GEOPHYSICAL DATA ANALYSIS

Class Notes by Bob 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:

$$\sum_{n=-\infty}^{\infty} R_X(n)^2 = \int_{-½}^{½} S_X(f)^2 \, df < \infty \tag{1.1}$$

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

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

(We run into a notational problem, that the hat $\hat{}$ 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:

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

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.

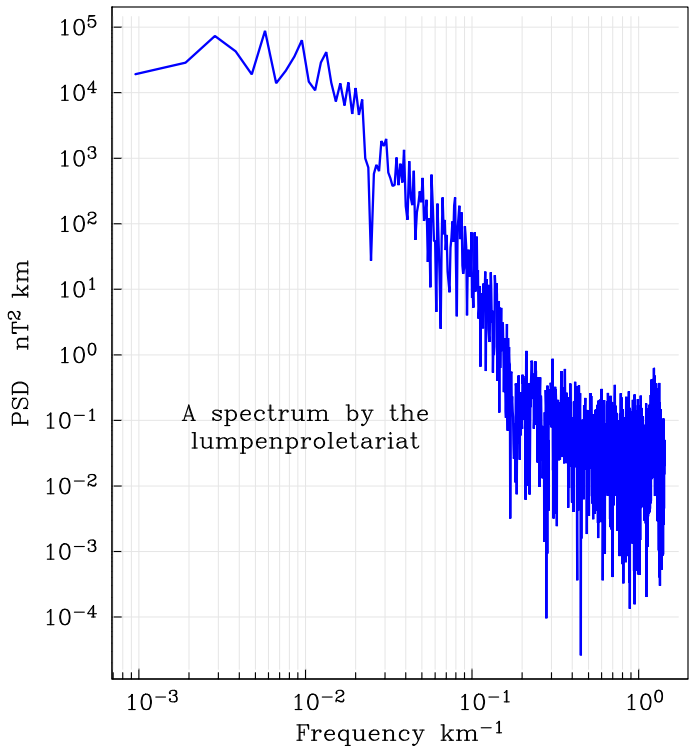


Figure 0: A primitive periodogram spectrum of the Project Magnetic X field shown in earlier Chapters

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

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

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 $f = \pm ½$. 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):

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

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:

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

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

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

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

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

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

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:

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

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 $\frac{1}{2} N \sigma^2$, 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 $\pm\frac{1}{2}$ will be intended by this phrase), the DFT comprises iid Gaussian RVs $\sim N(0, \frac{1}{2} N \sigma^2)$; 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 $\frac{1}{2} N \sigma^2$, 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^2_K, 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 $\mathrm{var}\,[A_m] = \frac{1}{2}N\sigma^2$. 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 $\rho^2 = \mathrm{var}\,[A_m] = \frac{1}{2}N\sigma^2$. 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 $\frac{1}{2}N$ 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:

$$\mathcal{E}\,[\hat{S}_X(f)] = \sum_{n=-N+1}^{N-1} \left(1 - \frac{|\,n\,|}{N}\right) \int_{-½}^{½} S_X(f')\, \mathrm{e}^{2\pi i f' n}\, df'\, \mathrm{e}^{-2\pi \mathrm{i} f n} \tag{4.6}$$

$$= \int_{-½}^{½} S_X(f') \left\{ \sum_{n=-N+1}^{N-1} \left(1 - \frac{|\,n\,|}{N}\right) \mathrm{e}^{-2\pi \mathrm{i}(f-f')n} \right\} df' \tag{4.7}$$

$$= \int_{-½}^{½} S_X(f')\, F_N(f'-f)\, df' \tag{4.8}$$

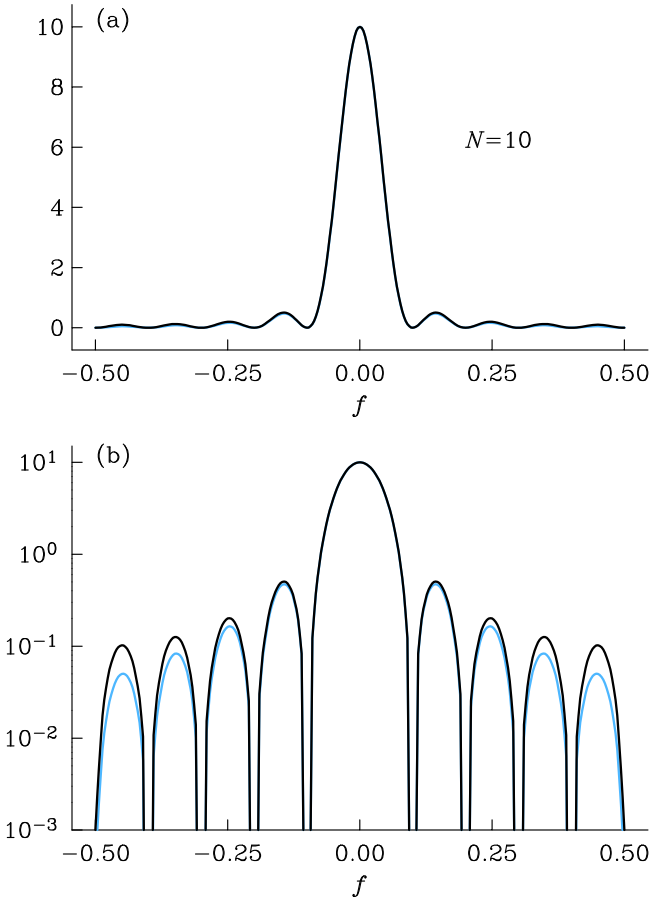
where the sum in (4.8) can be evaluated exactly, and is called the **Fejer kernel**:

