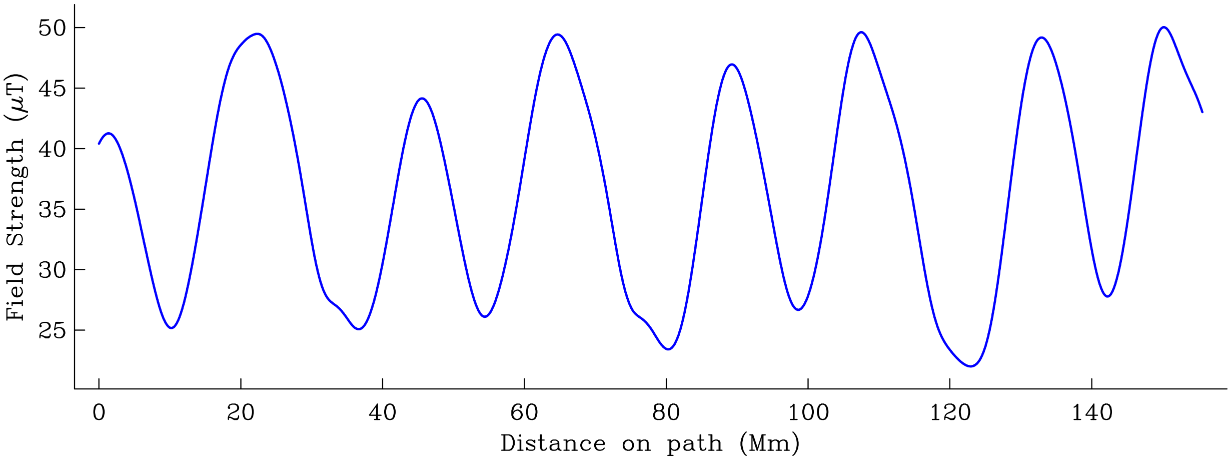
To improve the poor variance of the periodogram we must expect to average in some way. The most natural approach, which we will examine briefly, is simply to average together a number of estimates made at neighboring frequencies. The periodogram has the virtue that the spectral estimates at the frequencies $m\Delta t$ are uncorrelated for white noise, and this remains approximately true even for other smooth spectra. Therefore when we average K consecutive spectral estimates together with a uniform weight, we reduce the variance by a factor of K: from (4.1)

$$\mathrm{var}\,[<\hat{S}_X(f)>_K] \approx \frac{S_X(f)^2}{K}\,. \tag{5.1}$$

You will see that averaging introduces a smoothing of the spectrum, a new bias of its own. Such **loss of spectral resolution** is inescapable; we must balance the desire to see detail in the PSD against the need for statistical reliability.

Another technique, still in wide use, is **Welch's method** or **section averaging**. Here one splits the original record up into K segments of equal length, and makes spectral estimates from each one. To the extent that the sections are long enough, the data series are approximately independent, and thus each one provides a statistically independent estimate of the PSD at each frequency, and these are then averaged. Notice, just as in the case of frequency-domain averaging, one loses spectral resolution, because now the interval between consecutive frequencies in the estimated PSD becomes K/N instead of $1/N$ as it was with the whole record. To reduce bias the sections are each tapered, a process we will discuss next, and it is also part of the method that the segments should overlap to make maximum use of the information. We will not examine the method in detail because it is now obsolete in my opinion.

Figure 2: Total field observation on 4 complete orbits of Magsat.



In Figure 2 I show the total geomagnetic field strength as measured during four orbits of Magsat. There are 4096 measurements, taken at a sampling interval of 38 km along track. I want to use this fairly extreme example for illustration. In Figure 3 we see two periodogram estimates in the lowest wavenumber part of the spectrum; the Nyquist wavenumber is $0.0132\,\mathrm{km}^{-1}$. The orange curve is the raw periodogram, which is as predicted extremely noisy. In fact above $k = 0.0003\,\mathrm{km}^{-1}$ the raw periodogram becomes smooth, an effect of spectral leakage. Also shown is the result of averaging 11 consecutive periodogram estimates: a great improvement results in the variance.

Averaging is the traditional way to reduce variance and it is no surprise that the variability in the PSD here as been brought down considerably. But what of the bias? It is no coincidence that the bias-producing kernel in (4.8) is approximately the square of the sinc function, well-known as the FT of a box-car function. By the technique we used in section 4 we can show a more general result: suppose that the data sequence is multiplied by a weight series $w(n)$ and the original periodogram estimate is replaced by

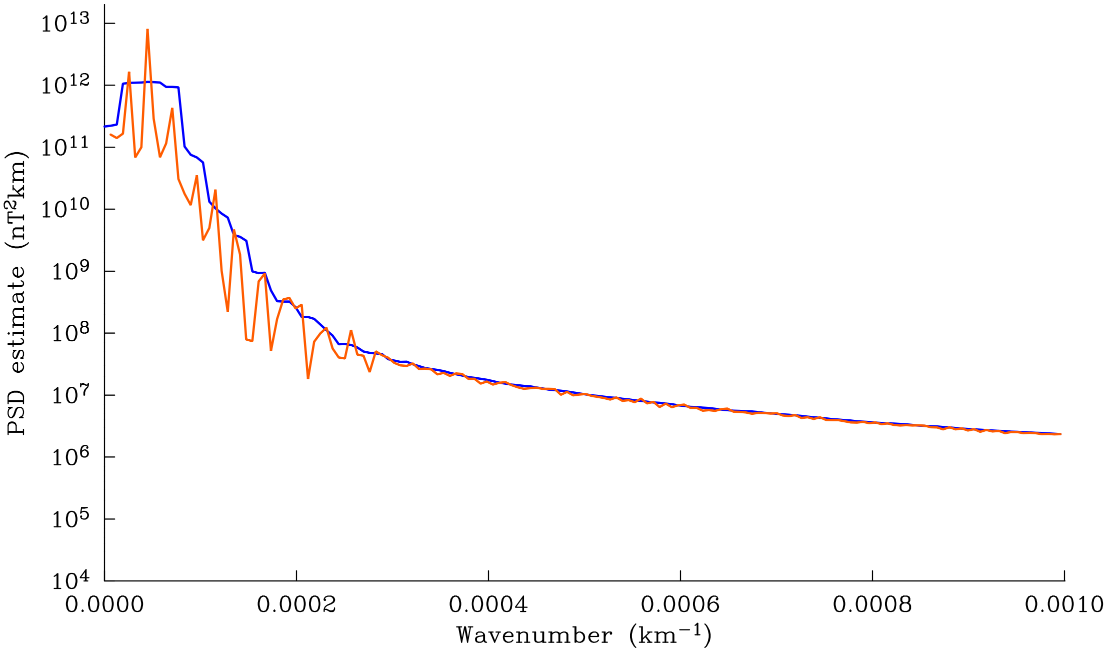
$$\hat{S}_X(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} w(n)\,x_n \mathrm{e}^{-2\pi i n f}\right|^2 . \tag{5.2}$$

Then the new expected value of the estimator comes out as

$$\mathcal{E}\,[\hat{S}_X(f)] = \int_{-1/2}^{1/2} S_X(f')\,W(f'-f)\,df' \tag{5.3}$$

where the convolving function in the frequency domain is given

Figure 3: Raw periodogram of data in Figure 2, shown orange; 11-point average blue.



approximately by

$$W(f) = \frac{1}{N} \,|\, \hat{w}(f) \,|^2 = \frac{1}{N} \,|\, \mathcal{F}\,[w] \,|^2 \tag{5.4}$$

that is, the squared magnitude of the FT of the weight series. In the case of the raw periodogram, the weight series comprises a set of N ones: suppose we take $w(t)$, the function of continuous time w in (5.4), to be:

$$w(t) = \mathrm{box}(t/N - ½)\,. \tag{5.5}$$

Then

$$\hat{w}(f) = N \mathrm{e}^{\pi \mathrm{i} f N} \mathrm{sinc}\,(Nf) \tag{5.6}$$

and because $|\,\mathrm{e}^{\pi \mathrm{i} f N}\,| = 1$ (5.4) agrees exactly with the approximation $N\,\mathrm{sinc}^2(Nf)$ mentioned earlier and plotted as the blue curve in Figure 1.

To improve the bias of spectral leakage, we deliberately choose $w(n)$ to be a function whose FT falls away faster than sinc^2, which for large frequencies has f^{-2} behavior. The key is to have smooth behavior in the transitions at $n = 0$ and $n = N - 1$, the ends of the interval, because the leakage is a consequence of poor convergence of the Fourier transform of a function with a discontinuity. Let us introduce a continuous time approximation for convenience in the use of (5.6): let $t = n$ and $T = N - 1$ so that the observations are on the time interval $(0, T)$. Nowadays the function $w(t)$ is called a **taper.** In the earlier literature it was known as a "data window," and in some fields called the "apodizing function." There are many choices of suitable $w(t)$, and a large number have been given names in the literature: see Priestley for definitions of the Daniell, Bartlett, Parzen, Tukey-Hamming, Tukey-Hamming and Bartlett-Priestley tapers! *They are all obsolete,* for reasons we will soon come to. For purposes of illustration, we will look at two simple examples: First the sine taper

$$w_A(t) = \begin{cases} (2/T)^{½} \sin\,(\pi t/T), & 0 \le t \le T \\ 0, & \text{otherwise}\,. \end{cases} . \tag{5.7}$$

The leading constant factor is to insure unit area under $|\,\hat{w}\,|^2$, which we need if the convolution (4.8) is to get the right answer for white noise. Then the FT is

$$\hat{w}_A(f) = (½T)^{½}\,[\mathrm{sinc}\,(fT + ½) + \mathrm{sinc}\,(fT - ½)]\,. \tag{5.8}$$

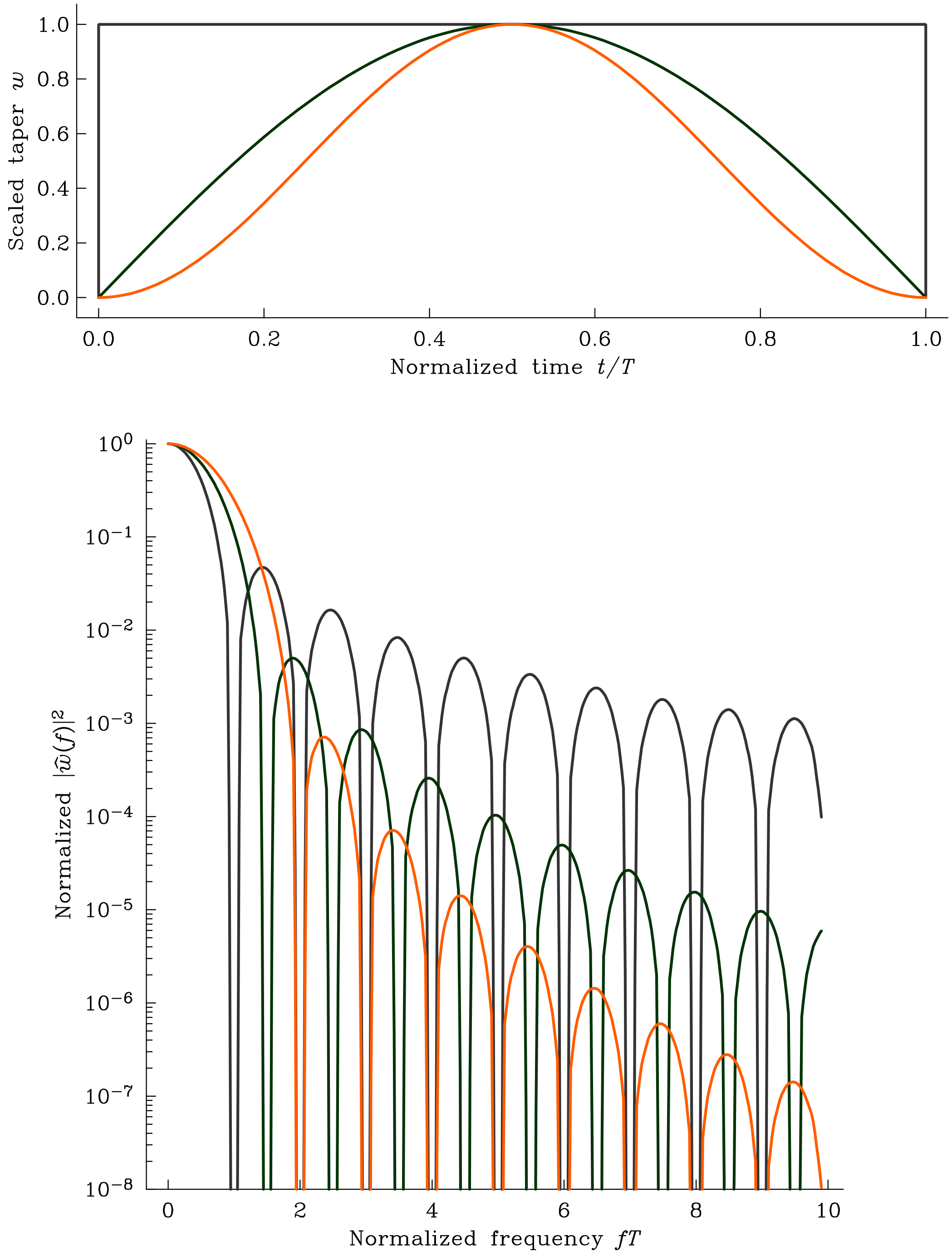
Next the sine-squared taper:

$$w_B(t) = \begin{cases} (8/3T)^{½} \sin^2 (\pi t/T), & 0 \le t \le T \\ 0, & \text{otherwise} \end{cases} \tag{5.9}$$

$$\hat{w}_B(f) = \left(\frac{2T}{3}\right)^{½} [\mathrm{sinc}\,(fT) + ½\mathrm{sinc}\,(fT - 1) + ½\mathrm{sinc}\,(ft + 1)]\,. \tag{5.10}$$

We plot the convolving functions $|\hat{w}_A|^2$ and $|\hat{w}_B|^2$ in Figure 4, only for $f \geq 0$, along with the sinc$^2(fT)$. In the plots I have scaled the function to unity at $f = 0$ for easy of comparison.

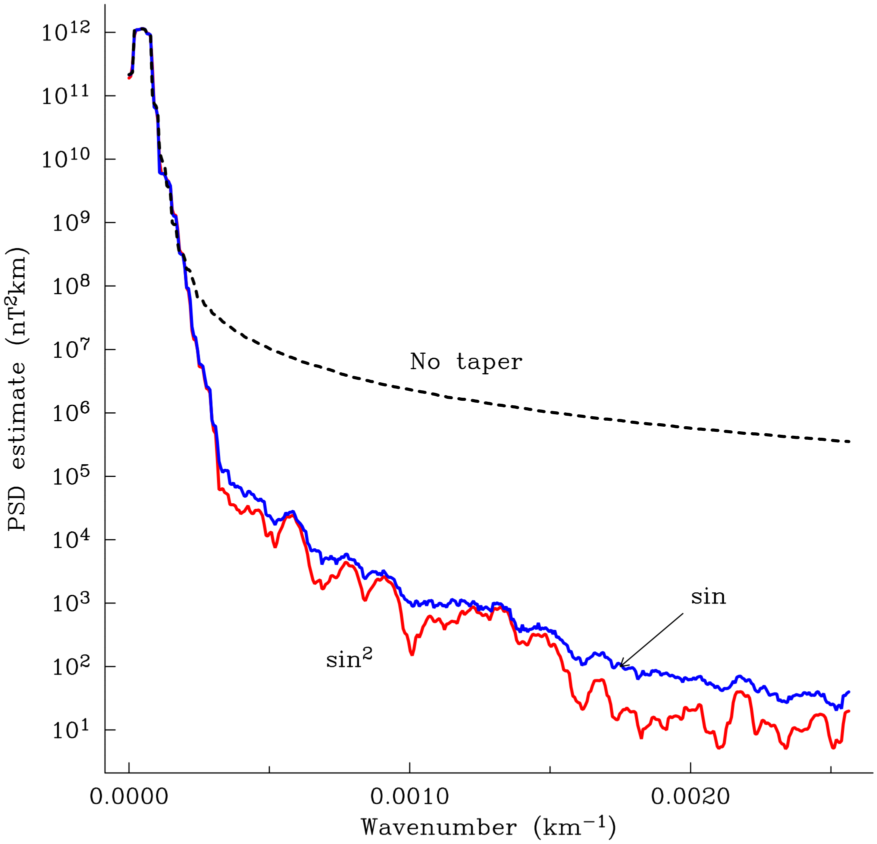
Figure 4: Top panel. Three tapers, box-car (or no taper) gray; sine taper, dashed; sine-squared taper black. Lower panel. Corresponding convolving weight functions with frequency.



The graphs in the lower panel show on a log scale how much more rapidly the smoother tapers decay away. In that respect, the convolving function $W(f)$ is a better approximation to a delta function, because the functions corresponding to the sine taper are nearly zero away from the center; and the quality is even better for the sine-squared taper. But notice that the central peak is wider than that of the sinc-squared for both functions, so in this respect $|\hat{w}(f)|^2$ is a poorer approximation. Suppression of leakage from large peaks into low-amplitude parts of the spectrum turns out to be **much** more important than the loss of resolution introduced by this factor. We have already had to sacrifice resolution by averaging for the improvement in variance, a sacrifice well worth making.

