Chapter 13
Estimating the Power Spectral Density

13.1. Introduction

We now turn to the problem of actually estimating the power spectral density of a stochastic process. Note that this is, in a fundamental sense, more difficult than the kind of parameter-estimation problems we discussed earlier in the section on statistical inference. In those problems we were trying to estimate, or set bounds on, a finite set of parameters; whereas now we are trying to estimate a function $h(f)$, which cannot in general be described by such a finite set of numbers. We will, given finite data, inevitably have to settle for incomplete resolution of the details of this function.

We assume that the process is discrete, with sample interval $\Delta t = 1$ for simplicity. We also assume that (in principle at least) the process is infinite in length, though in practice we will have only N terms. We also assume that there are no purely periodic components (i.e., lines in the spectrum), that the spectral density $h(f)$ is continuous, and that the process is ergodic.

This last term has not been used before, so we explain it now. All the theoretical results we have derived have relied on our ability to take the expected value $\mathcal{E}[.]$ over repeated realizations from an abstract statistical process. But in many practical situations, we have only a single realization and no means of obtaining more. How then can we get a good estimate of a property that requires averaging?

If we can assume that averaging in time is the same as averaging over realizations we have a possible way out: we could cut our single realization into pieces and assume that each piece is an independent realization of the original process. An **ergodic** process is one for which long term temporal averages give the same result as ensemble averages. As you might suspect the concepts of stationarity and ergodicity are rather closely linked. One can imagine that if the autocorrelation function falls off fast enough then the idea that temporal sections of a single realization provide independent samples will be reasonable. For a Gaussian stationary process all we require is that the mean and covariance (over time) be ergodic. For any discrete process, a sufficient condition for ergodicity is that

$$\int_{-\frac{1}{2}\Delta t}^{\frac{1}{2}\Delta t} h_X(f)^2 df < \infty$$

and, as we have already required, that there are no periodic components to the spectrum.

13.2. Parametric Spectral Estimation

We first touch on a very popular method of spectral estimation, with a very large literature. We describe it briefly because we think it is not a good way to estimate the spectrum for geophysical cases. Basically, it consists of modelling the spectrum by a relatively small number of parameters, determined in the time domain. This would be an excellent method of spectrum estimation if we knew that the process we were examining in fact had a spectrum of the assumed form– but we almost never do know this, and if it is not true, such spectral fitting will give answers that may be badly biased, and which at best must be regarded as determined as much by our assumptions as by the data. In particular, it becomes very difficult to make realistic estimates of the uncertainty of the estimated spectrum.

The basic approach derives from earlier results on the effects of filters on the spectrum: we showed that if we had a continuous process X that could be regarded as a filtered version of another process Y

$$X = g * Y$$

then their spectral densities would be related by

$$h_X(f) = |G(f)|^2 h_Y(f)$$

Suppose now that $h_Y(f) = \sigma^2$, i.e. , Y is a continuous white noise process. Then $|G(f)|^2$ would give us the spectrum for X. Thus one way to find the spectrum for X is to find a convolution γ that transforms X into white noise; i.e. ,

$$\gamma * X = g^{-1} * X \approx \text{ white noise}$$

Then $\Gamma(f) = \mathcal{F}[\gamma]$ and the spectrum would be

$$h_X(f) = \left| \frac{1}{\Gamma(f)} \right|^2$$

This idea is usually implemented by assuming that γ can be represented by a finite-length autoregressive (AR) or moving average (MA) filter. In the AR case this is equivalent to modeling the process in the time domain by

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} + \ldots + a_k X_{t-k} = \epsilon_t \quad \text{for all } t$$

with ϵ_t independent and identically distributed (and usually Gaussian). The parameters $a_i,\ i = 1, \ldots, k$ are estimated using maximum likelihood or least squares. Once we have found these, it is straightforward to compute the transfer function (using the methods of Chapter 11) and thus find the spectrum. Note that even if we know the process, we still have the problem of deciding what order (k) the filter must have been. A number of semi-automatic methods have been proposed for choosing k, based on repeated solution at various k's. Maximum entropy, Yule-Walker estimation, Burg's method, and other methods, are all variations on this procedure.