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

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

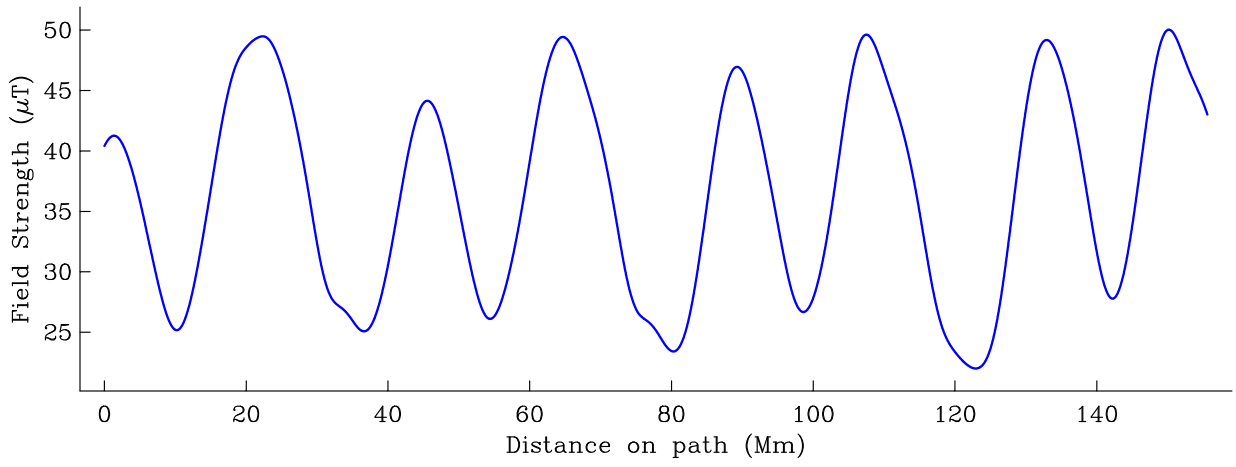
To improve the poor variance of the periodogram we must expect to average in some way. The most natural approach, which we will examine briefly, is simply to average together a number of estimates made at neighboring frequencies. The periodogram has the virtue that the spectral estimates at the frequencies $m\Delta t$ 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)

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

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

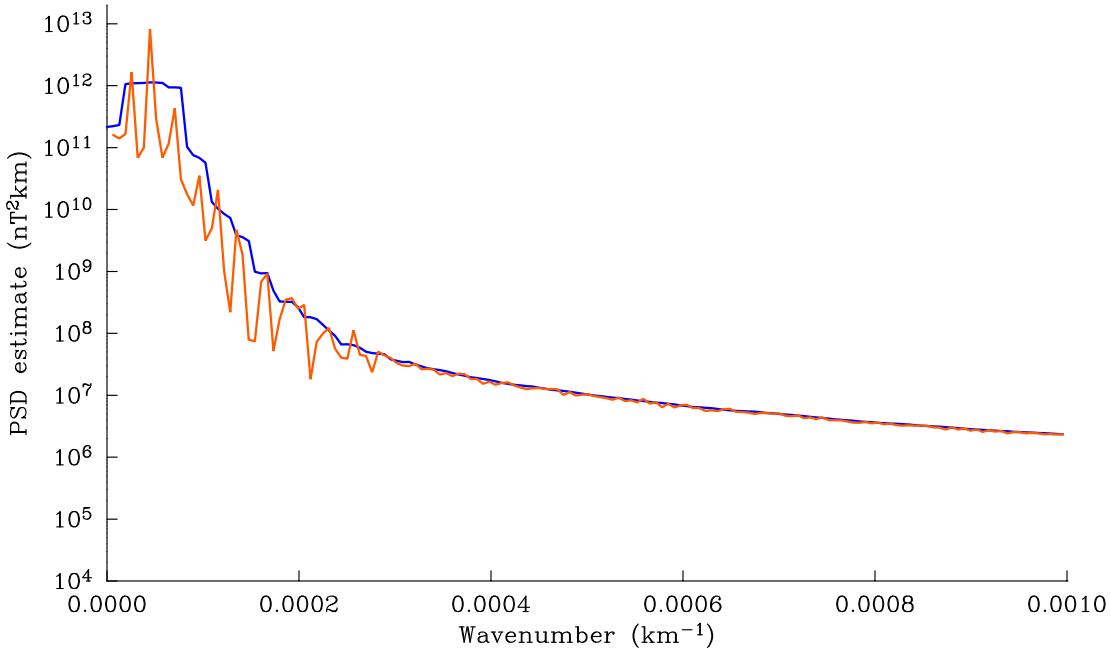
$$\hat{S}_X(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} w(n)\, x_n e^{-2\pi i n f}\right|^2 . \tag{5.2}$$

Then the new expected value of the estimator comes out as

$$\mathcal{E}\,[\hat{S}_X(f)] = \int_{-½}^{½} S_X(f')\, W(f'-f)\, df' \tag{5.3}$$