How effective in practice is the introduction of a taper like w_A or w_B? For many spectra the reduction in bias gives astonishing results, and that is the case for the Magsat fields. In Figure 5, I show the spectra estimates that result from tapering with w_A, the sine taper, and w_B the sine-squared taper. What the Figure reveals is that for $k > 0.0005\,\mathrm{km}^{-1}$ spectral leakage in the periodogram estimate has artificially raised the level of the PSD by **3 to 4 orders of magnitude.** Admittedly this is a spectrum with a gigantic dynamic range. None-the-less, the smoothed periodogram totally misrepresents the field behavior at shorter wavelengths.

Figure 5: PSD estimates of Magsat data for smoothed periodogram and with two tapers w_A, the sine taper, and w_B the sine-squared taper.



Examples like this are easy to find, though not usually so dramatic as this particular example.

Our next topic is a brief description of the modern approach to spectral estimation, in which tapers play a central role, not only in bias suppression but also, strange to say, in variance reduction also.

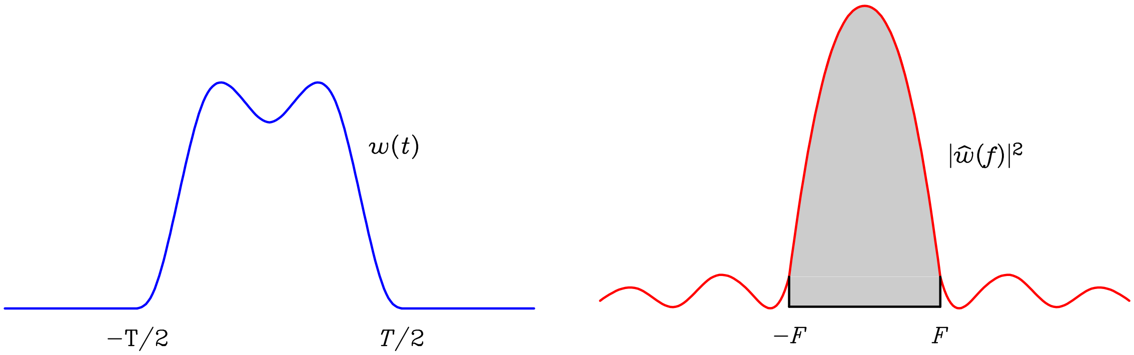
6. The Perfect Taper

The idea of tapering prompts us to ask the following deceptively simple question: What function $w(t)$, vanishing outside the interval $(-T/2, T/2)$, has the greatest concentration of energy in its Fourier transform? I have turned to continuous time for this discussion, because it is simpler to illustrate the ideas; of course there is an entirely parallel theory for discrete processes, which is the theory we need for practical calculations. Recall in (5.4) the continuous model was used to compute the approximation $|\hat{w}(f)|^2$ for convolving kernel $W(f)$ in (5.3). I have also shifted the time origin to the center of the interval. Here I am drawing on, but not exactly following, Percival and Walden pp 75-116. The question is still too vague to be answered, but we make it more precise by defining a measure of **spectral concentration**:

$$C[w] = \frac{\displaystyle\int_{-F}^{F} |\hat{w}(f)|^2 \, df}{\displaystyle\int_{-\infty}^{\infty} |\hat{w}(f)|^2 \, df} \,. \tag{6.1}$$

We know the original signal is limited between $-T/2$ and $T/2$; so we **choose** a bandwidth F and ask that the taper w have maximum concentration for that F; clearly from (6.1), no w can make $C[w]$ exceed unity. It makes no sense to select F as small as $1/T$, since that's the lowest

Figure 6: The integral for concentration.



frequency accessible from a record of length T. The idea is to pick ahead of time the interval of frequency averaging we would be willing to settle for, and then to find the best taper for that problem. Observe in Figure 4 how as the spectral leakage improves, the width of the region averaged near $f=0$ gets broader.

Suppose we pick the bandwidth $F = 1/T$. How well do the three tapers we have looked at so far perform according to the measure in (6.1)? This is of course just an exercise in integration. The answers are 0.902 for no taper at all, 0.97 for the sine taper, and 0.918 for the sine-squared taper. Somewhat surprisingly perhaps the sine taper is the best in this case. If we choose $F = 4/T$ the numbers for $C[w]$ are 0.9748, 0.99973, and 0.999986; if we are willing to average over this bandwidth, the sine-squared taper is clearly superior.

How is this problem of maximum concentration solved? Suppose, instead of normalizing by the total power, we simply set it to unity as a side condition. Then we introduce a Lagrange multiplier ν and look for stationary points of the functional

$$U[\hat{w}] = \int_{-F}^{F} \hat{w}(f)\,\hat{w}(f)^*\,df - \nu \int_{-\infty}^{\infty} \hat{w}(f)\,\hat{w}(f)^*\,dt\ . \tag{6.2}$$

To solve this we need to introduce the parent function $w(t)$, whose Fourier transform is $\hat{w}(f)$, along with the fact that w vanishes outside $(-T/2, T/2)$. We use Parseval's Theorem for the second term in (6.2):

$$\int_{-\infty}^{\infty} \hat{w}(f)\,\hat{w}(f)^*\,df = \int_{-T/2}^{T/2} w(t)^2\,df\ . \tag{6.3}$$

Then we insert (6.3) and the definition of $\hat{w}$, namely,

$$\hat{w}(f) = \int_{-T/2}^{T/2} \mathrm{e}^{-2\pi \mathrm{i} ft}\,w(t)\,dt \tag{6.4}$$

into equation (6.2):

$$U = \int_{-F}^{F} df \int_{-T/2}^{T/2} \mathrm{e}^{-2\pi \mathrm{i} ft}\,w(t)\,dt \int_{-T/2}^{T/2} \mathrm{e}^{2\pi \mathrm{i} ft'}\,w(t')\,dt' - \nu \int_{-T/2}^{T/2} w(t)^2\,dt \tag{6.5}$$

$$= \int_{-T/2}^{T/2} dt \int_{-T/2}^{T/2} dt'\,w(t)\,w(t') \int_{-F}^{F} \mathrm{e}^{-2\pi \mathrm{i} f(t-t')}\,df - \nu \int_{-T/2}^{T/2} w(t)^2\,dt \tag{6.6}$$

$$= \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} w(t)\,w(t')\,\frac{\sin 2\pi F(t-t')}{2\pi(t-t')}\,dt\,dt' - \nu \int_{-T/2}^{T/2} w(t)^2\,dt\ . \tag{6.7}$$

This equation is a Hilbert space equivalent of the one we encountered at the beginning of the year for the principal axes of inertia, to maximize the moment of inertia, I. In abstract notation (6.7) is

$$U[w] = (w, Kw) - \lambda(w, w) \tag{6.8}$$

for the linear operator K on $L_2(-\frac{1}{2}T, \frac{1}{2}T)$ where

$$(K\,w)(t) = \int_{-T/2}^{T/2} \operatorname{sinc}(2(t-t'))\, w(t')\, dt' \,. \tag{6.9}$$

The stationary points of (6.8) are obtained by differentiating with respect to w. (This is Gateaux differentiation, which we will see in inverse theory). The result for (6.8) is that we seek the solution to the eigenvalue equation:

$$Kw = \lambda w \tag{6.10}$$

which is explicitly:

$$\int_{-T/2}^{T/2} \frac{\sin 2\pi F(t-t')}{2\pi(t-t')}\, u_n(t')\, dt' = \mu_n\, u_n(t), \quad | t | \le T/2, \quad n = 0, 1, 2, \cdots \tag{6.11}$$

for eigenvalues μ_n and the corresponding eigenfunctions $u_n(t)$. **The eigenvalues of the system corresponds to concentration factors C in (6.1) for the appropriate eigenfunction.** The largest eigenvalue, μ_0, gives the largest concentration, and the corresponding eigenfunction u_0 is the optimal taper for the specified values of F and T.

Before describing some of the remarkable properties of the solutions to (6.11), it helps to make a change of variables to remove the apparent dependence on the two parameters T and F; as you can imagine, the complete family of solutions is parameterized by a single dimensionless number. Let $x = 2t/T$, $y = 2t'/T$ and $p = FT$; then (6.11) becomes

$$\int_{-1}^{1} \frac{\sin \pi p(x-y)}{\pi(x-y)}\, \psi_n(y)\, dy = \mu_n\, \psi_n(x), \quad | x | \le 1 \,. \tag{6.12}$$

and $u_n(t) = \psi_n(2t/T)$. This is almost equation (33) of Percival and Walden, in a slightly more readable notation.

When you decide on a bandwidth F, this fixes p, which is called the **time-bandwidth product** of the system under study; for practical problems we always choose $p > 1$, because, as we mentioned, you cannot expect to get good concentration into a frequency band narrower than $1/T$. Then it can be proved there are infinitely many distinct, real, positive eigenvalues:

$$1 > \mu_0 > \mu_1 > \mu_2 > \cdots \tag{6.13}$$

and as $n \to \infty$, $\mu_n \to 0$. Since the concentration function $C[\psi_n] = \mu_n$, the first eigenfunction, ψ_0 is the best performing taper for the particular value of p in question; we will soon see the other eigenfunctions in the sequence have a role in estimation too. This is partly because the eigenfunctions (and of course the corresponding tapers) are mutually orthogonal on $(-1, 1)$:

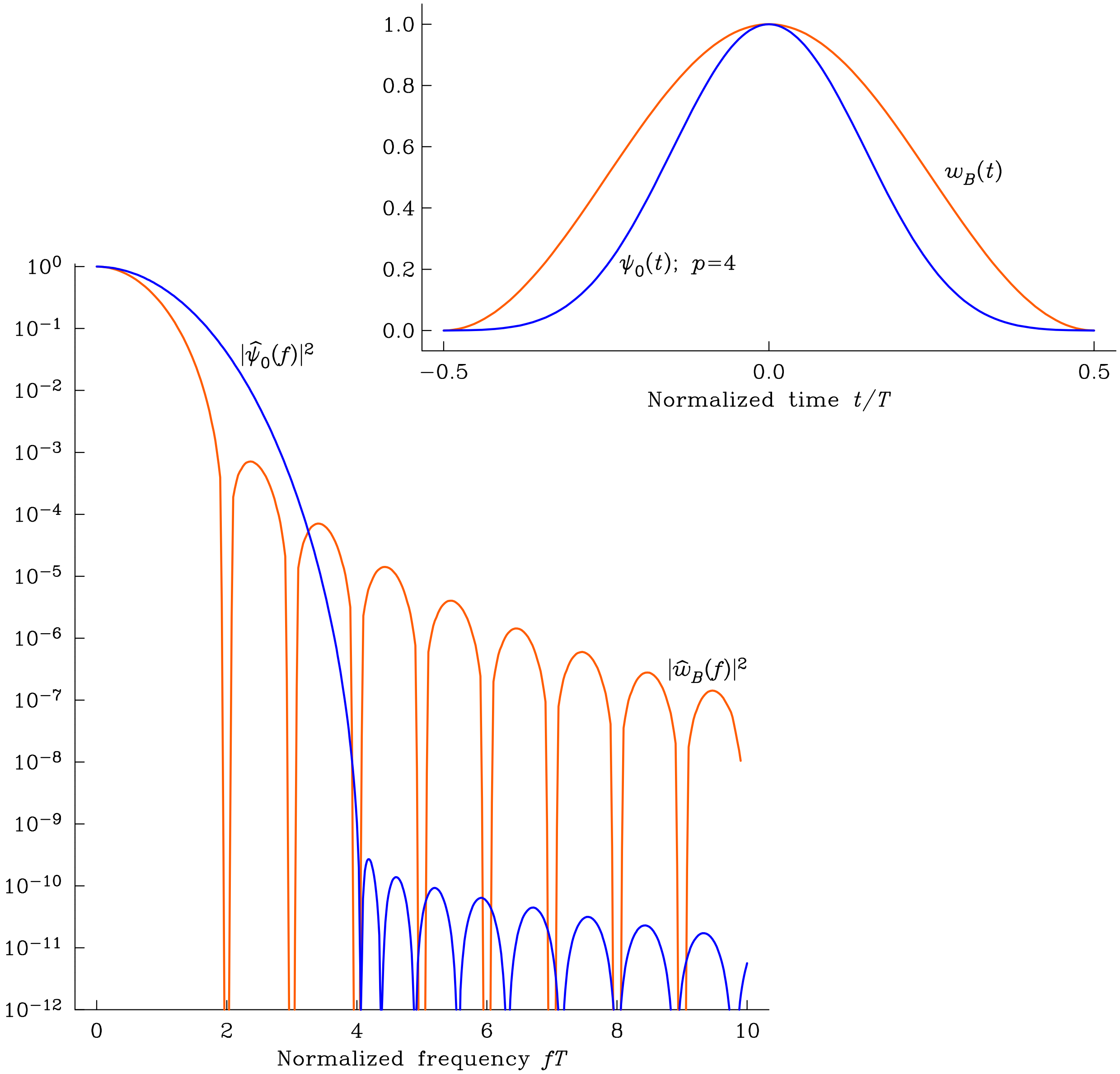
$$\int_{-1}^{1} \psi_m(x)\,\psi_n(x)\,dx = 0, \quad m \neq n \tag{6.14}$$

which is a general property of self-adjoint eigensystems like (6.12). More unusual, and harder to prove, is the following fact about the eigenvalues: for $n \leq 2p$, the eigenvalues μ_n (and hence the spectral concentrations) *are all very close to unity*, then fall suddenly to very small values. This means for time-bandwidth product p there are about $2\,p$ very good tapers, that have strong rejection outside the desired band. There are hosts of other properties – Percival and Walden list eight on pp 79-80; but they don't mention in that list where the functions ψ_n get their name, which is **prolate spheroidal wavefunctions**. A simpler name has been suggested: **Slepian functions,** after David Slepian who invented this application of them. The self adjoint integral operator in (6.12) commutes with a certain second-order differential operator describing wave motion in prolate spheroidal coordinates! Commuting operators share eigenfunctions, hence the name. And in fact, computation of ψ_n is greatly facilitated by this commutation.

How good is the perfect taper? Let us study the case $p = 4$, which you will recall corresponds to $F = 4/T$, and for which we obtained $C[w]$ of 0.999986 for the best, sine-squared taper. The largest eigenvalue of (6.11) with $p = 4$ is $0.99999\,999942$, that is nine 9s! Figure 7 shows the optimal taper ψ_0 for the time-bandwidth product, $p = 4$, along with the sine-squared taper. Below are the squared Fourier transforms of these $|\hat{w}|^2$. Observe how the weight function for ψ_0 is nearly two orders of magnitude smaller than the one for sine-squared in the rejection band, and larger in most of the pass band. It is one of those remarkable facts that $\hat{\psi}_0$, the Fourier transform of ψ_0, is equal to a constant $\times \psi_0(2\pi f/p)$, that is, a stretched version of the original function; when the argument of the stretched function is outside $(-1, +1)$, we simply use the left side of (6.12) to extend it! And the relationship is the same between ψ_n and $\hat{\psi}_n$.

We have looked at only the continuous-time/continuous-frequency theory, which first appeared in the 1980s and was invented by Slepian. All the results, including our (6.1) which motivates the whole idea, are only approximations for a true discrete and finite time series. But an exact theory precisely corresponding to the optimization we have just discussed can be carried out – unfortunately it introduces another parameter (the number of points in the time record) over which the family of functions varies, when we want the absolute best out of the theory. The people who have developed these ideas such as David Thomson (and Percival and Walden) seem to believe in the principle that the more subscripts and superscripts you hang on a variable, the clearer the notation. This is an error: the truth is, *The utility of a mathematical notation is proportional* *to the amount of information it* **hides**. So the treatment of Thomson is generally agreed to be hard to read, and Percival and Walden are no better. We will not go into the thicket of the optimal discrete prolate spheroidal sequences as they are called – it is enough that you understand the general idea.

Figure 7: Best taper and its FT for $p=4$; also shown, performance of $w_B(t)$, the sine-squared taper.