There is a useful, if subsidiary, role for these techniques, which is to provide what is called **prewhitening**. As we will see, the methods of spectral estimation we prefer have the defect that they can be biased if the power spectral density covers a large range, with values at different frequencies differing by many orders of magnitude. This is not uncommon for actual data. The methods just mentioned can be used to devise a filter which, when used on the data, gives a series whose spectrum is much closer to being flat. We may then estimate the spectrum of this new series, with much less fear of bias, and find the spectrum of the actual data by dividing the estimated spectrum by the filter response.

13.3. The Raw Periodogram

Earlier we showed that our two definitions for the power spectral density function were equivalent

$$h_X(f) \;=\; \mathcal{F}[R_X(k)] \;=\; \lim_{N\to\infty} \frac{1}{N}\mathcal{E}\left[\left|\sum_{n\,=\,0}^{N-1} X_n e^{-2\pi i n f}\right|^2\right]$$

The second definition, in terms of the expected value of the Fourier transform of the process, suggests an estimator

$$\hat{h}_X(f) \;=\; \frac{1}{N}\left|\sum_{n\,=\,0}^{N-1} x_n e^{-2\pi i n f}\right|^2$$

This is known as the **periodogram** estimator. In this simple form it is not particularly good, but modifications of it form the basis of almost all efficient modern spectral methods. We therefore need to spend some time understanding the periodogram well.

13.4. 3 The Periodogram for White Noise

We start with the simplest possible case: the periodogram for a discrete stationary process with Gaussian statistics. That is, we suppose the process X_k to be independent and identically distributed Gaussian random variables, with zero mean and variance σ^2, so $X_k \;\sim\; N(0,\sigma^2)$. Then the autocovariance is

$$R_k \;=\; Cov[X_j, X_{j+k}] \;=\; \begin{cases} \sigma^2 & k \;=\; 0 \\ 0 & k \;\neq\; 0 \end{cases}$$

and the actual spectral density function is

$$h_X(f) \;=\; \sigma^2 \quad \text{for all } f \;\in\; [-\frac{1}{2},\frac{1}{2}]$$

We have seen this process above, and named it a discrete white noise process.

We now evaluate the periodogram values $\hat{h}_X(f)$ at $N+1$ equally spaced frequencies f_m; we write $\hat{h}_X(m) \;=\; \hat{h}_X(f_m)$, where $f_m \;=\; m/N$ for $m \;=\; 0\;\pm 1,\; \pm 2, \ldots, \pm N/2$ (we take N to be even); the limiting frequencies are thus $\pm\frac{1}{2}$, just at the Nyquist band. This may seem like a numerical convenience based on the FFT, but we shall see shortly that there are other reasons. Then the real and imaginary parts of the DFT are

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

and the periodogram estimate is

$$\hat{h}_X(m) \;=\; (A_m^2 + B_m^2)/N$$

It is easy to characterize the statistical distributions of A_m and B_m. The x_k are a sample of X which is an independent and identically distributed (iid) Gaussian. Both A_m and B_m are just linear combinations of Gaussians, so they too will be samples from a random variable with Gaussian distribution. Therefore, to characterize A_m and B_m, all we need to know is the mean, variance, and covariance of A_m and B_m.

Remembering that X_k was zero-mean, we have

$$E[A_m] \;=\; Re \sum_{n\,=\,0}^{N-1} \mathcal{E}[X_n] e^{-2\pi inm/N} \;=\; 0$$

and similarly $\mathcal{E}[B_m] \;=\; 0$. For the variance,

$$Var[A_m] \;=\; \mathcal{E}[A_m^2] \;=\; \mathcal{E}[(Re \sum_{n\,=\,0}^{N-1} X_n e^{-2\pi inm/N})(Re \sum_{l\,=\,0}^{N-1} X_l e^{-2\pi ilm/N})]$$

$$= \sum_{n\,=\,0}^{N-1} \sum_{l\,=\,0}^{N-1} E[X_n X_l] cos\frac{2\pi nm}{N} cos\frac{2\pi lm}{N} \;=\; \sum_{n\,=\,0}^{N-1} \sigma^2 cos^2 \frac{2\pi nm}{N} \tag{1}$$