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:

$$w(t) = \text{box}(t/N - \tfrac{1}{2}) \, . \tag{5.5}$$

Then

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

and because $|e^{\pi i f N}| = 1$ (5.4) agrees exactly with the approximation $N \, \text{sinc}^2(Nf)$ 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

$$w_A(t) = \begin{cases} (2/T)^{1/2} \sin(\pi t/T), & 0 \le t \le T \\ 0, & \text{otherwise} \, . \end{cases} \, . \tag{5.7}$$

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

$$\hat{w}_A(f) = (\tfrac{1}{2}T)^{1/2} \, [\text{sinc}\,(fT + \tfrac{1}{2}) + \text{sinc}\,(fT - \tfrac{1}{2})] \, . \tag{5.8}$$

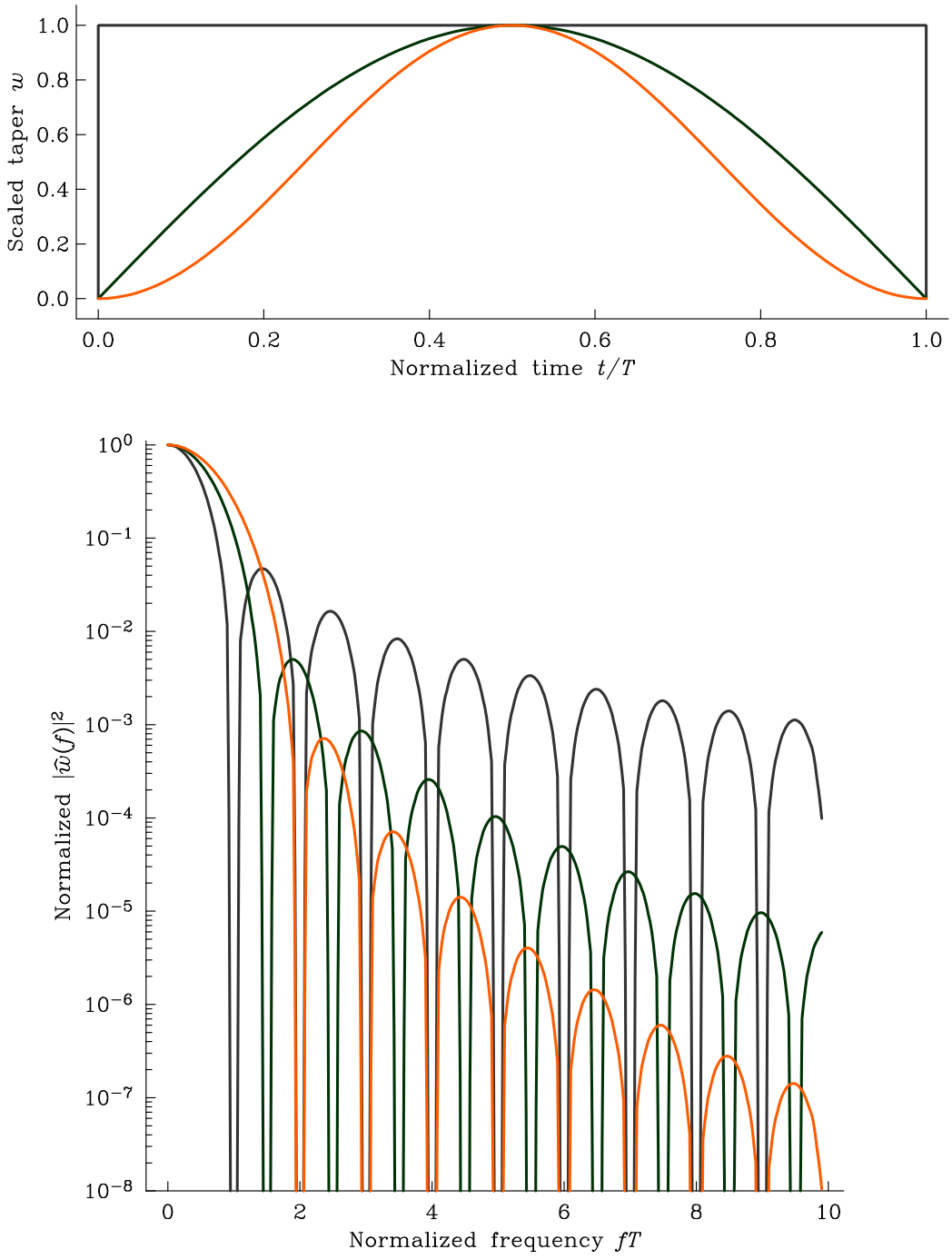
Next the sine-squared taper:

$$w_B(t) = \begin{cases} (8/3T)^{1/2} \sin^2(\pi t/T), & 0 \le t \le T \\ 0, & \text{otherwise} \end{cases} \tag{5.9}$$

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

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 $\mathrm{sinc}^2(fT)$. In the plots I have scaled the function to unity at $f = 0$ for easy of comparison.