7. Spectral Estimation: Multitapers

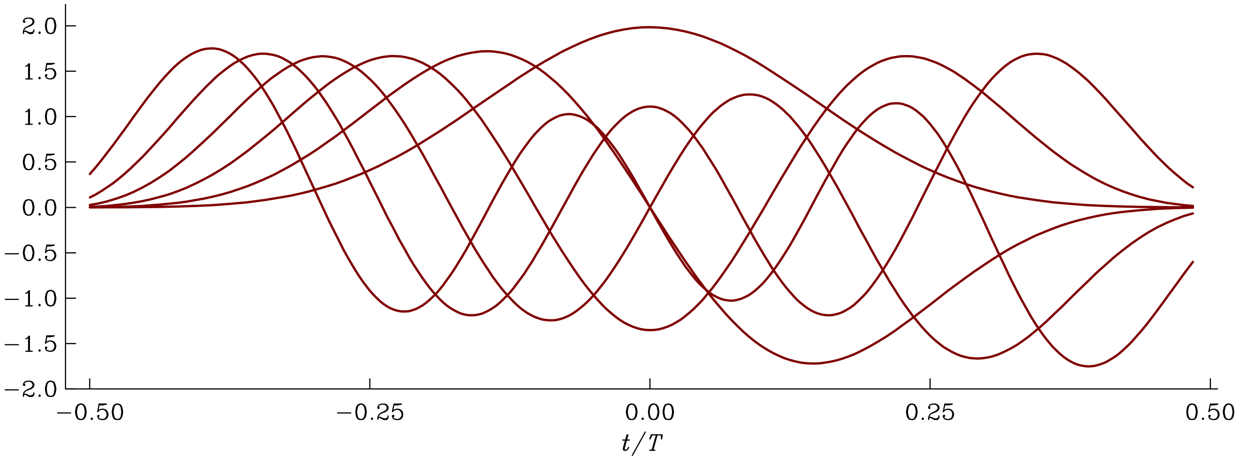
At this point we have focussed on the taper with the best concentration, $\psi_0(t)$, with eigenvalue μ_0. But the theory tells us there are infinitely many eigenfunctions for K associated with orthogonal tapers. How good are the eigenvalues for these functions? The Table lists those for $p = 4$.

n	Eigenvalue μ_n
0	0.99999999942487
1	0.99999997246259
2	0.99999878976974
3	0.99996755459638
4	0.99941008235158
5	0.99250455019311
6	0.93665243143508
7	0.69883581857698
8	0.29937483065771
9	0.06424183042118

The spectral concentration of these tapers is excellent out to $n = 4$ at least. It is fair to say that these tapers are nearly as good as the optimal taper, and yet we apparently have no use for them. David Thomson (op. sit.) observed, however, that *if a set of tapers is orthogonal, the spectral estimates made with them are uncorrelated.* Thus we can made separate estimates with each taper, and average them together, in this way reducing the variance by averaging over the independent estimates. This is the method of **Multitaper Estimation.**

To summarize: a series of tapers is computed, based on the preassigned bandwidth F of interest; then a tapered periodogram is made for each one and they are averaged together. How is the number of tapers K chosen? There are a number of recipes, some based on estimating the

Figure 8: The first six eigentapers with $p = 4$.

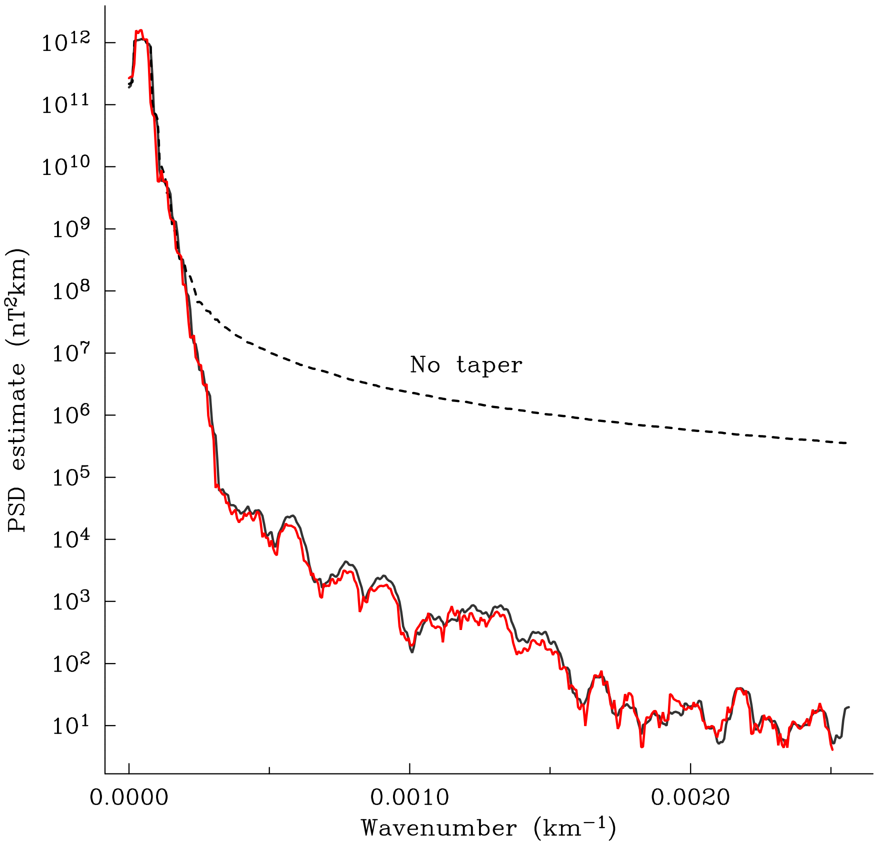


variance improvement by adding a further term, but by and large the answer is usually around $1.2p$, a factor times the time-bandwidth product, because after this number the concentration falls off so rapidly. Figure 8 shows the first six tapers for the time bandwidth product $p=4$. Notice how the higher order tapers are applying more emphasis to values near the ends of the series, thus overcoming an old objection that such data were being accorded insufficient weight in the single taper theory.

In Figure 9 we see the result of a multitaper estimation on the Magsat data. Here, after some experimentation, I used a time-bandwidth product $p = 6$ (this gives $F = 0.000038\,\mathrm{km}^{-1}$) and five tapers. The result is barely distinguishable from the sine-squared taper (shown dashed) even though the spectral concentration of the fifth taper is 0.9999999826. We can conclude from this that spectral leakage is not a problem for the sine-squared taper in this case, and that are we indeed looking at the true spectrum in the low-amplitude band in the graph. It may be disappointing that the multitaper estimate does not outperform the simple sine-squared taper here, but there will be situations where it does.

As you will appreciate by consulting Percival and Walden there is a great deal I have not covered. For example, the details of how to calculate

Figure 9: Magsat PSD estimates with Thomson multitapers, $p=6$ and 5 tapers, shown red; sine-squared taper grey.



the tapers: there is a very clever idea that allows the solution of the matrix eigensystem to be formulated into a three-term recurrence scheme (like the one used for computing spherical harmonics) which makes practical the use of these Thomson tapers even for very long time series – the computation time only rises as N, the number of points, not N^3 as a naive approach would give. The Matlab Signal Processing Toolbox provides code for computing the Slepian functions. I have Fortran for them. They are also included in my spectral estimation program *PSD*.

The Thomson/Slepian multitapers are without doubt the best way to estimate spectra when there is very large dynamic range, such as a sharp fall-off or a strong peak (true spectral lines should **always** be removed separately before spectral analysis begins). But because of the need to choose an averaging bandwidth that is fixed across the whole spectrum, the averaging may be too severe in one frequency interval (where the spectrum is varying rapidly), and not enough in other parts (where the spectrum is relatively flat). A completely different multitaper method was given by Riedel and Sidorenko (*IEEE Trans. Sig. Proc.*, 43, pp 188-195, 1995) which we discuss next.

8. Local Bias Minimization

When one tapers the time series, one convolves the true spectrum with a function that essentially averages the local behavior, smoothing it to some extent. If there are peaks, or troughs, this introduces a **local bias,** as Figure 10 below illustrates. When the spectrum has a moderate dynamic range, so that spectral leakage is not too significant, the local bias can be a problem. We can ask, What tapers will minimize the local bias? Of course we need a measure. Riedel and Sidorenko (1995) use a quadratic approximation as follows. The bias β is the difference between the expected value and the true value of the PSD:

$$\beta = \mathcal{E}\,[\hat{S}(f_0)] - S(f_0) = \int_{-½}^{½} S(f)\, W(f - f_0)\; df \; - S(f_0) \tag{8.1}$$

where W is the convolving function; see (5.3)-(5.4). The area under $W(f)$ is always chosen to be unity, and hence (8.1) can be written

$$\beta = \int_{-½}^{½} [S(f) - S(f_0)]\, W(f - f_0)\; df\,. \tag{8.2}$$

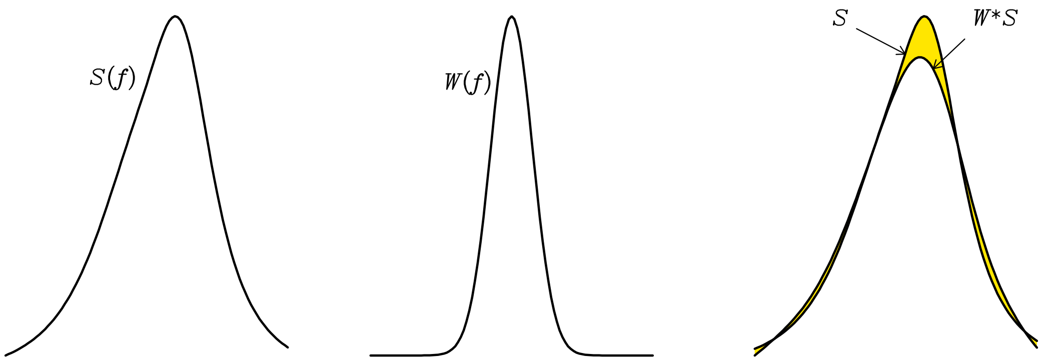
For practical tapers, W also dies away to zero fairly fast; R&S then say we can approximate the factor in brackets with a local Taylor series:

$$S(f) - S(f_0) = (f - f_0)\, S'(f_0) + ½(f - f_0)^2\, S''(f_0) + O(f - f_0)^3\,. \tag{8.3}$$

If we substitute (8.3) into (8.2) and integrate, because $W(f)$ is always an even function of f the odd derivative terms vanish and we obtain the approximation:

$$\beta = \int_{-½}^{½} ½ S''(f_0)\, W(f - f_0)\, (f - f_0)^2\; df \tag{8.4}$$

Figure 10: Smoothing of true PSD introduces bias. The convolution flattens peaks and widens the flanks.



and the error depends on $S^{iv}(f_0)$. This is only valid if $S''(f_0) \neq 0$ or course. You can see in Figure 10 how the discrepancy between the true S and $W * S$ is greatest where the second derivative of the PSD is largest in magnitude, just as predicted by (8.4).

Now we can proceed to ask for the time-limited taper $w(t)$ that minimizes $\beta[w]$, just as Slepian minimized $C[w]$. We omit the details, and simply note that a similar eigenvalue problem is produced whose eigenfunctions are a family of orthogonal functions, just as the prolate taper functions are. Notice, that here there is no band-width parameter, like F, to choose. But an amazing thing happens: to a remarkable degree of approximation, **those orthogonal functions are the sines.** In continuous time

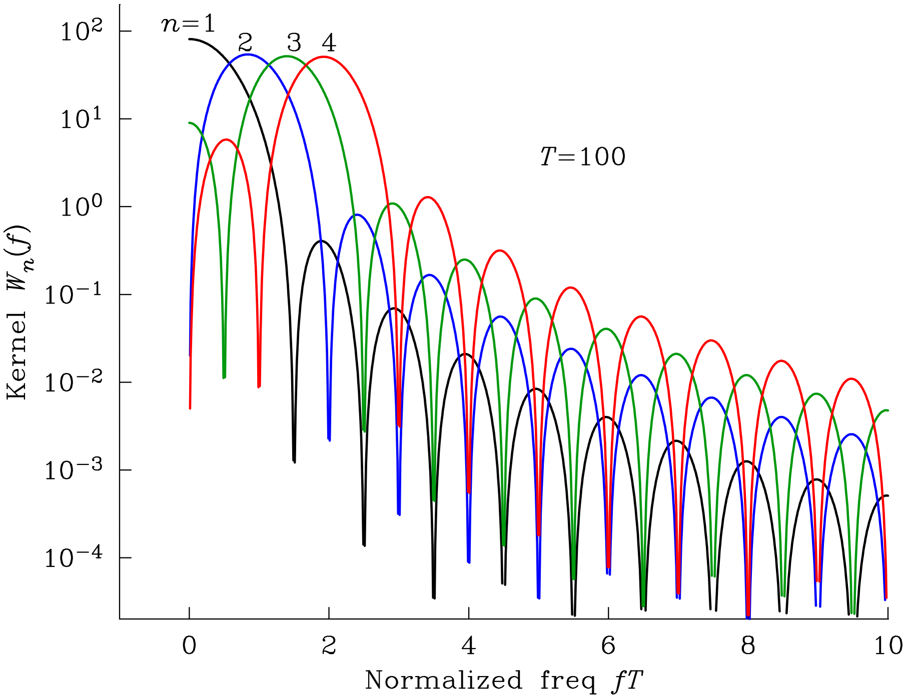
$$\phi_n(t) = \left(\frac{2}{T}\right)^{1/2} \sin \frac{\pi n t}{T}\,, \quad 0 \leq t \leq T\,. \tag{8.5}$$

These simple functions are an orthogonal set that can be used together to form estimates of the spectrum, and by averaging the statistically independent estimates we can reduce variance. The remaining question is, How many of the ϕ_n should be used?

Before answering that question, let us plot the convolving functions in frequency corresponding to the eigentapers in time. Recall

$$W_n(f) = |\hat{\phi}_n(f)|^2\,. \tag{8.6}$$

Figure 11: Several convolving kernels for sine tapers.



We find the following simple result:

$$W_n(f) = \frac{2n^2 T}{(n+2fT)^2} \operatorname{sinc}(fT - \tfrac{1}{2}n)^2 \,. \tag{8.7}$$

Thus when $fT \gg n$ the kernel decays like $n^2/2\pi^2 T^3 f^4$ times a squared sine function. The peak of $W_n(f)$ is roughly at $f = n/2T$ for $n > 1$ and is at $f = 0$ when $n = 1$. These functions are illustrated in Figure 11, best viewed in color. Unlike the Slepian-Thomson tapers, which attempt to concentrate inside the bandwidth F, these functions spread out over a wider and wider frequency band as n increases, because there is no fixed scale parameter corresponding to F. We obviously have a much poorer rejection of spectral leakage.

Suppose we consider averaging together K spectral estimates, based on the first K tapers in (8.5). Then R&S show by integrating (8.4) with the W_n inserted that approximately

$$|\beta| = \frac{|S''(f_0)|\, K^2}{24T^2} \tag{8.8}$$

so that the bias increases as the square of the number of tapers included. On the other hand the variance decreases like $1/K$ because we are averaging independent estimates We find

$$\operatorname{var}[\hat{S}(f_0)] = \frac{S(f_0)^2}{K} \,. \tag{8.9}$$

We would like to keep both of these undesirable properties small, so R&S suggest looking for the minimum value of a linear combination:

$$L = \beta^2 + \operatorname{var}[\hat{S}(f_0)] \tag{8.10}$$

which is called the **mean square error** or MSE in statistics. Here is a simple calculus problem: What value of K makes L smallest?

$$L = \frac{S''(f_0)^2 K^4}{576T^4} + \frac{S(f_0)^2}{K} \,; \quad \frac{dL}{dK} = \frac{S''(f_0)^2 K^3}{144T^4} - \frac{S(f_0)^2}{K^2} \tag{8.11}$$

which gives

$$K_{opt} = \left(\frac{12T^2 S(f_0)}{|S''(f_0)|} \right)^{2/5} \,. \tag{8.12}$$

This formula provides a way of estimating the best number of tapers to average **at each frequency.** Where the spectrum is smooth, we take a lot of tapers and beat down the variance as much as possible; where there are narrow peaks or troughs, we sacrifice variance for good resolution by using only a few tapers. There is no need to guess a suitable value for the bandwidth of smoothing that best suits the spectrum, as required with Slepian multitapers. There are two problems, however.

First we don't really know S or S'' at any frequency, and so we cannot calculate K_{opt}; this is a chicken-and-egg problem. We solve it by simply guessing some value for the number of tapers to be used at all frequencies, as first estimate. The pilot estimate is then used in (8.12). Since the estimates are noisy, we will need to smooth some more to get a reasonably reliable approximation for S''; but we know from K_{opt} a local band-width over which the spectrum should be relatively smooth, so this provides a guide for the range of smoothing needed. With the new K_{opt} we find another estimate of S, and the process can be repeated. In practice things settle down in two or three steps, but there is no proof of convergence that I know of.

The second difficulty is not easily overcome in a pleasing way. The whole theory depends on $S''(f_0) \neq 0$. When the second derivative vanishes, the bias calculation is invalid, and the number of tapers predicted by the theory is infinite. Because the spectral estimates are noisy, $S''(f)$ passes through zero quite often just through random variations. At present my code (called PSD), looks for runaway growth in K_{opt} and limits the increase as a function of f. This ad hoc process seems to work quite well, but it would be better if there were a more defensible theoretical approach.

There are of course a lot of other details to be worked out in the creation of a useful, reliable espectral estimate procedure. For example, I haven't mentioned **prewhitening** except in passing: see Section 9 for more about this useful idea. The sine multitaper approach seems to provide a very convenient way of performing spectral estimation, because it doesn't require the user to guess various parameters, like F or the number of sections to be averaged, if you use Welch's method. It is very fast for a reason I haven't mentioned yet: Because the tapers are all sine functions, *you need only Fourier transform the original time series once!* Unlike every other method, that requires many FTs, here the different estimates can be made simply by combining the Fourier coefficients found by the FFT in different ways. While with today's computers speed is often not an issue, it is still convenient with very long records to get an answer in less than a second, while the prolate method makes you wait a lot longer. The prolates still win if there is a very large dynamic range, or for really short series, however.