To proceed beyond this, we need the orthogonality relationships for the sums of sines and cosines, which we simply state without demonstration. In particular, we have for N even that

$$\sum_{n\,=\,1}^{N} cos\frac{2\pi nm}{N} cos\frac{2\pi nl}{N} \;=\; \begin{cases} 0 & 0 \,\le\, m \,\ne\, l \,\le\, N \\ N/2 & 0 \,<\, m \,=\, l \,<\, N/2 \\ N & m \,=\, l \,=\, 0 \text{ or } N/2 \end{cases}$$

from which we obtain, substituting into (1), that

$$Var[A_m] \;=\; \begin{cases} N\sigma^2 & \text{for } \; m \;=\; 0,\; \pm N/2 \\ N\sigma^2/2 & \text{for } \; m \;=\; \pm 1,\; \pm 2, \ldots,\; \pm N/2 - 1 \end{cases}$$

and also that $Cov[A_m, A_l] \;=\; 0$ for all $m \;\ne\; l$. We may go through similar calculations for the variance of B_m, and use the fact that

$$\sum_{n\,=\,1}^{N} sin\frac{2\pi nm}{N} sin\frac{2\pi nl}{N} \;=\; \begin{cases} 0 & 0 \,\le\, m \,\ne\, l \,\le\, N \\ N/2 & 0 \,<\, m \,=\, l \,<\, N/2 \\ 0 & m \,=\, l \,=\, 0 \text{ or } N/2 \end{cases}$$

to obtain

$$Var[B_m] \;=\; \begin{cases} 0 & \text{for } \; m \;=\; 0,\; \pm N/2 \\ N\sigma^2/2 & \text{for } \; m \;=\; \pm 1,\; \pm 2, \ldots,\; \pm N/2 - 1 \end{cases}$$

and $Cov[B_m, B_l] \;=\; 0$ for all $m \;\ne\; l$. The final orthogonality relationship is

$$\sum_{n\,=\,1}^{N} sin\frac{2\pi nm}{N} cos\frac{2\pi nl}{N} \;=\; 0 \qquad \text{for all } l, m$$

from which we obtain $Cov[A_m, B_l] \;=\; 0$ for all $m,\; l$. Thus, A_m and B_m are zero-mean statistically independent Gaussian random variables, with (except at the Nyquist frequency) equal variances of $N\sigma^2/2$. Note that the statistical independence is a function of our choice of frequency spacing, since this enabled us to use the orthogonality relations in the above calculations; if we computed these quantities at more closely-spaced frequencies they would not be independent.

Now we need to find out how the statistical distribution of $\hat{h}_X(m) = (A_m^2 + B_m^2)/N$ is related to those of A_m and B_m. We already know (see Chapter 4) that the sum of the squares of ν independent and identically distributed zero-mean Gaussian random variables are distributed according to the chi-square (χ^2) distribution which has the following normalized pdf:

$$p(x) = \begin{cases} [2^{\nu/2}\Gamma(\nu/2)]^{-1} x^{\nu/2-1} e^{-x/2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$ and for any positive integer $\Gamma(k) = (k-1)!$ by repeated integration by parts.

In our case $\nu = 2$. Let $\rho^2 = N\sigma^2/2$, the variance of A and B. We now need the distribution for the sum of the squares of two independent and identically distributed Gaussian random variables with equal variance. In other words, if we let $\chi^2 = A^2 + B^2$, we want the distribution function $\Psi(\chi^2)$, which is given by

$$\Psi(\chi^2) = Pr(A^2 + B^2 \leq \chi^2) = \int_0^\chi \frac{1}{2\pi\rho^2} e^{-r^2/2\rho^2} 2\pi r dr = 1 - e^{-\chi^2/2\rho^2}$$

where to do the integral we have gone from A and B to polar coordinates. The pdf is given by

$$\psi(\chi^2) = \frac{d\Psi}{d\chi^2} = \frac{1}{2\rho^2} e^{-\chi^2/2\rho^2}$$

Thus, we have found that a χ^2 distribution with 2 degrees of freedom is an exponential distribution; this is the sampling distribution for the periodogram. To look for bias and consistency of the periodogram we need the mean and variance of this distribution. The results are well known:

$$\mathcal{E}[\chi^2] = 2\rho^2 \qquad Var[\chi^2] = 4\rho^4$$

which means that for frequencies not at zero or the Nyquist ($m \neq 0, N/2$), we get

$$\mathcal{E}[\hat{h}_X(f_m)] = \frac{1}{N} N \frac{2N\sigma^2}{2} = \sigma^2 \qquad \text{and} \qquad Var[\hat{h}_X(f_m)] = \frac{1}{N^2} 4 \left(\frac{N\sigma^2}{2} \right)^2 = \sigma^4$$

and for $m = 0, N/2$

$$\mathcal{E}[\hat{h}_X(f_m)] = \sigma^2 \qquad \text{and} \qquad Var[\hat{h}_X(f_m)] = 2\sigma^4$$

The good news is that $E[\hat{h}] = h$: the periodogram estimate of the power spectral density function for a white process is unbiased. The bad news is twofold. First, this estimate has a very large variance; taking the standard deviation, $(Var[\hat{h}_X(m)])^{\frac{1}{2}}$, as a measure of the uncertainty in our estimate for $h_X(f)$, we see that this is σ^2: the uncertainty of the estimate is the same as the estimate itself. And worse yet, this uncertainty is independent of N: as $N \rightarrow \infty$, the variance $Var[\hat{h}]$ remains constant, so the periodogram is not a consistent estimator.

An intuitive way to see why this would be so is to note that, for N data, the periodogram gives us $N/2$ spectral estimates. We might thus expect that each estimate has about the same amount of information, equivalent to two data, no matter how large (or small) N is. For many estimation problems, we have a fixed number of parameters and can expect them to be better estimated as we add more data (though this is not always true). Since here we are trying to estimate a function (in this case the power spectral density function) there is, in principle, no limit to the number of parameters needed– but the example of the periodogram shows us that unless we apply some limits, we cannot form a good estimate.

13.5. The Periodogram for Non-White Noise: General

The periodogram estimate for the uncorrelated Gaussian process is a very special case. How can we carry these results over to the more general problem of spectrum estimation for a general linear process with a continuous spectrum?

We make a heuristic argument based on the idea already used in Section 13.1: any general process with a continuous spectrum can be regarded as a filtered version of a white noise process. We can write

$$Y_k = \sum_{j=-\infty}^{\infty} g_j X_{k-j}$$

Then

$$h_Y(f) = |G(f)|^2 h_X(f)$$

with

$$G(f) = \sum_{k=-\infty}^{\infty} g_k e^{-2\pi i k f}$$

We might expect that the periodogram estimates $\hat{h}_Y(f)$ and $\hat{h}_X(f)$ would be related in the same way as the spectral densities h_X and h_Y; this is in fact true. It is possible to show that the periodogram is an asymptotically unbiased estimator for smooth (i.e., continuous h_Y) spectra. The corresponding result for the variance of the estimator also holds; i.e.,

$$Var[\hat{h}_Y(m)] \rightarrow [h_Y(m)]^2$$

13.6. The Periodogram for Non-White Noise: Bias

While periodogram estimates are asymptotically unbiased, it turns out that they can be very biased for finite N (which is all we ever have). We can see the form that the bias takes by representing the periodogram

$\hat{h}_X(f)$ as a Fourier transform of the estimated autocovariance function. A discrete analog of the proof of the equivalence of our two definitions for the power spectral density shows that

$$\hat{h}_X(f) \;=\; \frac{1}{N}\sum_{k\,=\,1-N}^{N-1}\left[\sum_{j\,=\,1}^{N-|k|} X_j X_{j+k}\right] e^{-2\pi i k f}$$

which in terms of the unbiased estimate of the autocovariance function

$$\hat{R}_X(k) \;=\; \frac{1}{N-|k|}\sum_{j\,=\,1}^{N-|k|} X_j X_{j+k}$$

gives that

$$\hat{h}_X(f) \;=\; \sum_{k\,=\,1-N}^{N-1}\left(\frac{1-|k|}{N}\right)\hat{R}_X(k)e^{-2\pi i k f} \tag{2}$$