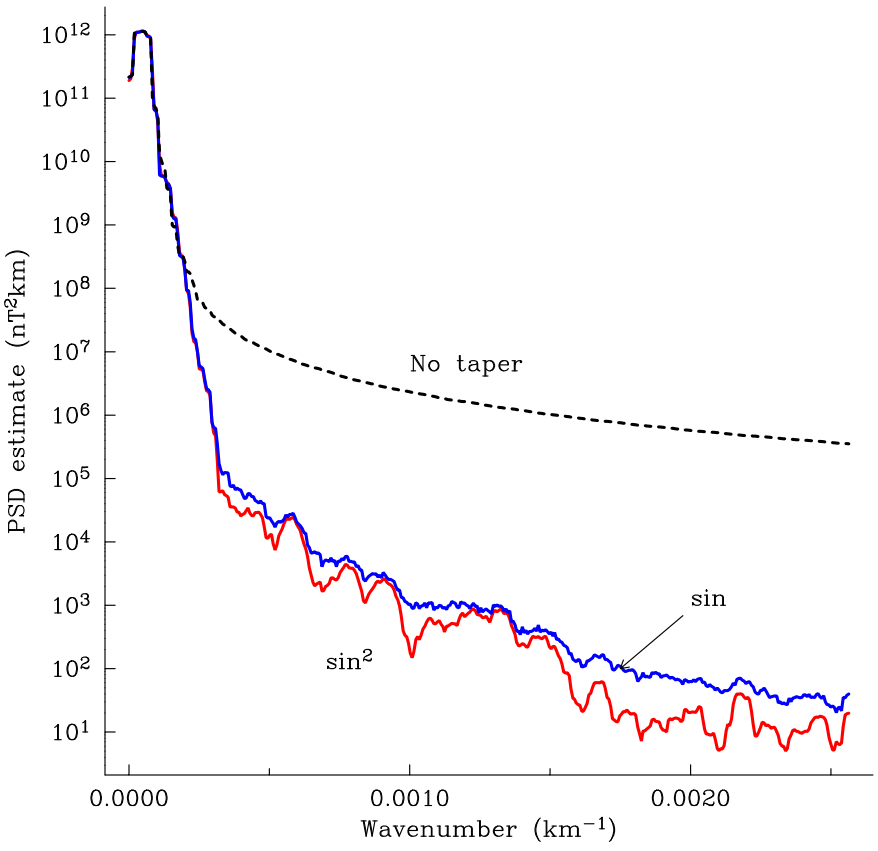
Figure 4: Top panel. Three tapers, box-car (or no taper) gray; sine taper, dashed; sine-squared taper black. Lower panel. Corresponding convolving weight functions with frequency.



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.

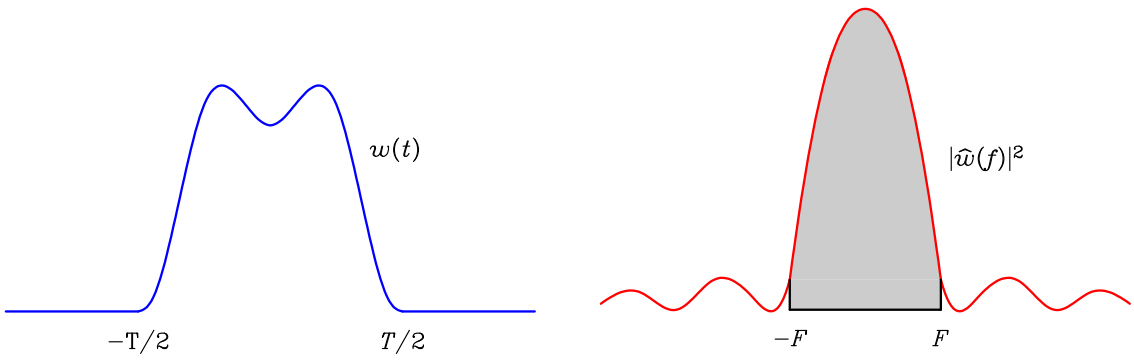
6. The Perfect Taper

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**:

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

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,

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

into equation (6.2):

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

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

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

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 $(-1, 1)$:

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

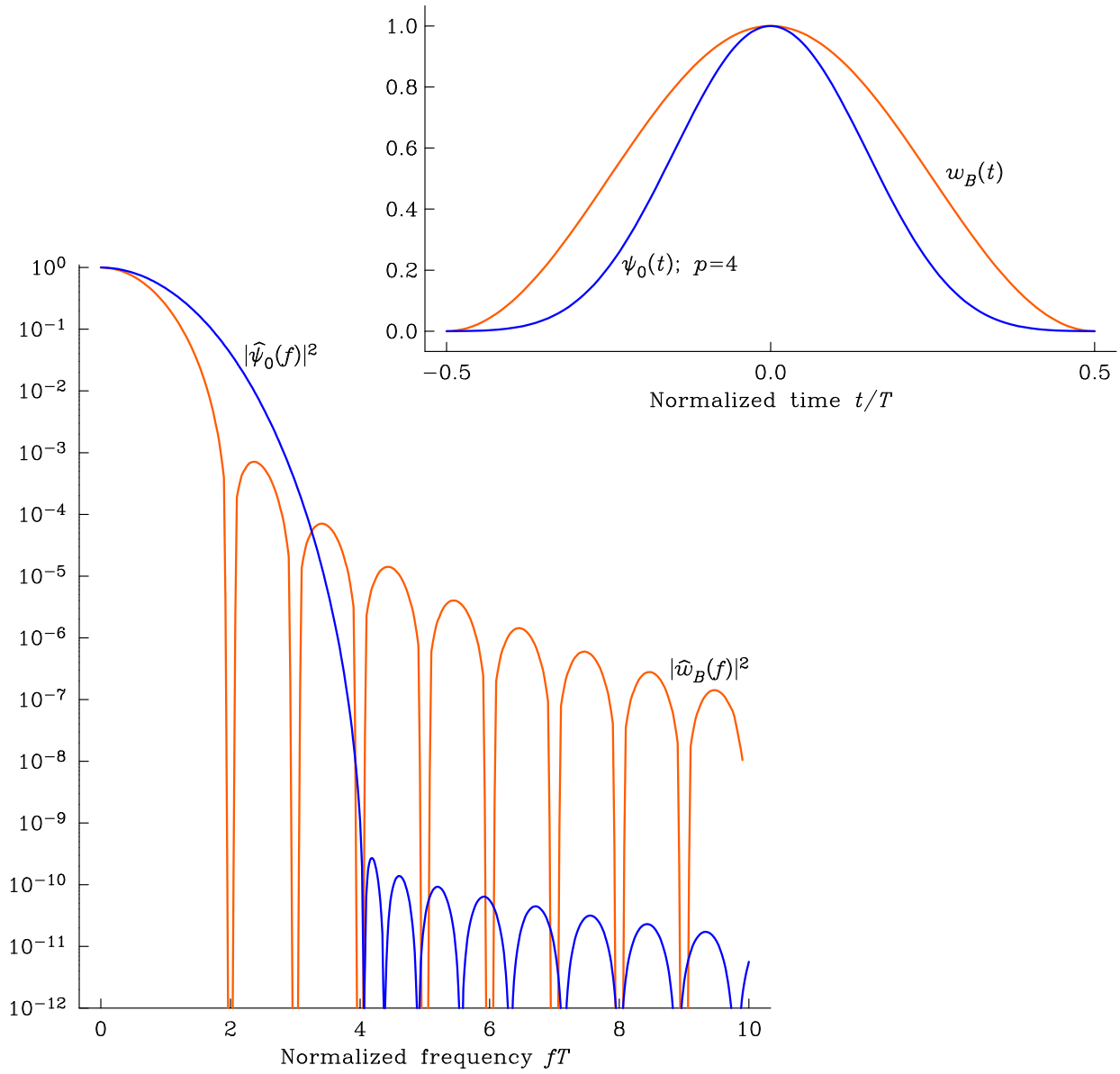
which is a general property of self-adjoint eigensystems like (6.12). More unusual, and harder to prove, is the following fact about the eigenvalues: for $n \leq 2p$, the eigenvalues μ_n (and hence the spectral concentrations) *are all very close to unity*, then fall suddenly to very small values. This 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

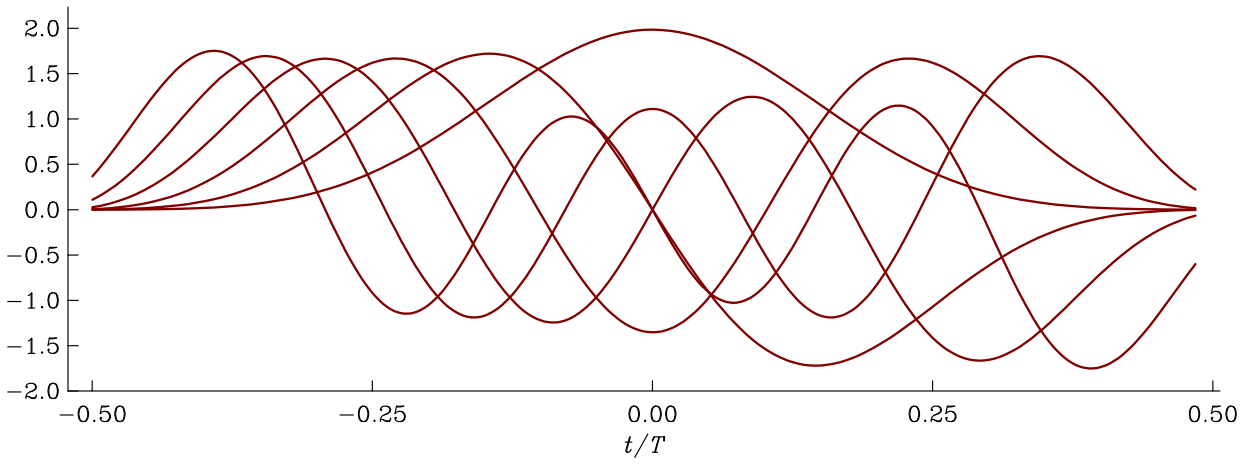
At this point we have focussed on the taper with the best concentration, $\psi_0(t)$, with eigenvalue μ_0. But the theory tells us there are infinitely many eigenfunctions for K associated with orthogonal tapers. How good are the eigenvalues for these functions? The Table lists those for $p = 4$.

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 $p = 4$.