References

Percival, D. B., & Walden, A. T., Spectral analysis for physical applications - Multitaper and conventional univariate techniques, Cambridge, 1993.

Priestley, M. B., Spectral Analysis and Time Series, Academic Press, New York, 1981.

Rice, J. A., Mathematical statistics and data analysis, Brooks-Cole Pub. Co., Monterey, CA, 1988.

Riedel, K. S., & Sidorenko, A., Minimum bias multiple taper spectral estimation, IEEE Trans. Sig. Proc., 43, 188-195, 1995.

Thomson, D. J., Spectrum estimation and harmonic analysis, *Proc. IEEE,* v 70, pp 1055-96, 1982.

9. Prewhitening

Most geophysical spectra with large dynamic range are red, meaning that the energy is concentrated at the low frequencies, often falling as an inverse power or exponentially. The Slepian-Thomson tapers can provide excellent suppression of spectral leakage to the higher frequencies, but they often distort the low-frequency spectrum because of the flattening due to the (relatively broad) convolution introduced. On the other hand, the sine-multitapers may do a reasonably good job at the low end, but spectral leakage can corrupt the higher frequency estimates because the sines are not very good at leakage protection. One technique particularly helpful in these circmstances is **prewhitening**.

The idea, which we have mentioned briefly before (Notes, Chap 3, Section 2), is to design a really simple filter arranged to have a response so that upon application to the original record, the output series is nearly a white process. The original series is filtered in the time domain (usually with a convolution filter) and a spectrum estimated for that series, then the effect of the filter is undone in the frequency domain. Here are the details.

We set up a model in which there are k weights from which the stocahstic process X_n is generated autoregressively from white noise Y_n via:

$$X_n = Y_n + a_1 X_{n-1} + \cdots a_k X_{n-k} \tag{9.1}$$

Given a process X_n, we wish to recover the unknown weights a_k in this model: we multiply (9.1) through by X_j and take the expectation:

$$\mathcal{E}\,[X_j X_n] = \mathcal{E}\,[X_j Y_n] + a_1 \mathcal{E}\,[X_j X_{n-1}] + \cdots a_k \mathcal{E}\,[X_j X_{n-k}] \tag{9.2}$$

Now a truly white noise Y_n is uncorrelated with anything except itself at zero lag, so the first term on the right vanishes. The rest of the terms are given by autocovariances of the process:

$$R_X(j-n) = \sum_{m=1}^{k} R_X(j-n+m)\, a_m \tag{9.3}$$

If we perform this on for $j-n = 1, 2, \cdots k$ we have a square linear system of equations for the unknown weights. These are called the **Yule-Walker equations**. But wait a minute: we don't know the $R_X(j)$s! In a practical scheme we make a pilot estimate of the PSD S_X with a guess for the number of sine tapers, then take its digital FT for preliminary values of $R_X(j)$ with which we can solve (9.3). It will not be a large system, since typically $k < 10$.

Next we use the weights a_m to filter the original series X_n as follows: we rearrange (9.1) to give Y_n:

$$Y_n = X_n - \sum_{m=1}^{k} a_k X_{n-k} \tag{9.4}$$

which applies to X_n a simple convolution filter of length $k+1$. The series

Y_n is called the **prewhitened** version of the original series X_n. If the idealized model (9.1) were exact, Y_n would be a white noise but (9.1) is only an approximation, maybe not even a very good one. Nonetheless, it is certain that when the original X_n had a red spectrum, the corresponding Y_n will have a less severe concentration at low frequency in its PSD, and then spectral leakage will be a less serious problem.

How is the spectrum $S_Y(f)$ related to $S_X(f)$? That was answered in Section 4 of Chapter 2: from (9.4)

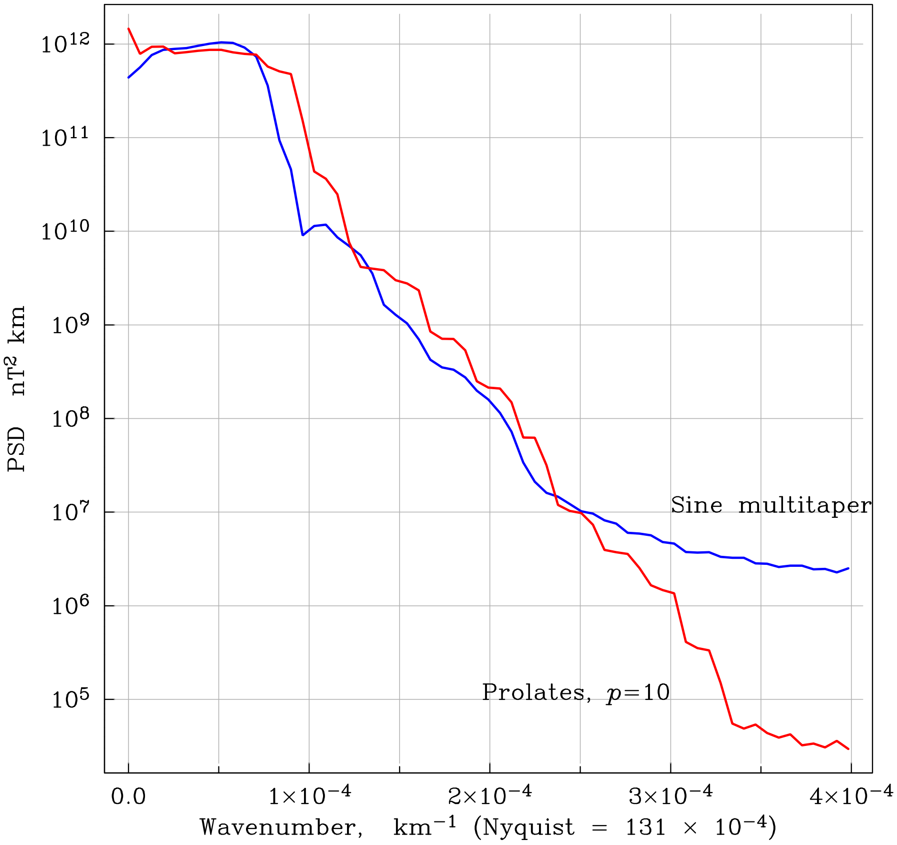
$$S_Y(f) = |\hat{a}(f)|^2 S_X(f) \tag{9.5}$$

$$= \left| 1 - \sum_{m=1}^{k} a_m \, \mathrm{e}^{-2\pi imf} \right|^2 S_X(f) \tag{9.6}$$

the discrete version of the famous result. So finally, we make a PSD estimate on the prewhitened record and obtain $\hat{S}_Y(f)$, then we rearrange (9.6) to give the spectral estimate of the original series:

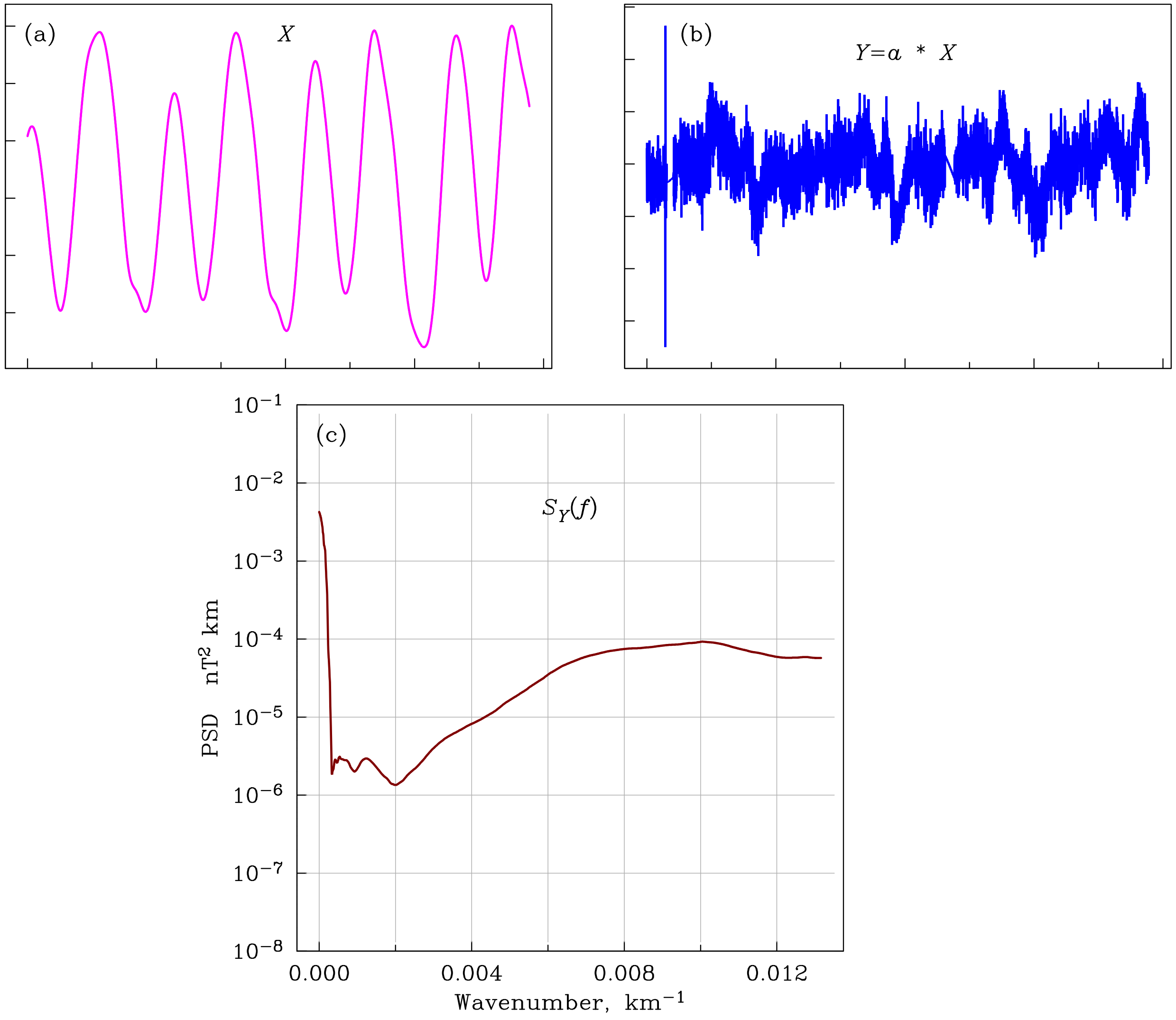
$$\hat{S}_X(f) = \frac{\hat{S}_Y(f)}{\left| 1 - \sum_{m=1}^{k} a_m \, \mathrm{e}^{-2\pi imf} \right|^2} \tag{9.7}$$

Figure 12: Low end of PSD for Magsat field data.



We conclude with an example. Let us return to the Magsat total field magnetic data of Section 5. In Figure 12 we see the lowest wavenumber portion of the PSD according to two estimates, the prolate mutitapers with time-bandwidth 10, and the adaptive sine multitaper method. Both suggest a rather flat PSD near $k = 0$, which I assert is due to bias. Notice spectral leakage is apparent in the sine multitaper estimate. Using one of these estimates we find a set of just 4 weights by solving the Yule-Walker equations, and apply that series as a filter to X. The results appear in Figure 13, where I've plotted the original series and the prewhitened record. The change is remarkable: with just four weights the original data series, which is smooth, almost sinusoidal in appearance, is converted into an almost random series. To the eye it is not a white noise perhaps, but the advertised "whitening" has clearly been effective. Something very obvious to the eye is the spike near the beginning of the prewhitened series, which is followed by several low-amplitude values.

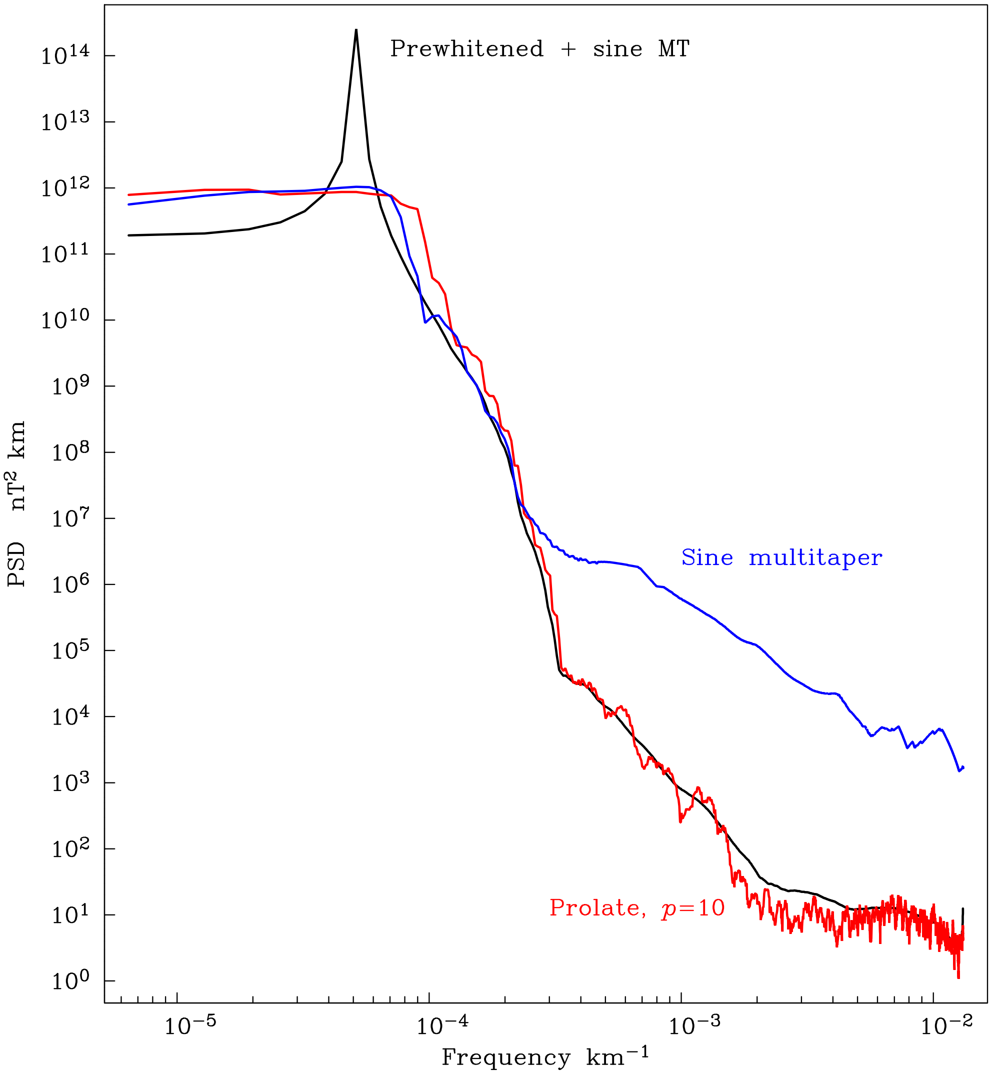
Figure 13: (a) Original time series; (b) Prewhitened series with 4 coefficients; (c) PSD of prewhitened record



This is apparently a data glitch that has been incompletely edited, something entirely invisible in the original record. Prewhitening is very good way of inspecting data to look for departures from regularity and stationarity.

Next we take the sine multitaper spectrum of the prewhitened record, which is shown on the previous page. The PSD is very far from white. But the dynamic range, while it is four orders of magnitude, is still enormously less than the 11 orders of magnitude exhibited by the estimated spectrum of X shown in Figure 9. Finally, we divide the prewhitened spectrum to form the estimate $S_Y(f)/|\hat{a}(f)|^2$ plotted below. To show the whole spectrum, which covers a huge range in PSD and frequency, I have used log-log scales now. Observe how the sine multitaper estimate of the prewhitened series is just as good as the prolate multitaper estimate in avoiding the spectral leakage from low wavenumbers, but

Figure 14: PSD estimates on a log-log scale.



it offers much smaller variance. At the lowest wavenumbers we see a new peak, of almost 3 orders of magnitude, emerging where the other two estimators gave as a very flat spectrum.