To find the bias in $\hat{h}_X$ we need its expected value

$$\mathcal{E}[\hat{h}_X(f)] \;=\; \sum_{k\,=\,1-N}^{N-1}\left(\frac{1-|k|}{N}\right)\mathcal{E}[\hat{R}_X(k)]e^{-2\pi i k f} \;=\; \sum_{k\,=\,1-N}^{N-1}\left(\frac{1-|k|}{N}\right)R_X(k)e^{-2\pi i k f}$$

because $\hat{R}$ is the unbiased form of the autocovariance estimator. But since the power spectral density $h_X(f)$ is a Fourier transform (from discrete time to continuous frequency) of R_X, we can write the inverse transform

$$R_X(k) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} h_X(f')e^{2\pi i k f'}df'$$

Now, we substitute this expression for R_X into the equation for $\mathcal{E}[\hat{h}_X]$:

$$\begin{aligned}
\mathcal{E}[\hat{h}_X(f)] \;&=\; \sum_{k\,=\,1-N}^{N-1}\frac{(1-|k|}{N})\int_{-\frac{1}{2}}^{\frac{1}{2}} h_X(f')e^{2\pi i k f'}df'e^{2\pi i k f} \\
&=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} h_X(f')\left\{\sum_{k\,=\,1-N}^{N-1}(\frac{1-|k|}{N})e^{-2\pi i (f-f')k}\right\}df' \\
&=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} h_X(f')F_N(f'-f)df'
\end{aligned} \tag{3}$$

where we have defined a function $F_N(f)$ which is called the Fejer kernel. This can be evaluated exactly:

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

The Fejer kernel is N times the square of the Dirichlet kernel encountered earlier:

$$D_N(f) \;=\; \frac{sin\pi N f}{N sin\pi f}$$

The Dirichlet kernel corresponds to a boxcar function ($\Pi(k/N)$), while the Fejer kernel goes with a triangle or convolution of two boxcars in the time domain

$$\Lambda_{N/2}(k) \;=\; \frac{1}{N}\Pi(k/N) * \Pi(k/N) \;=\; \begin{cases} 1 - |k|/N & k \;\leq\; N \\ 0 & k > N \end{cases}$$

As we saw in Chapter 8, for N large the Dirichlet kernel approaches the sinc function so we can approximate the Fejer kernel by $Nsinc^2$ for large N, corresponding to finer and finer sampling. The approximation is adequate even for modest values of N.

Equation (3) tells us that, for a nonwhite power spectral density, the expected value for the raw periodogram estimator is biased: $\hat{h}_X(f_0)$ is the convolution of the Fejer kernel with $h_X(f)$ at surrounding frequencies. Unless the spectrum is completely flat there will be bias. If the true spectrum $h_X(f)$ has a strong peak or a deep trough the results may be very deceptive: the low parts of the spectral estimate will be raised by energy actually present in the high parts. This phenomenon is known as **leakage** .

13.7. Improving the Periodogram

The raw periodogram is an extremely poor estimate of the power spectral density function. While it is asymptotically unbiased, it is not unbiased for nonwhite spectra and finite N, and may reach the asymptote only very slowly. It also has two other undesirable features:

1. It is not a consistent estimate of $h(f)$ in the sense that $Var[\hat{h}(f)]$ does not tend to zero as $N \to \infty$.

2. As a function of f, $\hat{h}(f)$ fluctuates wildly; to see that this is what to expect, remember that for any two fixed neighboring frequencies f_1, f_2, $Cov[\hat{h}_N(f_1), \hat{h}_N(f_2)]$ decreases as N increases.

We now discuss ways to improve it.

13.7:1 Reducing Bias: Tapering

Equation (3) shows that the bias is due to convolution of the true spectrum with the Fejer kernel. If we can find a way to replace the Fejer kernel with a function with smaller sidelobes then the leakage will be less and contamination from distant frequencies will not be such a problem.

One can guess (we will not prove) that multiplication of the original data series by an appropriate taper (also called a data window) might do the job. The Fejer kernel results from an implicit boxcar taper which truncates the data to a series N long. If we choose a taper which dies smoothly away to zero at the ends of the data series, then it will have smaller side lobes in the frequency domain. If the weights used in the taper are w_k, the expected value of the estimated spectrum becomes

$$\mathcal{E}[\hat{h}_X(f)] \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} W(f' - f) h_X(f') df'$$

where we have approximately

$$W(f) \;=\; |\mathcal{F}[w]|^2$$

The weights w_k are chosen so that $W(f)$ falls off rapidly. As was the case in digital filter design, protection against bias from distant frequencies comes at the expense of broadening the central frequency lobe.