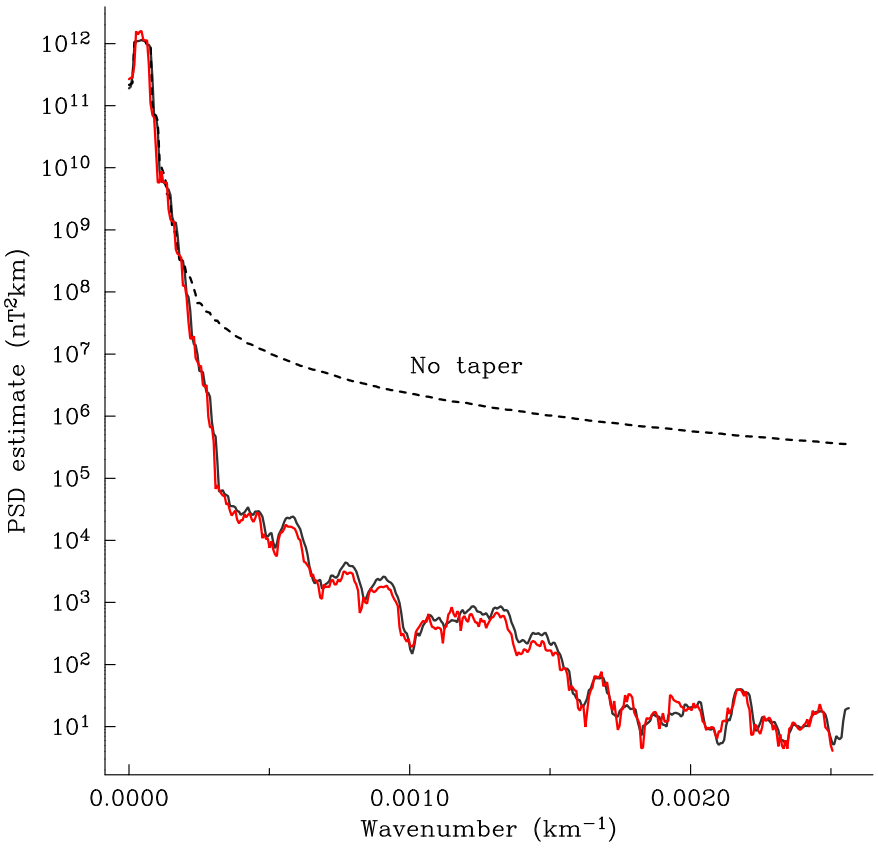


variance improvement by adding a further term, but by and large the answer is usually around $1.2p$, a factor times the time-bandwidth product, because after this number the concentration falls off so rapidly. Figure 8 shows the first six tapers for the time bandwidth product $p=4$. Notice how the higher 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

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

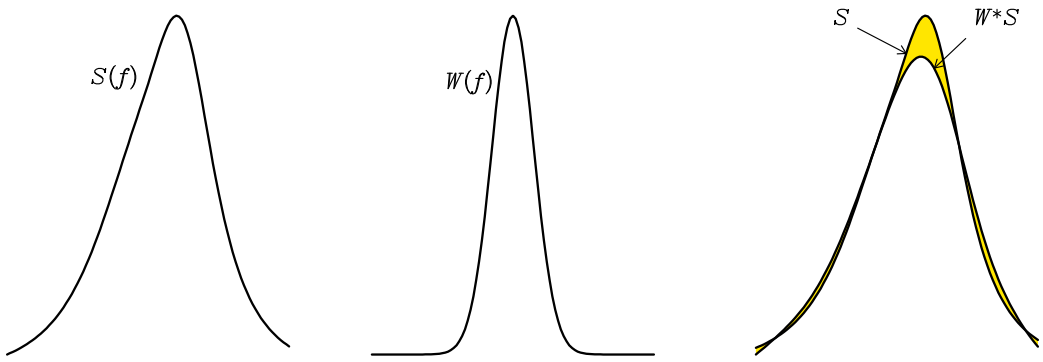
For practical tapers, W also dies away to zero fairly fast; R&S then say we can approximate the factor in brackets with a local Taylor series:

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

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:

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

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

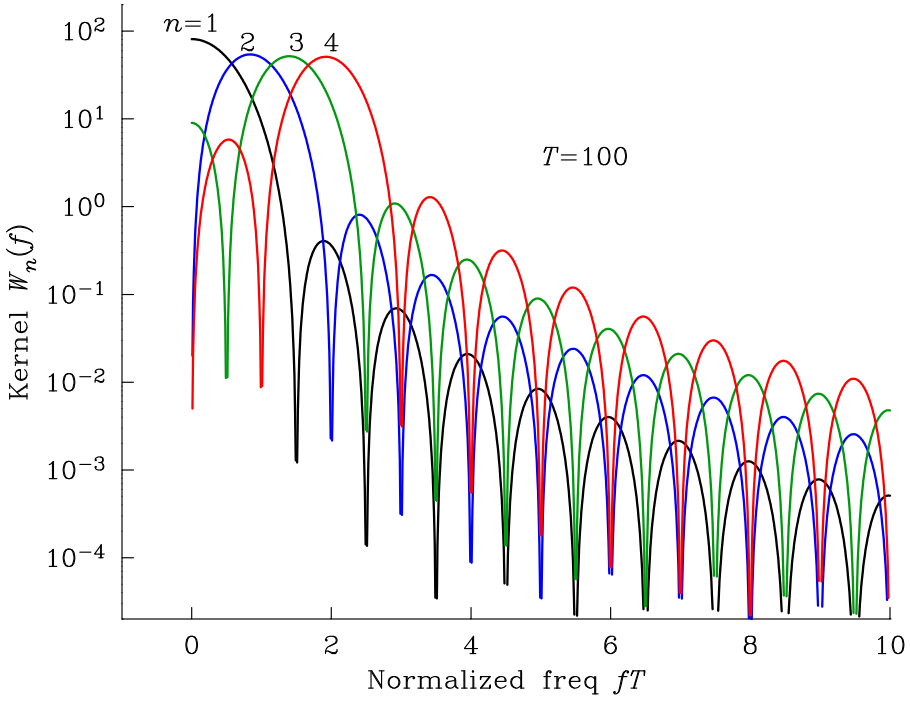
$$\phi_n(t) = \left(\frac{2}{T}\right)^{1/2} \sin \frac{\pi n t}{T}, \quad 0 \le t \le T. \tag{8.5}$$