GEOPHYSICAL DATA ANALYSIS

Robert L. Parker

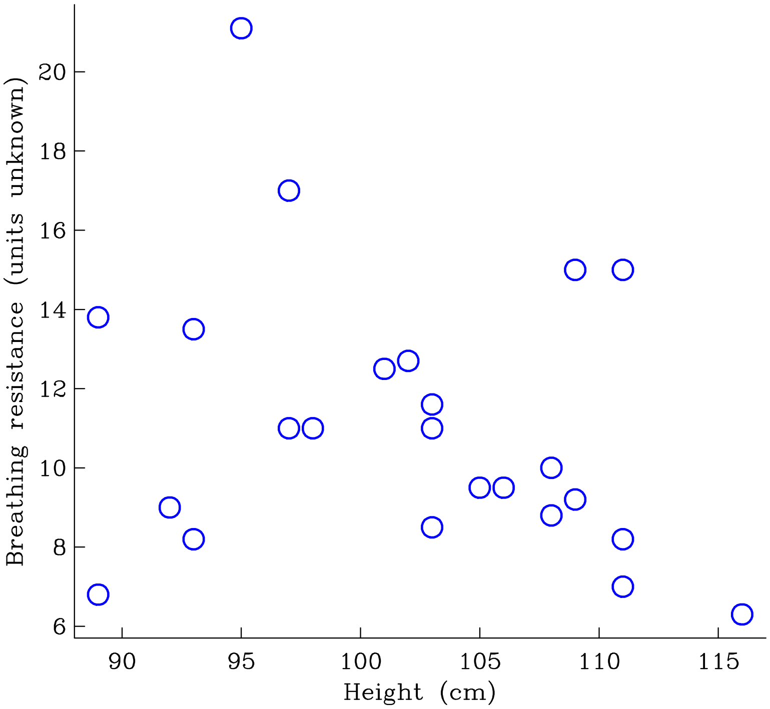
CHAPTER 4: MULTIVARIATE and MULTIDIMENSIONAL SPECTRA

1. Random Data Pairs

In geophysics and every other branch of physical science we encounter time series in pairs that are related to one another. Examples: the vertical ground motion, and the output voltage of a seismometer; two components of a vector field (say the vertical and the horizontal) as the instrument measures values along a path in space; the topographic height and the observed strength of gravity as seen along a profile (here time has been replaced by distance); the north electric field and the east magnetic field at a site. You can easily think of dozens more yourself. One of the characteristics of the apparently random geophysical list above is that in every case one can discover a theory capable of predicting the relationship between the signal pairs, and a practical need for a data analysis technique to estimate the quantitative behavior. We will consider both these aspects

First, however, let us briefly consider how pairs of variables are traditionally analyzed when there is no independent variable, like time, against which the system is evolving. A first example (from John Rice's book *Mathematical Statistics and Data Analysis*, 2nd ed, Duxbury Press, 1995) concerns breathing resistance in 24 children with cystic fibrosis tabulated as a function of their heights; see Fig 1. We plot this pair of

Figure 1: Scatter plot of medical data.



parameters as a point on a graph, (called a **scatterplot**) and if we have enough data we may see a pattern emerging: very roughly, a taller child has a lower breathing resistance. In this case there is no theoretical model, just a statistical tendency. We don't expect a perfectly linear relationship, but it would be nice to have a measure of common tendency, and a test to say if a given sample really exhibits such a tendency or not. (Biologists and social scientists, who usually lack predictive theories, depend on this measure a great deal.)

In cases of relations like this we would like to measure the degree of association of one variable with another. The commonly used one is the **correlation coefficient** ρ, or its square. For a pair of random variables X and Y, jointly distributed we can calculate:

$$\rho_{XY} = \frac{\mathcal{E}\,[(X - \mathcal{E}\,[X])(Y - \mathcal{E}\,[Y])]}{\sqrt{\mathrm{var}\,[X]\,\mathrm{var}\,[Y]}} \tag{1.1}$$

where $\mathcal{E}$ is the expectation, and var is the variance. That is something you can calculate for a theoretical distribution like a joint Gaussian. The numerator you will recognize as your old friend the **covariance** between X and Y. A natural estimator of ρ is:

$$\hat{\rho}_{XY} = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\left(\sum_j (x_j - \bar{x})^2 \sum_k (y_k - \bar{y})^2 \right)^{1/2}} \tag{1.2}$$

where x_j and y_j are the N data samples and $\bar{x}$ and $\bar{y}$ are the sample means. We can shown that, for both the parameter and its estimator, ρ is a number lying between −1 and 1. In the N-dimensional space $\mathbb{R}^N$ containing the data sample, (1.2) is the inner product between the two sample vectors, divided by their Euclidean norms. Schwarz's inequality $(\mathbf{x}, \mathbf{y}) \leq ||\mathbf{x}||\ ||\mathbf{y}||$, shows that $|\rho_{XY}| \leq 1$. The inequality also shows that ρ is ±1 only if the two variables are perfectly correlated and the scatterplot would be a perfect straight line; conversely zero corresponds to no statistical connection between the two.

In our example I calculate ρ to be −0.2603. Is this significant, or would a random collection of data pairs commonly come up with a number as big as this in magnitude? To answer this question we will assume that X and Y are jointly distributed with a Gaussian law. The theoretical sampling distribution for $\hat{\rho}$ in general is very complicated; see Priestley, Chap 9 for references. But in one case it turns out to be relatively easy to compute – just the case we have: we test the null hypothesis that the true ρ is zero. Then (Kendall and Stuart, *Advanced Theory of Statistics* Vol 2, p 316) it can be shown that the variable t defined by

$$t^2 = \frac{(N-2)\,\hat{\rho}_{XY}^2}{1 - \hat{\rho}_{XY}^2} \tag{1.3}$$

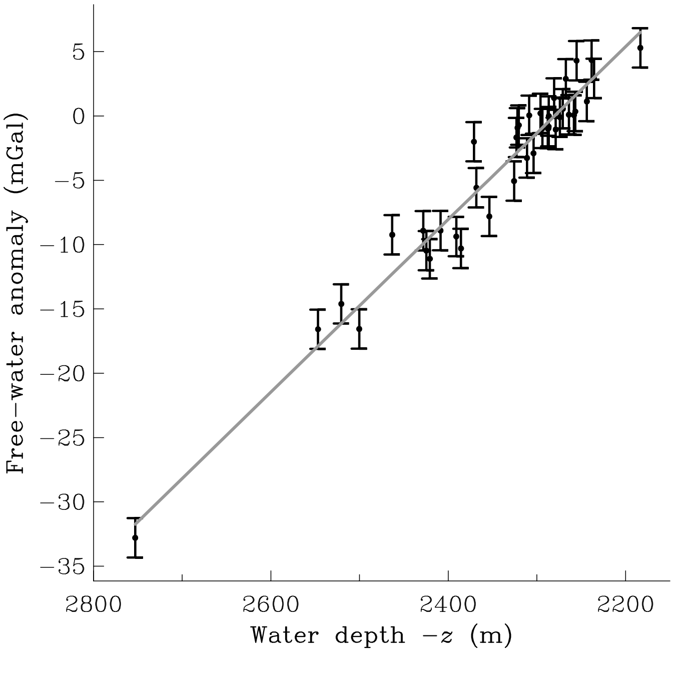
follows the **Student t-distribution** with $N-2$ degrees of freedom. Thus we can find the probability that even though the true correlation coefficient of the distribution is zero, the estimator, by chance gets bigger than t. Plugging in the numbers for Fig 1 we get $t=1.264$, and from the t tables, I find the probability of its being this big by chance is about 0.15; this implies an 85% confidence there is a relationship.

We continue to look at the case of a pair of variables, without an underlying evolutionary process in time. Often in the physical sciences we have a reason for believing in a linear relationship between two variables based on some theory, but that there is a random component affecting one (or both) of the observations. Fig 2 is a geophysical application from the PhD thesis of Mark Stevenson, who worked for Mark Zumberge. This plot (from Fig 3 of Stevenson et al., *JGR* 99, 4875-88, 1994) concerns gravity survey made on the Vance-Cleft Overlapping Rift Zone on the Juan de Fuca ridge. The two variables are: (y-axis) the observed gravity anomaly measured on the seafloor, corrected for the earth's vertical gravity gradient and the presence of seawater; (x-axis) the depth of the observation site. For smooth terrain we expect the two to be connected by:

$$\Delta g = 2\pi G\, \rho_c\, z + c \tag{1.4}$$

where G is Newton's gravitational constant and ρ_c is the density of the crust. (What is c in this equation?) Finding the crustal density from the slope of the line, is known as **Nettleton's Method**. The gravity variable is much less accurately known than the water depth because of instrument shaking due to water currents and the factor that the seafloor density really isn't constant; so the fit is not exact (of course). Here the

Figure 2: Sea-floor gravity versus bathymetry.



appropriate statistical model is not one of a pair of random variables with joint two-dimensional Gaussian distribution, but rather:

$$Y_j = \alpha x_j + \beta + N_j \tag{1.5}$$

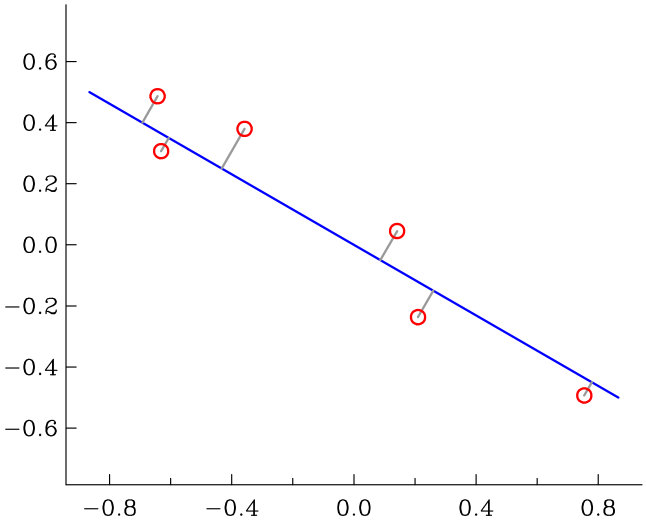
where N_j are random variables, ideally distributed identically with zero mean, uncorrelated and maybe with known variance. (Why did I write Y in upper case but x lower case?) Though you can calculate the correlation coefficient here, the graph makes it clear there is a strong relationship and the real interest centers on estimation of the slope of the line, which leads to the constant ρ_c in (1.4); also we would like a figure for the uncertainty of the estimate. You should recognize a good application of the **least squares** method. I won't go over that material here, since you must have seen it several times already. However, I want just to mention one result. The least-squares estimator of the slope is given by:

$$\hat{\alpha} = \frac{\sum_j (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_j (X_j - \bar{X})^2} = \hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \ . \tag{1.6}$$

Remember, for uncorrelated random variables, (1.6) has the least variance of all unbiased linear estimators (Gauss-Markov Theorem). If the variables are Gaussian in addition, (1.6) is the best unbiased estimator of α, period.

Notice that in (1.5) that **all** the noise is attributed to the y variable, in the example, the gravity measurement. But what if both coordinates have random components? Then things get considerably more complicated, depending on what can be assumed about the noise. You will easily convince yourself that the estimate of the slope is different if one assumes all the noise is x instead of y. The next simplest case is the one in which the noise in x and y is taken to be the same. But observe this really only makes any sense when x and y are measured in the same units; it is meaningless to say the water depth in meters has the same uncertainty

Figure 2a: Distances used in TLS.



as the gravity anomaly in m s^{-2} ! In this special case we perform what is called a **total least squares** estimate: we minimize the sum of squares of the Euclidean distances from the points in the plane to the line.

If the straight line is $y = \alpha x + \beta$, the TLS estimator for the intercept is

$$\hat{\beta} = \frac{1}{N} \sum_j Y_j - \frac{\hat{\alpha}}{N} \sum_j X_j = \hat{\hat{X}} - \hat{\alpha} \hat{\hat{Y}} \tag{1.7}$$

where $\hat{\alpha}$ is the slope estimate given by the solution of this quadratic equation:

$$0 = -c_2 + c_1 \hat{\alpha} + c_2 \hat{\alpha}^2 \tag{1.8}$$

$$\begin{aligned} c_2 &= \sum_j (X_j - \hat{\hat{X}})(Y_j - \hat{\hat{Y}}) \\ c_1 &= \sum_j [(X_j - \hat{\hat{X}})^2 - (Y_j - \hat{\hat{Y}})^2] \end{aligned} \; . \tag{1.9}$$

A quadratic equation usually has two solutions of course; in thus case, one corresponding to the best fit, the other to the worst.

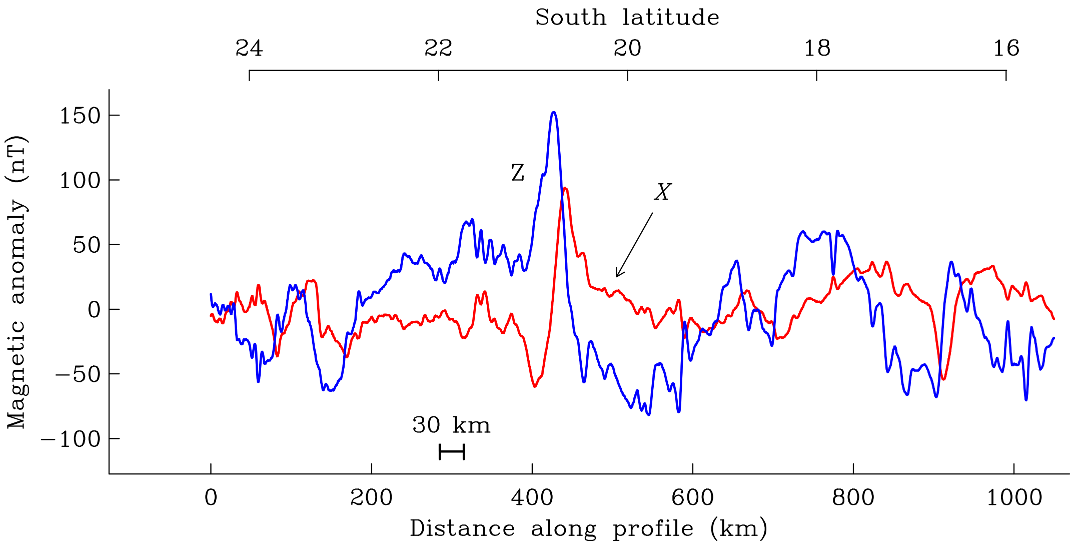
Exercise

In (1.5) suppose it is known that $\beta = 0$, so that the true straight lines runs through the origin. Solve the problem of estimating the slope α by the least-squares method. Give an explicit formula like (1.6) for the estimator.

2. Pairs of Stationary Signals

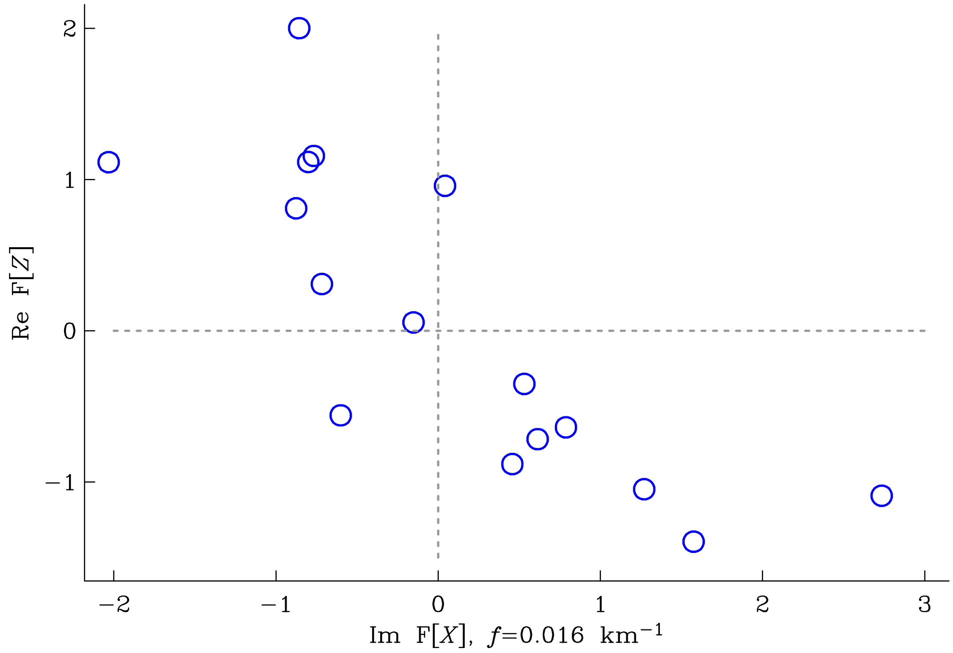
Suppose now we have a pair of signals, possibly related, both of them assumed to be stationary processes, for example the pair in Fig 3. These

Figure 3: Two components of an aeromagnetic signal.



are the track-parallel X, and vertical Z, magnetic anomaly components as measured on a high-flying aircraft traveling roughly northward in the southeastern Pacific over Nazca Plate. There are two ways of thinking about the power spectrum. The more intuitive one asks us to imagine filtering the time series through a very narrow band filter to isolate a given frequency component – then the variance of the random signal that results, normalized by the bandwidth of the filter is the power spectral density. The corresponding idea for handling pairs of signals is very similar: imagine taking the Fourier transform of an identical time sample of each series and selecting a single frequency; do this again and again with new realizations. Now form a *complex* scatterplot of the Fourier amplitude of series 1 against 2. Each point on the plot is one Fourier transform. If the series are related, the cloud of points will tend to be elongated. We can compute the correlation coefficient using the complex version of (1.2) and taking the absolute value; this is called the **coherency** at the selected frequency. You can also form the (complex) slope, and this is the **transfer function** between the two signals. To give you the flavor, I have plotted the imaginary part of $\mathcal{F}\,[X]$ versus the real part of $\mathcal{F}\,[Z]$ at one particular wavenumber: see Fig 4. I made separate realizations by dividing the original series into shorter pieces. You see the strong (negative) correlation between the two components; we will see why we should expect this behavior in a moment. To do a proper job of showing you the complete complex scatterplot I would have to plot every component against every other (six plots as we omit self-correlations, and the system is symmetric). But the frequency domain provides a basis for forming numerical estimates.