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:

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

Thus when $fT \gg n$ the kernel decays like $n^2/2\pi^2T^3f^4$ 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

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

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

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

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

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

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)

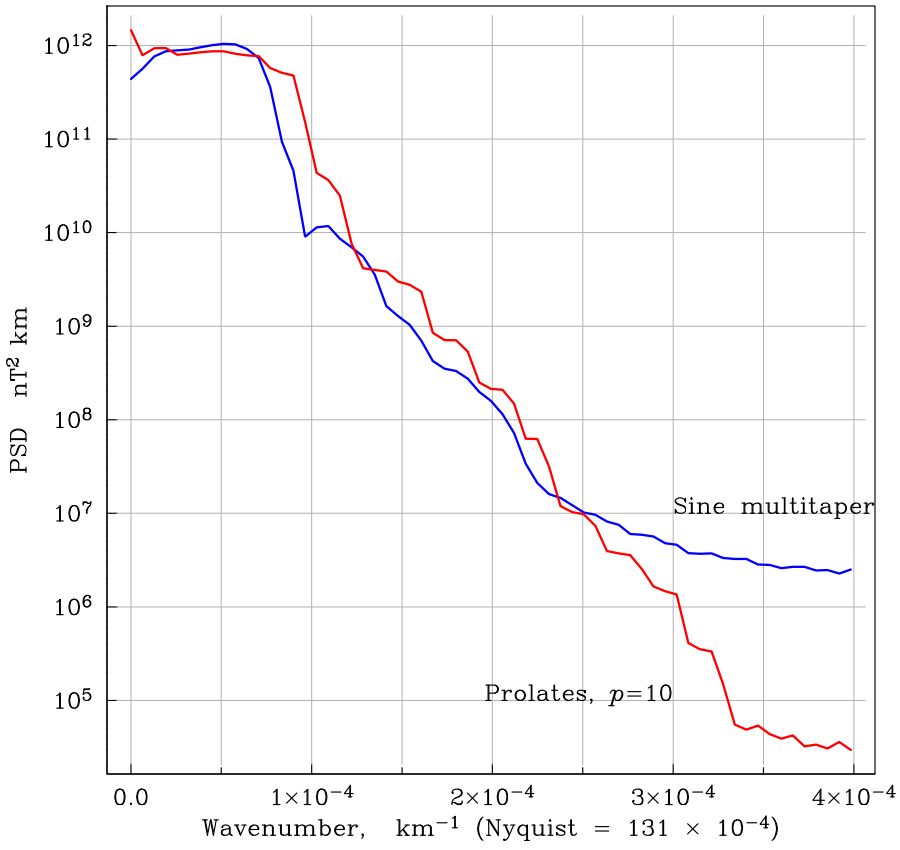
$$S_Y(f) = |\hat{a}(f)|^2 S_X(f) \tag{9.5}$$

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

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:

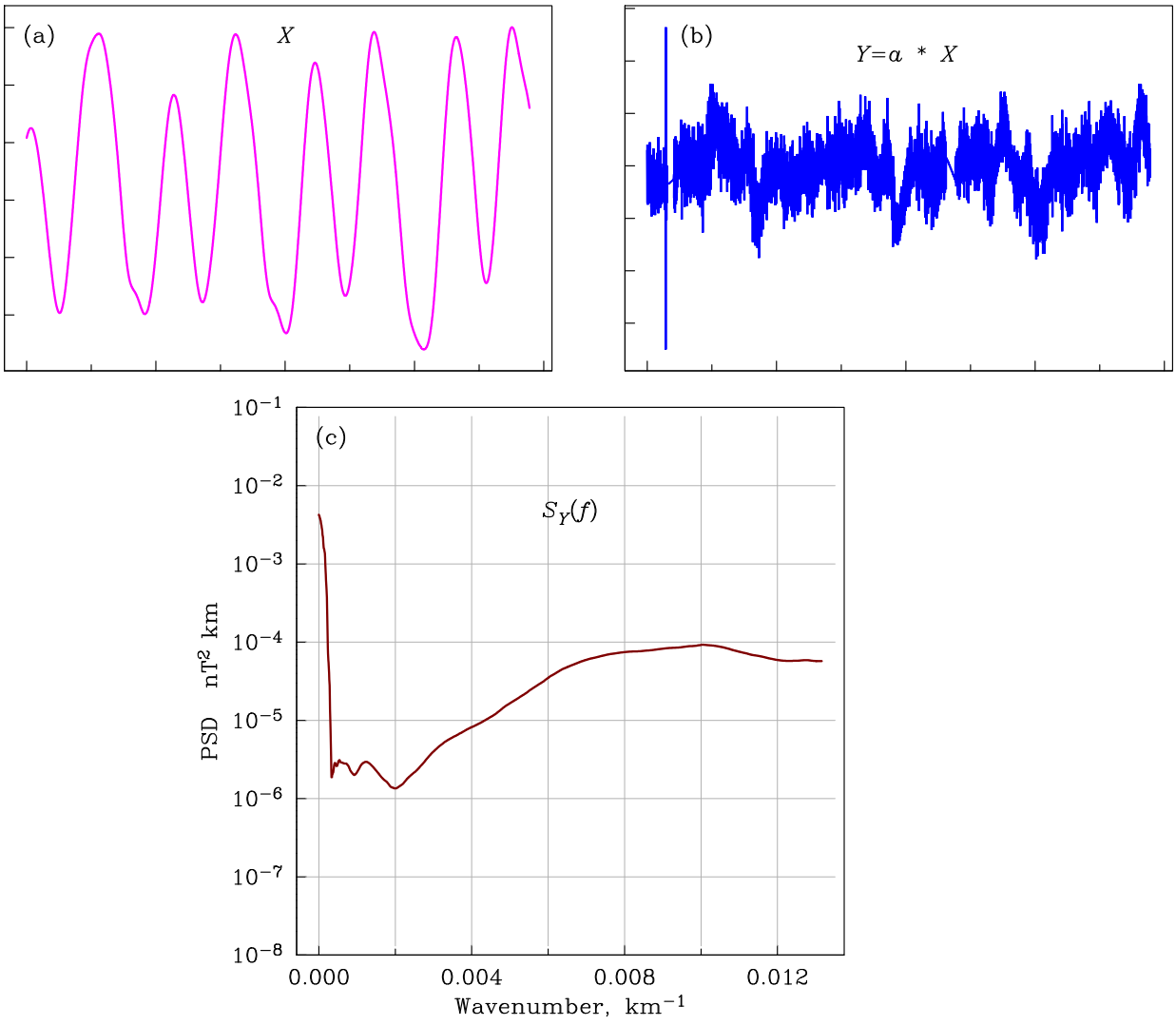
$$\hat{S}_X(f) = \frac{\hat{S}_Y(f)}{\left| 1 - \sum_{m=1}^{k} a_m e^{-2\pi i m f} \right|^2} \tag{9.7}$$