Figure 4: Scatter plot of two components of DFTs from Fig 3 at a fixed wavenumber.



Before discussing the estimation process, we develop in more detail the alternative of looking at spectra as Fourier transforms of covariance functions. Let two real stationary stochastic processes be X_1 and X_2 both with zero mean, and define the **cross-covariance** between them:

$$R_{21}(s) = \mathrm{cov}\,[X_1(t),\ X_2(t+s)] = \mathcal{E}\,[X_1(t)\,X_2(t+s)] \tag{2.1}$$

and in general the **covariance matrix** is

$$R_{kj}(s) = \mathcal{E}\,[X_j(t)\,X_k(t+s)], \quad j,\,k = 1, 2. \tag{2.2}$$

Notice the (somewhat illogical looking) reversal of the order of the indices in the function and under the expectation. Because the processes are stationary, R_{jk} depends only on s even though t appears in the equation. Also some trivial algebra gives $R_{12}(s) = R_{21}(-s)$.

We have already seen that the **power spectral density** (PSD) is defined by taking the Fourier transform of the covariance function, in the present case choosing the diagonal elements of the matrix. Similarly, we define the **cross-spectrum** between X_1 and X_2 by

$$S_{12}(f) = \mathcal{F}\,[R_{12}](f) = \int_{-\infty}^{\infty} R_{12}(t)\,\mathrm{e}^{-2\pi \mathrm{i} ft}\,dt\ . \tag{2.3}$$

Unlike the PSD, which is always real and nonnegative, the cross-spectrum is usually a complex function of frequency. Traditionally the real and imaginary parts are given names:

$$S_{12} = C_{12} + \mathrm{i}Q_{12} \tag{2.4}$$

where C_{12} is called the **cospectrum** and Q the **quadrature** spectrum. Then the **coherency** spectrum is defined by

$$\gamma_{12} = \frac{|\,S_{12}\,|}{\sqrt{S_{11}\,S_{22}}} = \frac{\sqrt{C_{12}^2 + Q_{12}^2}}{\sqrt{S_{11}\,S_{22}}}\ . \tag{2.5}$$

I might add that I personally prefer to use the square of γ_{12}, which is called the **coherence**. Equation (2.5) gives the correlation coefficient between the two signals as a function of frequency. Its equivalence to the more intuitive definition I gave earlier certainly requires proof; we don't have time here for it. Unfortunately, I don't find Priestley's (1981) demonstration (p 661) particularly convincing, as it depends on the orthogonal increment representation of the processes.

In addition to the coherence, we define the **phase spectrum** in the straightforward way as:

$$\Phi_{12} = \tan^{-1}\left(\frac{-Q_{12}}{C_{12}}\right). \tag{2.6}$$

The definition (2.3) is not a good way to estimate the cross spectrum from data time series, but it is usually the best way to make theoretical calculations as we will illustrate in short while.

Exercise

Suppose X_1 is a stationary stochastic process of time, and X_2 is defined by $X_2(t) = X_1(t+t_0)$, where t_0 is a constant. Find the coherence and phase spectrum of these two signals. How do these spectra depend on the PSD of X_1?

3. Estimation of Cross Spectra (Briefly)

Recall the standard modern way of estimating the PSD. It is based on the idea that the PSD is the limit of a finite Fourier transform. Let us assume that $\Delta t = 1$, so that the Nyquist frequency $f_{Nyq} = ½$. The periodogram estimate for the PSD is

$$\hat{P}_X(m\Delta f) = \left| \sum_{n=0}^{N-1} X_n \mathrm{e}^{-2\pi mn/N} \right|^2, \quad m = 0, 1, , \cdots ½N \tag{3.1}$$

where $\Delta f = 1/N$, and we have chosen the output frequency samples to take advantage of the FFT. As you should remember, this estimator has enormous variance, for white Gaussian noise the standard deviation as the same as the answer. And for non-white processes, the result can be strongly biased by spectral leakage. We ameliorate both these defects by using 3tapers. To improve the variance we must average over independent estimates. As you may recall we asserted that if the two tapers u_n and v_n are orthogonal (so that $\sum_n u_n v_n = 0$) then PSD estimates based on tapered versions of the time series $u_n X_n$ and $v_n X_n$ are independent estimates of P_X. So the modern strategy is to choose a set of mutually orthogonal tapers with good spectral leakage properties, and then to average the estimates together. That is **multitaper estimation**; see Percival and Walden, (1993), and Riedel and Sidorenko (1995).

Multitaper estimation can and should be used with cross spectra. The equivalent periodogram estimator for the cross spectrum between X_n and Y_n is just

$$\hat{P}_{XY}(m\Delta f) = \left(\sum_{n=0}^{N-1} X_n \mathrm{e}^{-2\pi mn/N} \right) \left(\sum_{n=0}^{N-1} Y_n \mathrm{e}^{-2\pi mn/N} \right)^* . \tag{3.2}$$

Tapered series can be formed and transformed with FFTs, but to preserve the independence the same taper set must be used for X_n and Y_n even though the two processes may be quite different, and the optimal set to suppress leakage would then be different too. When K sets are used the variance is reduced by the factor K^{-1}. Estimates of coherence and phase are usually based on the definitions

$$\hat{\gamma}_{XY} = \frac{| \hat{P}_{XY} |}{\sqrt{\hat{P}_X \hat{P}_Y}} \tag{3.3}$$

$$\hat{\Phi}_{XY} = \tan^{-1} \left(\frac{-\operatorname{Im}(\hat{P}_{XY})}{\operatorname{Re}(\hat{P}_{XY})} \right). \tag{3.4}$$

These are **not** unbiased estimators, even when the various spectra and cross spectra are. If the same tapers are used to estimates P_X and P_Y then it can be shown that $0 \le \hat{\gamma}_{XY} \le 1$, which is desirable; if different sets are used, the coherence can wander above unity.

Uncertainty estimates can be formed based on Gaussian statistics, but the modern way is the Jack-knife estimator, which makes no such assumptions. One exception is the case when we want to know if the coherence (the analog of the correlation coefficient) differs significantly from zero. Priestley (p 706) shows that under the Gaussian assumption for X_n and Y_n, if K independent estimates are averaged to find $\hat{\gamma}_{XY}$ then the statistic s defined by

$$s = \frac{2K\hat{\gamma}_{XY}}{1 - \hat{\gamma}_{XY}} \tag{3.5}$$

has an $F_{2,4K}$ distribution. This allows a simple test for the hypothesis that γ_{XY} is zero:

$$\Pr(| \hat{\gamma}_{XY} | \geq z) = (1 - z^2)^{K-1} \tag{3.6}$$

which is often the most important thing one wishes to know.

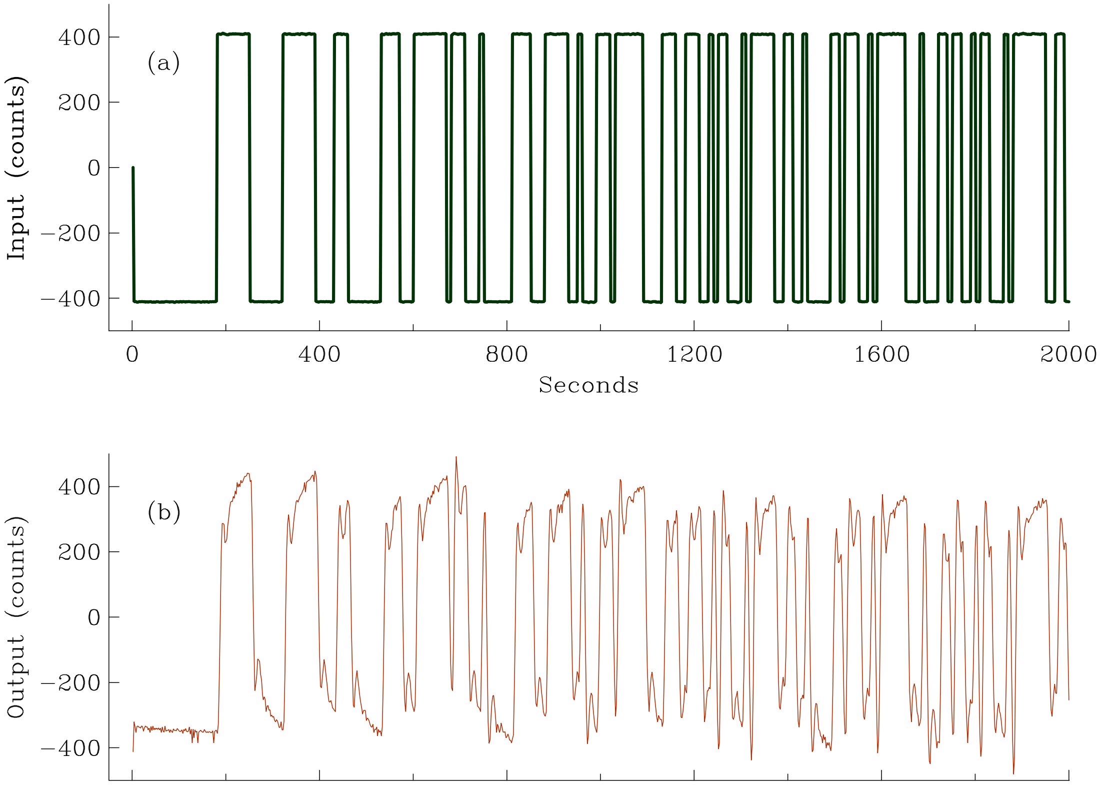
4. Example: Calibration — Convolution plus Noise

We return to time series – functions of a single independent variable, like time. Consider the calibration of a seismometer, by shaking it with a "random" signal generated by a computer

$$V = g * I + N \ . \tag{4.1}$$

Here V is the output voltage, g is the **impulse response** of the seismometer (it is the output for a delta function input), I is the input signal, and N is noise – it's the result of ordinary ground motion, from which one cannot isolate the system completely. (See the paper by Berger, et al., *Bull. Seism. Soc. Am., 69*, 271-88, 1979). Fig 5 shows the input (a), and output from a long-period seismometer, (b). The input is called a **random telegraph** signal – it switches randomly between one of two values. If the transition times were entirely uncorrelated I, would be a white noise but, because the transitions are allowed only on multiples of 5 data samples, this is not exactly true here; this input provides a fairly flat spectrum, without the large amplitudes; see Fig 6.

Figure 5: Input and output signals for a seismometer undergoing calibration.



If we possessed an infinite, noise-free record we could take the FT with time and obtain:

$$\hat{V} = \hat{g}\,\hat{I} \tag{4.2}$$

where $\hat{g}$ is called the **transfer function**, or simply the **frequency response** of the seismometer. At each frequency the output is just the input multiplied by a complex number, which varies with frequency. Our task here is to estimate $\hat{g}$ when V and I are stochastic processes which we have measured; the noise N is unknown. At first you might think all you have to do is to take the FFT of the time series, and divide both sides of (4.2) by $\hat{I}$ but, just as a naive technique is poor way of estimating the power spectrum, this is not the best estimator. Notice that in (4.1) we have the same least-squares situation as the gravity problem at the beginning: *the noise is confined to only one of the signals*. For stationary stochastic processes the answer is the analog of (1.6) as you might expect:

$$\hat{g} = \frac{S_{VI}}{S_I} \tag{4.3}$$

the cross spectrum of V with I divided by the PSD of I. The magnitude $|\,\hat{g}\,|$ is often called the **gain** of the system; then we can define the transfer function in terms of its gain and, through (2.6), its phase.

Here is the proof of (4.3). To make any progress we must assume that N in (4.1) is *uncorrelated* with the input I. Then we calculate the cross-covariance :

$$R_{VI}(s) = R_{IV}(-s) = \mathcal{E}\,[V(t)\,I(t-s)] = \mathcal{E}\,[(g * I + N)(t)\,I(t-s)] \tag{4.4}$$

$$= \mathcal{E}\,[(g * I)(t)\,I(t-s)] \tag{4.5}$$

$$= \mathcal{E}\,[\int dp\;\, g(p)\,I(t-p)\,I(t-s)] \tag{4.6}$$

$$= \int dp\;\, g(p)\,R_I(p-s) = \int dp\;\, g(p)\,R_I(s-p) \tag{4.7}$$

where in (4.7) we used the definition of R_I to evaluate $\mathcal{E}\,[I(t-p)\,I(t-s)]$, having moved the expectation under the integral. Now take the Fourier transform of (4.7), which is by definition the cross spectrum between V and I:

$$S_{VI}(f) = \int ds \int dp\; \mathrm{e}^{-2\pi \mathrm{i} sf}\, g(p)\,R_I(s-p) \tag{4.8}$$

$$= \int dp\, [\int ds\; \mathrm{e}^{-2\pi \mathrm{i}(s-p)f}\, R_I(s-p)]\; \mathrm{e}^{-2\pi \mathrm{i} pf}\, g(p) \tag{4.9}$$

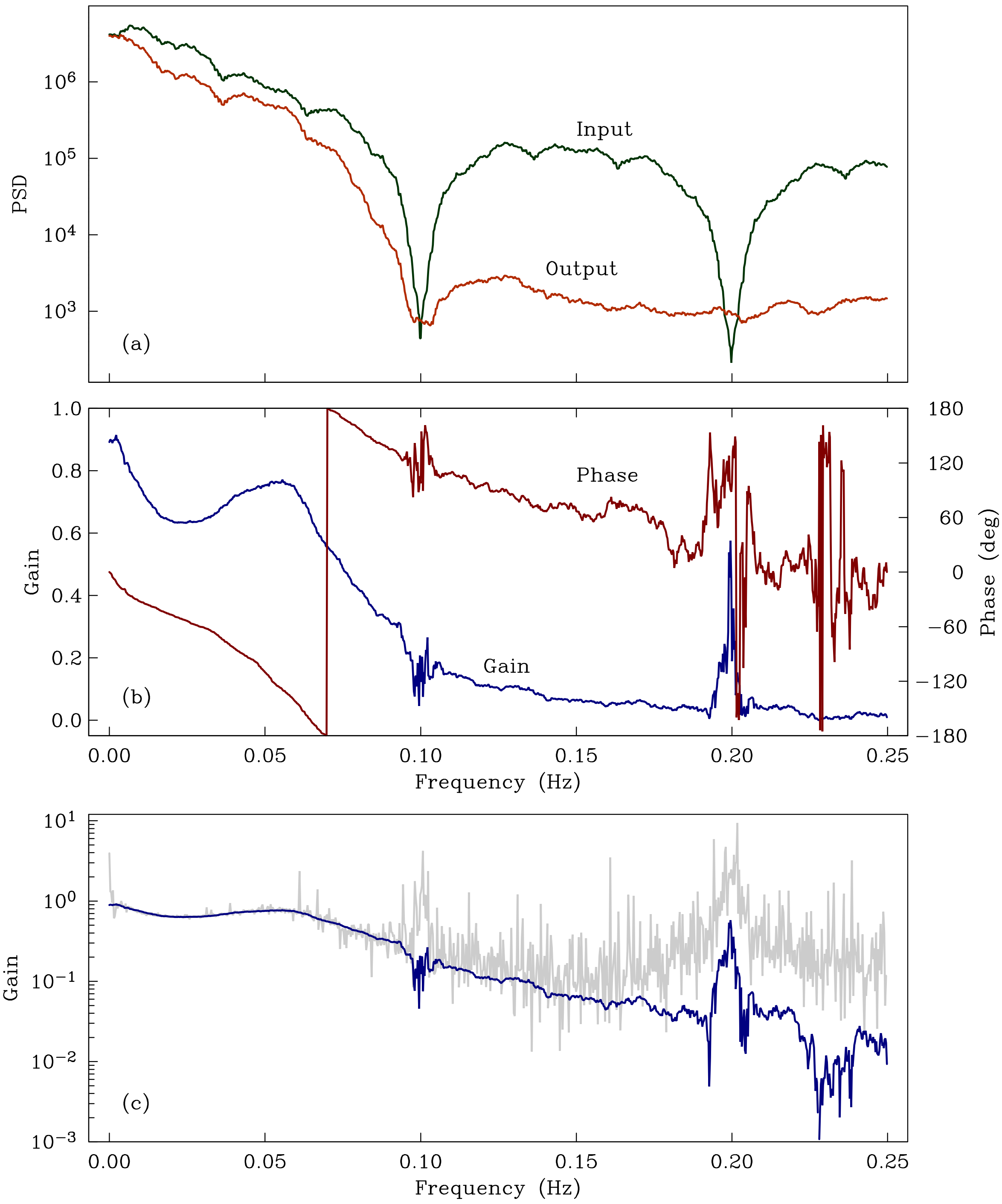
$$= S_I(f)\,\hat{g}(f)\;. \tag{4.10}$$

And this is just (4.3).

This is not of course a proper derivation of an *estimator*. What it shows is that if all the noise is in one of the stationary processes, the transfer function is given exactly by the ratio of the cross spectrum to the PSD of the input. It can be shown (see Priestley) that $\hat{g}$ is an unbiased estimator when estimates of S_{VI} and S_I are used rather than the exact functions.

Fig 6 shows in (a) the power spectra of input and output of our seismometer example. The random telegraph input has two very low power

Figure 6: Gain and phase characteristics of the seismometer.



"holes" in its PSD at 0.1 Hz, and again at 0.2 Hz. It is exactly at these frequencies the estimated phase and gain becomes unreliable, as can be seen in (b). The gain, here the magnitude of the seismometer frequency response, is in reality a very smooth function. The estimates are quite smooth, except in the neighborhood of the input power holes, and for $f>0.17$Hz where ground motion is having an effect. In the bottom panel, which has a log scale for gain, I have plotted in grey the naive estimate $|\hat{V}/\hat{I}|$, a simple ratio of DFTs. Note how very rough that estimate is, and how strongly biased — the average values are much too large.

The geophysical application most commonly associated with cross spectral analysis is geomagnetic and magnetotelluric sounding, where time series of electric and magnetic field components are the basis for estimates for the impedance of the Earth as a function frequency, from which electrical conductivity structure can be inferred. One complication here is that both signals are subject to noise, so that the approach described for calibration, while often used, is not strictly valid, and alternatives to avoid these problems sometimes involve a special experimental setup called a remote reference station (Gamble, et al., (1979); Egbert and Booker (1986).

Another important topic in both Earth and planetary sciences is cross-spectral analysis between topography and gravity data. The transfer function, called the **isostatic response**, is a function wavenumber, not frequency, yields insight into crustal and lithospheric strength. In this case it is plausible to assert that the gravity data are far more prone to error than the topography, and so a transfer function estimate like the one we have just done will be appropriate. See Chapter 5 of Watts (2001) for many details. However, for this work we may need to work from data over an area, rather than profiles, and that brings us to our next subject.

5. Stationary Processes in the Plane

We take a brief look at **multidimensional spectra**, that is spectra of functions of several variables. Up to now we considered on a single independent variable (usually identified with time). Suppose there is a stationary processes in the plane (like bathymetry, or the gravity field from random sources). It is simple to define the autocovariance of X, a signal in the plane, by following the single-variable recipe:

$$R_X(\mathbf{r}) = \mathcal{E}\,[X(\mathbf{s})\,X(\mathbf{s}+\mathbf{r})] \tag{5.1}$$

for $\mathbf{r}, \mathbf{s} \in \mathbb{R}^2$. Then the two-dimensional power spectral density of X is just the 2-D Fourier transform of R_X as you would expect:

$$S_X(\mathbf{k}) = \mathcal{F}_{\,2}[R_X](\mathbf{k}) = \int_{\mathbb{R}^2} R_X(\mathbf{x})\,\mathrm{e}^{-2\pi i\mathbf{k}\cdot\mathbf{x}}\,d^2\mathbf{x}\,. \tag{5.2}$$

Quite often we have measurements on a single profile, or a slice thorough a 2-dimensional random field. How do we collapse the 2-dimensional spectrum to get the observed profile spectrum? This is exactly what we have with the aeromagnetic data of Fig 3. We will prove the **Slice Theorem** for spectra and cross spectra. Then we'll apply it to magnetic data.

First we state without proof the Slice Theorem for ordinary Fourier transforms. Given the 2-dimensional Fourier transform of a function f:

$$\hat{f}(\mathbf{k}) = \mathcal{F}_{\,2}[f](\mathbf{k}) \tag{5.3}$$

we find the 1-dimensional FT of f along the line $y = 0$ by:

$$\mathcal{F}_{\,x}[f(x,0)](k) = \int_{-\infty}^{\infty} dk_y\ \hat{f}(k_{x,}\,k_y)\,. \tag{5.4}$$

In words, we integrate the 2-dimensional Fourier transform in the perpendicular direction in the wavenumber domain to get the corresponding 1-dimensional FT.

Exactly the same result holds for power spectra and cross spectra! We prove it for the PSD. A one-dimensional stationary process U is a sample on the line $x\hat{\mathbf{x}}$ of the process W in the plane $\mathbb{R}^2$. The autocovariance of W is R_W and then the autocovariance of U is

$$R_U(x) = R_W(x\hat{\mathbf{x}}) = R_W(x,\,0)\,. \tag{5.5}$$

By definition, the PSD of W is the 2-dimensional Fourier transform of the autocovariance R_W, that is

$$S_W = \mathcal{F}_{\,2}[R_W]\,. \tag{5.6}$$

Conversely:

$$R_W(\mathbf{x}) = \mathcal{F}_{\,2}^{-1}[S_W] = \int_{\mathbb{R}^2} d^2\mathbf{k}\ \mathrm{e}^{2\pi i\mathbf{x}\cdot\mathbf{k}}\ S_W(\mathbf{k})\,. \tag{5.7}$$

Substituting (5.7) into (5.5) gives

$$R_U(x) = \int\limits_{\mathbb{R}^2} d^2\mathbf{k}\ \mathrm{e}^{2\pi \mathrm{i} x \hat{\mathbf{x}}\cdot\mathbf{k}}\ S_W(\mathbf{k}) \tag{5.8}$$

$$= \int\limits_{-\infty}^{\infty} dk_x\ \mathrm{e}^{2\pi \mathrm{i} k_x x} \int\limits_{-\infty}^{\infty} dk_y\ S_W(k_x, k_y) \tag{5.9}$$

$$= \mathcal{F}_x^{-1}[\int\limits_{-\infty}^{\infty} dk_y\ S_W(k_x, k_y)]\ . \tag{5.10}$$

And again the PSD of U is the one-dimensional Fourier transform of R_U:

$$S_U(k) = \mathcal{F}_x[\mathcal{F}_x^{-1}[\int\limits_{-\infty}^{\infty} dk_y\ S_W(k_x, k_y)]] \tag{5.11}$$

$$= \int\limits_{-\infty}^{\infty} dk_y\ S_W(k_x, k_y)\ . \tag{5.12}$$

Equation (5.12) is the Slice Theorem for power spectra. The result is the same for cross spectra: simply integrate the 2-dimensional cross spectrum in the perpendicular direction in the wavenumber domain.

Suppose there are two stochastic processes U and V both derived from a third X via convolution:

$$U = g * X, \quad V = h * X\ . \tag{5.13}$$

How is the cross spectrum S_{UV} related to S_X, the PSD of X? The answer, which takes a few lines of simple algebra is:

$$S_{UV} = \hat{g} \cdot \hat{h}^* S_X \tag{5.14}$$

where the hat means Fourier transform. (When a hat decorates a function it denotes the Fourier transform; on a simple variable, it means a statistical estimate; I hope this convention does not cause confusion.) This result is valid for processes of a single variable or in the plane $\mathbb{R}^2$, or indeed higher dimensions. When $g = h$, so that $U = V$, we get the well-know result for the PSD of U:

$$S_U = |\hat{g}|^2 S_X\ . \tag{5.15}$$

The generalization of these results to three dimensional space is quite straightforward and we will not spend time on it. Another, more interesting generalization is to the sphere. Clearly for geophysics it is only an approximation to say that the surface of the Earth is a plane; we can ask what happens if the region is so large that we must account for curvature, or even larger still, so that the whole surface of the Earth is the domain of the random process. We will discuss that question later.

A specially important kind of stochastic process in the plane is the **isotropic process**, which is one with the property that the autocovariance depends only on $|\mathbf{r}|$, the distance between the two points. Then let

$$R_X(\mathbf{r}) = \rho(|\mathbf{r}|) . \tag{5.16}$$

The PSD is as always the Fourier transform of R_X and so

$$S_X(\mathbf{k}) = \mathcal{F}_2[R_X] = \sigma(|\mathbf{k}|) \tag{5.17}$$

where you will recall that the 2-D FT of a circularly symmetric function like ρ is another function with circular symmetry. (Here symbol σ is nothing to do with stanadard error.) The relation between ρ and σ is a **Hankel transform**:

$$\sigma(k) = \mathcal{H}[\rho] = \int_0^\infty \rho(r)\, J_0(2\pi kr)\, 2\pi r \; dr \tag{5.18}$$

and $J_0(x)$ is the Bessel function. We have encountered the Hankel transform in Fourier theory. We note that the Hankel transform is its own inverse: $\mathcal{H}[f] = \mathcal{H}^{-1}[f]$. When a process in a plane is isotropic it is only necessary to gather data on a straight profile to construct the PSD, because the autocovariance along a straight line in any direction is just the function $\rho(x)$. This allows us to bring in the Slice Theorem again.

Suppose one has observations along a single straight line from which we can calculate the profile spectrum $P_X(k_x)$; notice that because the stochastic process is isotropic, any line across the field can be designated the x axis. Then we can find the profile PSD in two ways: either as the one-dimensional FT of the autocovariance:

$$P_x(k_x) = \mathcal{F}_1[\rho] \tag{5.19}$$

or by the Slice Theorem:

$$P_X(k_x) = \int_{-\infty}^{\infty} S_X(k_x, k_y)\; dk_y = \int_{-\infty}^{\infty} \sigma(\sqrt{k_x^2 + k_y^2})\; dk_y \tag{5.20}$$

In practice, one may have observations on a profile and hence knowledge of P_X, which one would like to convert into the 2-D PSD function σ. The way is clear: we invert (5.19), then apply (5.18) to find σ. In fact the action of a 1-D FT followed by a Hankel transform can be combined into a new operation:

$$\sigma(k) = \mathcal{H}[\mathcal{F}_1^{-1}[P_X(k_x)]] = -\frac{1}{\pi} \int_k^\infty \frac{dP_X}{dk_x}\, \frac{dk_x}{\sqrt{k_x^2 - k^2}} . \tag{5.21}$$

I omit details of this derivation, which can be found in Bracewell, who calls (5.21) the **Abel transform**. Equation (5.21) takes us directly from a profile PSD to the 2-D PSD function. Equation (5.20) goes the opposite way.

6. Example: Magnetics over the Ocean

We can put all the 2-D material together for the aeromagnetic signals. The magnetic field components X and Z are related to V, the scalar magnetic potential in the plane at $z = 0$. Since the potential V is harmonic (it obeys Laplace's equation, $\nabla^2 V = 0$) above the sources, it can be upward continued from the plane $z = 0$. You should recall that for wavenumbers $\mathbf{k} \in \mathbb{R}^2$

$$\hat{V}(\mathbf{k}, z) = \hat{V}(\mathbf{k}, 0)\, \mathrm{e}^{-2\pi |\mathbf{k}| z} \tag{6.1}$$

where $\hat{V}$ is the FT of V in a plane of constant z:

$$\hat{V}(\mathbf{k}, z) = \mathcal{F}_2[V(\mathbf{x}, z)] = \int_{IR^2} \mathrm{e}^{-2\pi i \mathbf{k}\cdot\mathbf{x}} V(\mathbf{x}, z)\, d^2\mathbf{x}\,. \tag{6.2}$$

Thus knowledge of V on $z = 0$ is enough to determine V anywhere above this level, provided the sources all lie below the plane $z = 0$. If we take the inverse 2-D FT of (6.1), and use the Convolution Theorem (backwards), we find

$$V(\mathbf{x}, z) = V_0 * G = G * V_0 \tag{6.3}$$

where $V_0 = V(\mathbf{x}, 0)$ and

$$G(\mathbf{x}) = \mathcal{F}_2^{-1}[\mathrm{e}^{-2\pi |\mathbf{k}| z}] = \frac{1}{2\pi} \frac{z}{(z^2 + |\mathbf{x}|^2)^{3/2}}\,. \tag{6.4}$$

We need, not the potential, but the vector $\mathbf{B}$, and $\mathbf{B} = -\nabla V$. From the definition of convolution it follows that

$$B_z = Z = -\frac{\partial V(\mathbf{x}, z)}{\partial z} = -\frac{\partial G}{\partial z} * V_0 = G_Z * V_0 \tag{6.5}$$

and similarly for the other components.

So far we have been considering ordinary functions for V and Z. Now suppose that V_0 is a stationary stochastic process in the plane $z = 0$, a random field. Then from (6.5) so is the vertical field. Suppose the (two-dimensional) PSD of V_0 is $S_V(\mathbf{k})$. We calculate the (2-D) PSD of Z in the standard way, equation (27):

$$S_Z = |\hat{G}_Z|^2 S_V \tag{6.6}$$

where $\hat{G}_Z$ is the 2-D FT of G_Z. But by (6.4) G is the inverse FT of the exponential in (6.1), so it follows that

$$\hat{G}(\mathbf{k}) = \mathrm{e}^{-2\pi |\mathbf{k}| z} \tag{6.7}$$

$$\hat{G}_Z = -\frac{\partial \hat{G}}{\partial z} = 2\pi |\mathbf{k}|\, \mathrm{e}^{-2\pi |\mathbf{k}| z}\,. \tag{6.8}$$

In a very similar way we can calculate the spectrum of $X = B_x = -\partial V/\partial x$:

$$B_x = X = -\frac{\partial V(\mathbf{x}, z)}{\partial x} = -\frac{\partial G}{\partial x} * V_0 = G_X * V_0 \tag{6.9}$$

$$\hat{G}_X = 2\pi \mathrm{i} k_x \mathrm{e}^{-2\pi |\mathbf{k}| z} \tag{6.10}$$

$$S_X = |\hat{G}_X|^2 S_V \,. \tag{6.11}$$

And Y follows in exactly the same way, with

$$\hat{G}_Y = 2\pi \mathrm{i} k_y \mathrm{e}^{-2\pi |\mathbf{k}| z} \,. \tag{6.12}$$

Notice that the cross spectrum, say between X and Z, as given by (26) is

$$S_{XZ} = \hat{G}_X (\hat{G}_Z)^* S_V \,. \tag{6.13}$$

We have here a 2-D system in which the three processes X, Y, Z (which happen to be three components of a random vector) are each given by different convolutions of a single process V. This means the 3 components of the random magnetic vector are closely related. For example, because $|\mathbf{k}|^2 = k_x^2 + k_y^2$, it follows that

$$S_Z = S_X + S_Y \,. \tag{6.14}$$

This result says that the vertical component of **B** of a field generated from random sources is always larger on average than either of the other two.

And now back to that long magnetic profile over Pacific Ocean shown in part in Fig 3. We have all the pieces to be able to say something interesting. To look at the power and cross spectrum of the profile data we simply apply the Slice Theorem:

$$P_X(k_x) = \int_{-\infty}^{\infty} dk_y \; |\hat{G}_X(\mathbf{k})|^2 \; S_V(\mathbf{k}) \tag{6.15}$$

$$P_Y(k_x) = \int_{-\infty}^{\infty} dk_y \; |\hat{G}_Y(\mathbf{k})|^2 \; S_V(\mathbf{k}) \tag{6.16}$$

$$P_Z(k_x) = \int_{-\infty}^{\infty} dk_y \; |\hat{G}_Z(\mathbf{k})|^2 \; S_V(\mathbf{k}) \tag{6.17}$$

$$P_{XZ}(k_x) = \int_{-\infty}^{\infty} dk_y \; \hat{G}_X(\mathbf{k}) \, \hat{G}_Z(\mathbf{k})^* \; S_V(\mathbf{k}) \,. \tag{6.18}$$

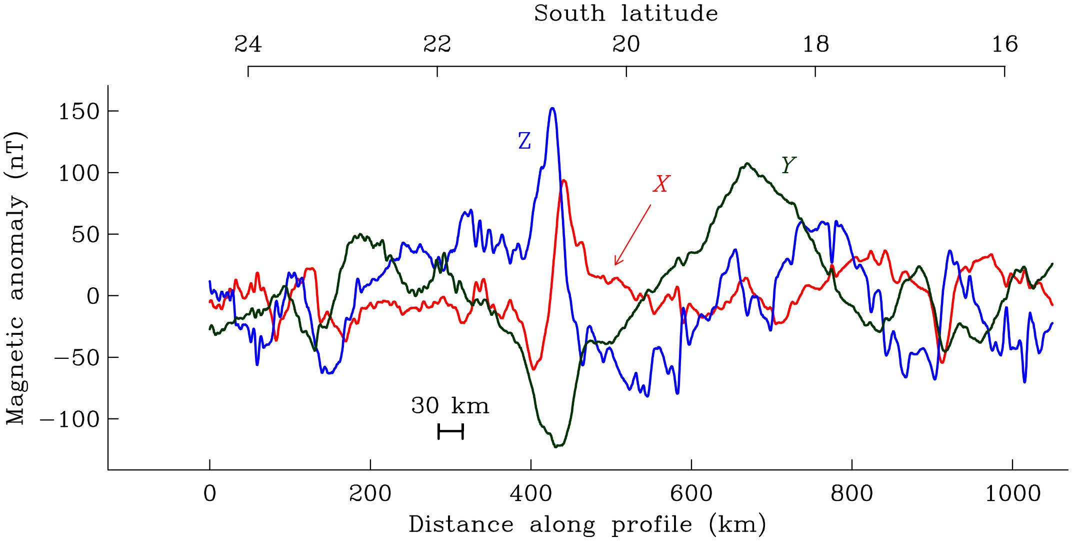
Observe that we have the same relationship between the profile PSDs of the components as in the 2-D system:

$$P_Z = P_X + P_Y \,. \tag{6.19}$$

This relationship has been called the **Power Sum Rule**. Another thing to notice in this example is that while $\hat{G}_Z$ is real, $\hat{G}_X$ (and $\hat{G}_Y$) is purely imaginary. Thus whatever the spectrum is for S_V, from (6.12) we see the phase spectrum between must be exactly $\pi/2$ and constant for all k_x. This is a very strong prediction. We'll return to this in a moment.