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.

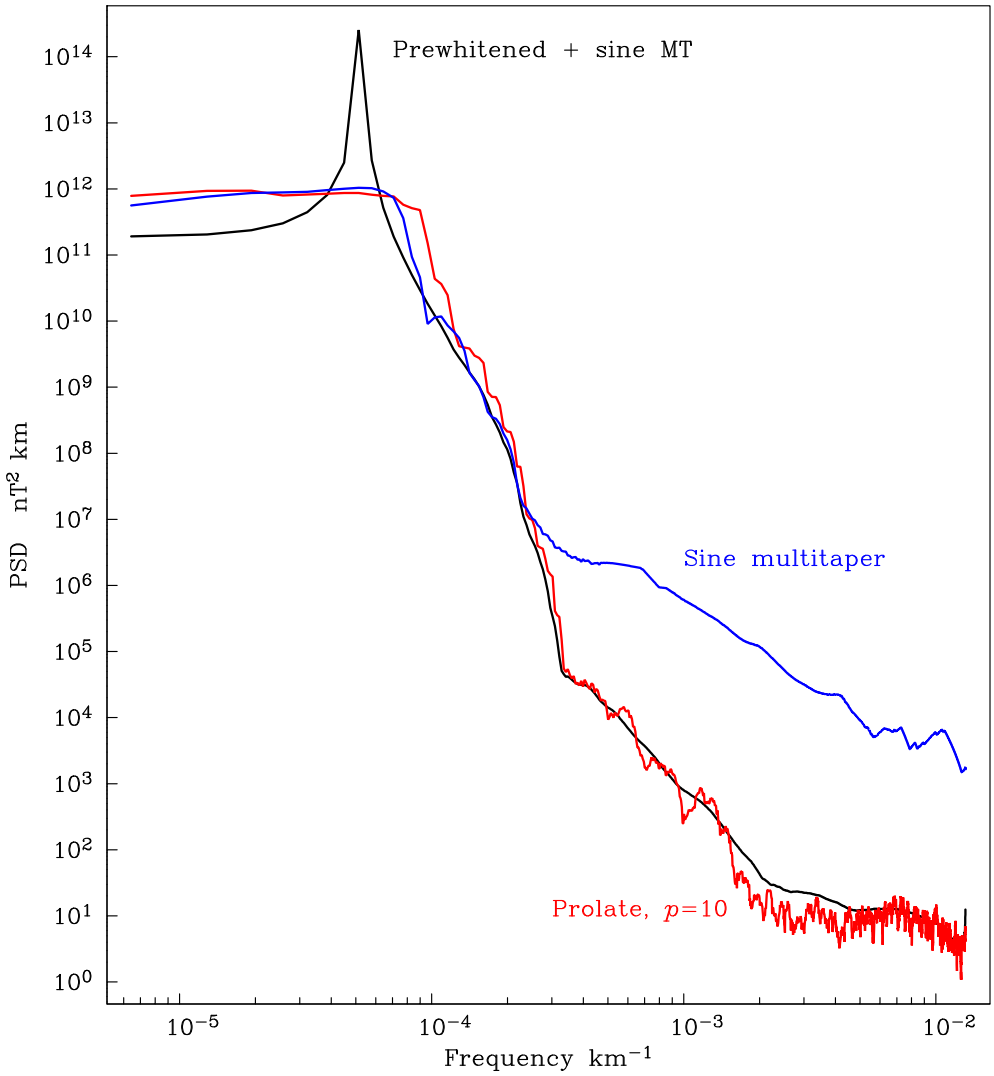
Figure 13: (a) Original time series; (b) Prewhitened series with 4 coefficients; (c) PSD of prewhitened record



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.