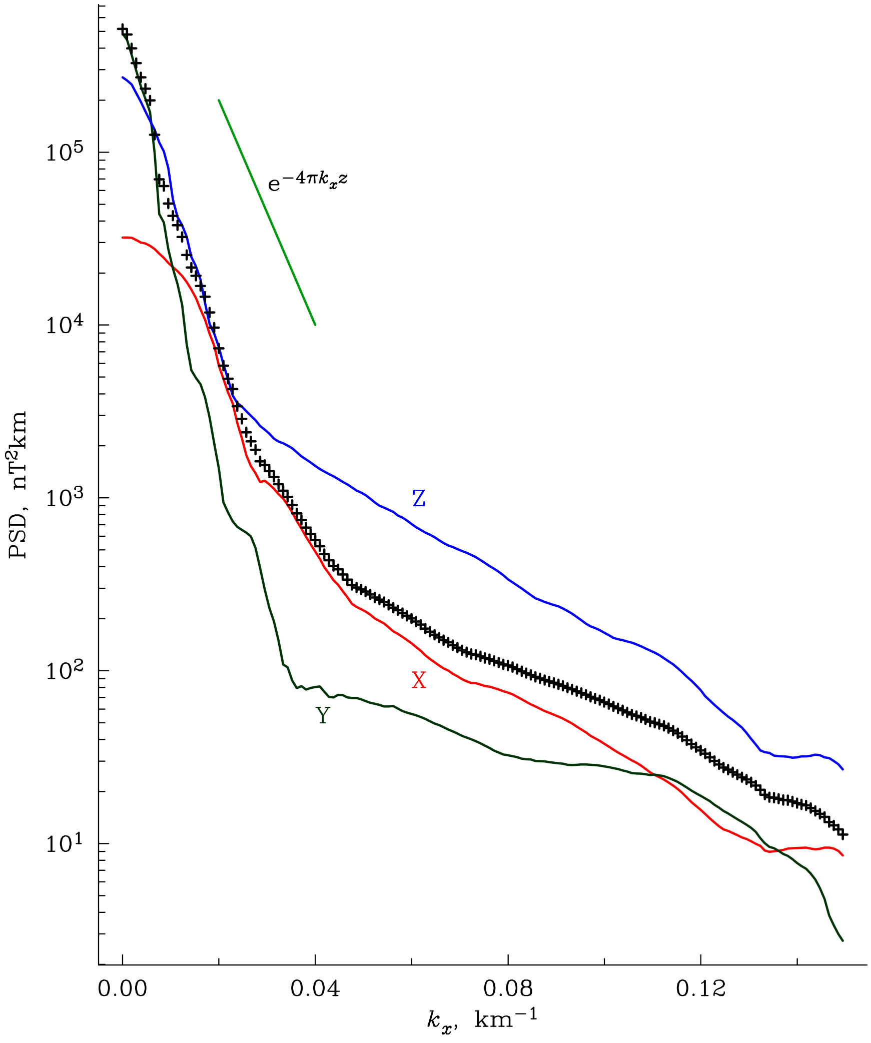
Looking at the one-dimensional spectra and cross spectra of measurements made through a random field of higher dimension is a very common practice in physical oceanography: the idea that you get a version that simply integrates over the unexplored directions is taken as obvious in that field, but this is not yet so in most of geophysics. Finally in this section, let's look at the estimated spectra for the data of Fig 7, which gives all three components (only X and Z were plotted before). This is part of great circle path flown from Easter Island to Bay St Louis.

Figure 7: Magnetic anomaly on a great circle path across the Pacific ocean.



This discussion is based closely on the paper of Parker and O'Brien *JGR*, 102, B11, pp 24815-24, 1997. The first interesting plot for us is Fig 8, the PSDs of all three components. Only the lowest tenth of the wavenumber spectrum are shown, since measurements of the field were made every 350 meters (What then is the Nyquist wavenumber?). All spectra fall off steeply, in fact exponentially to a first approximation, changing slope dramatically at about $k_x = 0.3$ km^{-1}. Approximate exponential decay like $\exp(-4\pi k_x z)$ is expected from (6.15)-(6.17) if S_V is fairly flat. And the slope of the PSDs is about right initially: aircraft height plus water depth equals $7+4$ km, so $z = 11$ km. It is plausible that the field spectrum falls of faster since S_V itself ought be a red spectrum, like most other geophysical PSDs. But clearly something happens when $k_x > 0.03$. Next look at the plus signs; these plot the value of $P_X + P_Y$. According to the Power Sum rule (6.19) the pluses should lie on top of the blue P_Z line. Again things go as we might hope until 0.03 km^{-1}.

Figure 8: PSDs of the three components of the magnetic anomaly shown in Fig 7.

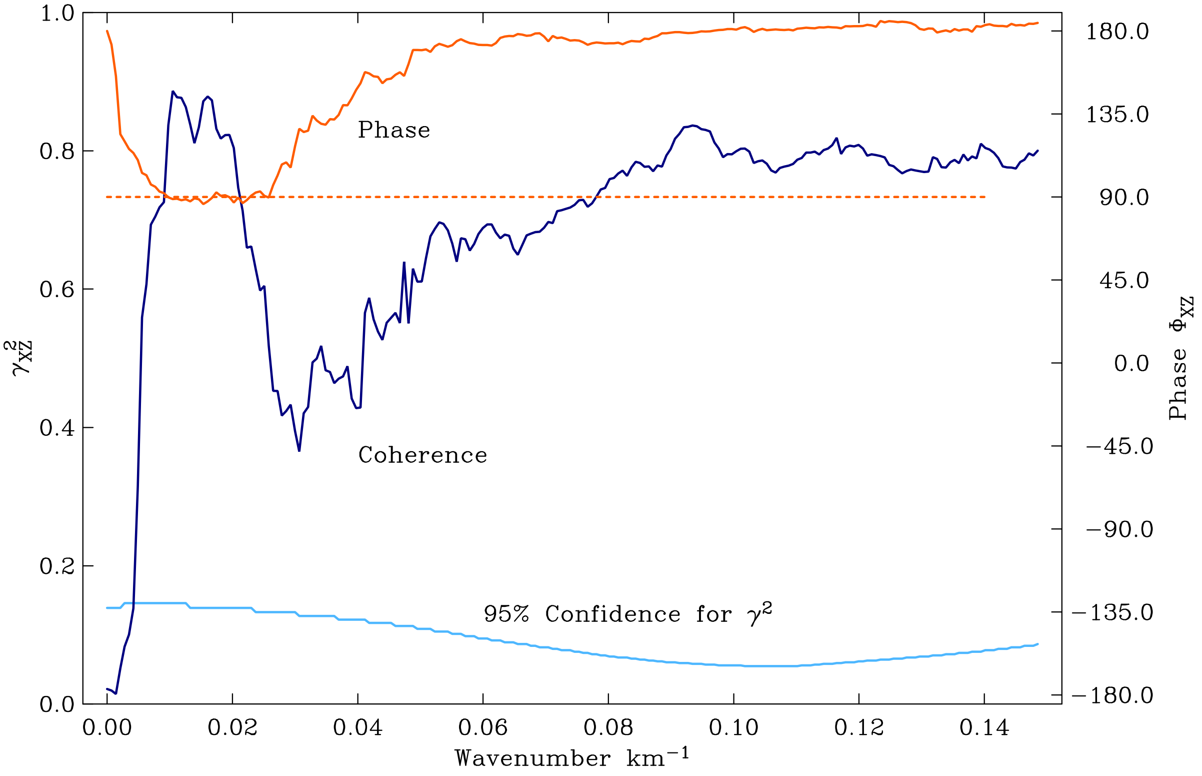


For further clues to solving this mystery we look at the cross spectra, here just at X and Z. In Fig 9 we plot the phase and coherence spectra. The phase is about 90° for the longer wavelengths although it wanders off right near $k_x = 0$, which may be an estimation artifact, or a real issue. The theory we did early predicts the 90° phase for potential fields with sources below the plane. Like the PSDs this shows that only for $k < 0.03\ \mathrm{km}^{-1}$ (length wavelength > 33 km) are magnetic signals of crustal origin.

As O'Brien and I explain in the paper quoted, the cause of the noise is error in the orientation of the magnetometers. Fig 7 shows that the anomalies are about 100 nT a fraction of a percent of the main geomagnetic field, which is over 30,000 nT and it is roughly horizontal pointing north. When the gyro-stabilized platform shakes, a tiny fraction of this horizontal field appears on the Z sensor and the other components are similarly corrupted, though less so; this is indicated clearly in Fig 8. Can you see that because the main field points upward in the southern hemisphere, the rocking platform causes a signal that is 180° out of phase between Z and X, just as we see in Fig 9; high coherence is also predicted by this mechanism.

One lesson from the power and cross spectra is that 90% of the bandwidth of the record from this Project Magnet data is devoted to noise! Only signal with wavelength less than 30 km are geophysical, and wiggles on a smaller scale must not be interpreted as geology. For interpretational purposes we should filter the traces with a low-pass filter. Indeed

Figure 9: Coherence and phase spectra between X and Z.



the very longest wavelength signals are also noise: on a flight of such long duration, the time varying magnetic field appear as long wavelength spatial signals, and these have not been properly corrected. Fig 8 shows the Y component has more power than the Z near $k_x = 0$, which the Power Sum rule says can never happen; the improper phase spectrum there also suggests a problem.

7. Stationary Processes on a Sphere

The following is one way of approaching the problem. It must be noted that, as usual in the literature involving spherical harmonics, everyone feels free to define different normalizations, so factors of π, l, $l+1$ and $2l+1$ appear in various places in the papers of various authors.

The proposition is that the statistics of a stationary random process on the unit sphere, called $S^2(1)$, cannot depend on the position $\hat{\mathbf{r}}$. We will assume that $V(\hat{\mathbf{r}})$, the process has zero mean, so that $\mathcal{E}\,[V]=0$ is obviously position independent. The second-order statistics are once again captured by the autocovariance:

$$\tilde{R}_V(\Delta) = \mathrm{cov}\,[V(\hat{\mathbf{r}}),\, V(\hat{\mathbf{s}})] \tag{7.1}$$

$$= \mathcal{E}\,[V(\hat{\mathbf{r}})\; V(\hat{\mathbf{s}})] \tag{7.2}$$

where Δ is the angle at the center of the sphere between the two unit vectors. On the sphere being independent of location means that every unit vector is the same as every other, that any reorientation of the sphere must leave $\tilde{R}_V$ unaltered. It is convenient to use the cosine of Δ rather than Δ itself in the definition:

$$R_V(\cos\Delta) = R_V(\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}) = \mathcal{E}\,[V(\hat{\mathbf{r}})\; V(\hat{\mathbf{s}})] \tag{7.3}$$

If this is the autocovariance, what is the corresponding PSD? How do we decompose a function on a sphere into different wavelength components? Answer: spherical harmonics! The analog of the relation that the autocovariance is the FT of the PSD is the expansion of R_V :

$$R_V(\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}) = \sum_{l=0}^{\infty} S_l\, P_l(\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}) \tag{7.4}$$

Here S_l is the variance of V in the part of the function with spherical harmonic degree l. Since $P_l(1)=1$ we see that

$$R_V(1) = \mathrm{var}\,[V] = \sum_{l=0}^{\infty} S_l \tag{7.5}$$

This result is the equivalent of the fact that the variance of random process on a line or in the plane is the area under the PSD. The sequence S_l is my candidate for the **Spherical Power Spectrum**.

The inverse of (7.4), since Legendre polynomials are orthogonal is

$$S_l = (l+\tfrac{1}{2}) \int_0^{\pi} R_V(\cos\theta)\, P_l(\cos\theta)\, \sin\theta\, d\theta \tag{7.6}$$

$$= \sqrt{(2l+1)\pi} \int_{S^2(1)} R_V(\hat{\mathbf{z}}\cdot\hat{\mathbf{s}})\; Y_l^0(\hat{\mathbf{s}})^*\; d^2\hat{\mathbf{s}} \tag{7.7}$$

where the spherical harmonic is normalized to $\int |\,Y_l^m\,|^2 = ||\,Y_l^m\,||^2 = 1$. Equations (7.6) and (7.7) are the analog to the fact that the PSD is the FT of the autocovariance.

An alternative definition proceeds as follows: imagine that the stationary process is written out as a spherical harmonic expansion on $S^2(1)$:

$$V(\hat{\mathbf{r}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} C_{lm} Y_l^m(\hat{\mathbf{r}}) \tag{7.8}$$

where C_{lm} must be random complex coefficients. Plug this into (7.3). It turns out the only way that you can get R_V to be independent of the vectors $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$ in (7.3) is to say that

$$\operatorname{cov}[C_{lm}, C_{nk}^*] = 0, \text{ unless } l = n, \text{ and } m = k \tag{7.9}$$

In other words the coefficients in the expansion must be uncorrelated random variables. Furthermore, we find that

$$\mathcal{E}\,[\,|\,C_{lm}\,|^2] = \sigma_l^2 \tag{7.10}$$

independent of the order m. Then from (7.3) and (7.8) we find

$$R_V(\hat{\mathbf{r}} \cdot \hat{\mathbf{s}}) = \mathcal{E}\,[V(\hat{\mathbf{r}})V(\hat{\mathbf{s}})] = \mathcal{E}\,[V(\hat{\mathbf{r}})V(\hat{\mathbf{s}})^*] \tag{7.11}$$

$$= \mathcal{E}\left[\sum_l \sum_m C_{lm} Y_l^m(\hat{\mathbf{r}}) \sum_n \sum_k C_{nk}^* Y_n^k(\hat{\mathbf{s}})^*\right] \tag{7.12}$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \mathcal{E}\,[\,|\,C_{lm}\,|^2] Y_l^m(\hat{\mathbf{r}}) Y_l^m(\hat{\mathbf{s}})^* \tag{7.13}$$

We used (7.9) to go from (7.12) to (7.13). Now we use the Spherical Harmonic Addition Formula together with (7.10)

$$R_V(\hat{\mathbf{r}} \cdot \hat{\mathbf{s}}) = \sum_{l=0}^{\infty} \frac{(2l+1)\sigma_l^2}{4\pi} P_l(\hat{\mathbf{r}} \cdot \hat{\mathbf{s}}) \tag{7.14}$$

Compare this to (7.4) and we see that

$$\sigma_l^2 = \frac{4\pi S_l}{2l+1} \tag{7.15}$$

The seemingly unnecessary factor of 4π here could be eliminated if we were willing to use spherical harmonics normalized as $\|\,Y_l^m\,\|^2 = 4\pi$.

The surprising fact is that on a sphere there are only isotropic stationary processes analogous to those described by equations (5.16-21), and none corresponding to the more general (5.1).

Spherical power spectra are used to describe the geomagnetic and gravitational potential fields, and also in seismology and geodynamics to characterize velocity and temperature structure at a particular radius within the Earth. One of the most successful uses of the stochastic model on the sphere has been geomagnetic field over geological time: if the dipole and a few of the other harmonics are excluded, the remaining field appears to be spatially stationary and to have a “white” spectrum (Constable and Parker, 1988).

References

Berger, J., D. C. Agnew, Parker, R. L., and W. E. Farrell. Seismic system calibration, 2. Cross-spectral calibration using random binary signals. Bull. Seism. Soc. Am., 69, 271-288, 1979.

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

Constable, C. G. and Parker, R. L., Statistics of the geomagnetic secular variation for the past 5 m. y., J. Geophys. Res., 93, B10. 11569-81, 1988.

Egbert, G. D., and Booker, J. R., Robust estimation of geomagnetic transfer functions. GJRAS, 87, 173-94, 1986.

Gamble, T. D., Clarke, J., and Goubau, W. M., Magnetotellurics with a remote magnetic reference. *Geophysics,* 44, 53-68, 1979.

Kendall, M. G., and Stuart, A., *Advanced Theory of Statistics* Vol 2, Griffin, London, 1966.

Parker,R. L., and O'Brien, M. S., Spectral analysis of vector magnetic field profiles, *JGR*, 102, B11, pp 24815-24, 1997.

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

Priestley, M. B., *Spectral Analysis and Time Series,* Academic Press, New York, 1981.

Rice, J. A., *Mathematical Statistics and Data Analysis,* Brooks-Cole Pub. Co., Monterey, CA, 1988.

Riedel, K. S., and Sidorenko, A., Minimum bias multiple taper spectral estimation, IEEE Trans. Sig. Proc., 43, 188-195, 1995.

Stevenson, J. M., Hildebrand, J. A., Zumberge, M. A. and Fox, C. G., An ocean bottom gravity study of the southern Juan de Fuca Ridge: J. Geophys. Res., 99, 4875-4888. 1994.

Watts, A. B., *Isostasy and Flexure of the Lithosphere*, Cambridge Univ. Press, Cambridge, 2001.