Another consequence of tapering is effectively to reduce the length of the data series. It should therefore not be surprising that the variance in the resulting estimate is always larger than for the untapered series (we discuss this further below). Our hope will be that the reduction in bias will result in a smaller overall mean square error. This will of course depend on the nature of the spectrum: if it is very nearly white the bias will be small to begin with, and perhaps not worth reducing– but in most cases this is not so.

13.7:2 Improving Consistency: Section Averaging

An obvious way to reduce the variance of the raw periodogram estimator is to obtain several independent periodogram estimates and then average them. There are a number of ways of doing this. One is to break the available time series into M parts of equal length, find the periodogram estimate on each one and then average the set together. This is consistent with the spirit of the definition of $h_X(f)$ and the ergodic assumption

$$h_X(f) \;=\; \lim_{N\to\infty} \frac{1}{N}\mathcal{E}\left[|\tilde{X}(f)|^2\right]$$

We use

$$\overline{\hat{h}_X(f)} \;=\; \mathcal{E}[\hat{h}_X(f)] \;\approx\; \frac{1}{M}\sum_{j\,=\,1}^{M} \hat{h}_X^{(j)}(f)$$

For sufficiently large M we might guess that it is reasonable to invoke the Central Limit Theorem and say

$$Var[\overline{\hat{h}_X(f)}] \;\approx\; \frac{Var[\hat{h}_X(f)]}{M} \;\approx\; \frac{h_X(f)^2}{M}$$

Clearly by making M large enough we can reduce the variance in the spectral estimate to any level required.

We can also show that section averaging is the best way, in the maximum likelihood sense, of reducing variance in the spectral estimate. Suppose we have M samples drawn from a χ_2^2 distribution, with unknown parameter ρ^2. Our spectral estimation problem is to find ρ^2 and we want to find the maximum likelihood estimator for it. The pdf for one sample χ_j^2 is

$$\psi(\chi^2) \;=\; \frac{1}{2\rho^2}e^{-\chi^2/2\rho^2}$$

Thus the joint log likelihood function for the M observations is

$$L(\vec{\chi}^2,\rho^2) \;=\; Mln\rho^2 - Mln2 - \frac{1}{2\rho^2}\sum_{j\,=\,1}^{M}\chi_j^2$$

We find the ML estimate from $\partial L / \partial \rho^2 = 0$ yielding

$$-\frac{M}{\rho^2} + \frac{\rho^4}{2} \sum_{j=1}^{M} \chi_j^2 = 0$$

i.e.,

$$\hat{\rho}^2 = \frac{1}{2M} \sum_{j=1}^{M} \chi_j^2$$

For the χ^2 distribution we had $E[\chi^2] = 2\rho^2$ which gives as the estimate

$$2\hat{\rho}^2 = \frac{1}{M} \sum_{j=1}^{M} \chi_j^2$$

making averaging the ML estimator for $E[\chi^2]$.

However, this averaging has a price (to some extent unavoidable). The raw periodogram provided an estimate at each of $N/2$ frequencies for $h_X(f)$. Section averaging reduces this by a factor of M. We thus reduce both the **resolution** of our spectral estimate, and also potentially increase the **bias** (because the Fejer kernel becomes broader).

13.7:3 Weighted Overlapped Section Averaging

Our suggested “simple” method for estimating the power spectral density combines the procedures of the previous two sections: we break the data into sections, taper each section, form the DFT of the result, and average over all sections. The tapering provides protection against bias; the averaging provides a lower-variance estimate than the raw periodogram. Because of the tapering applied to each section, it is appropriate to allow the sections to overlap. This method is known by the title given above, and also as Welch’s method, after its inventor.* We call this a “simple” method because it is not difficult to describe; while more information can be obtained using the multitaper method (to be described later), this method requires more background to explain and understand and more computation to estimate. While Welch’s method is now quite old, it remains very useful if there are large amounts of data relative to the level of detail in the spectrum that is needed.

For this method it is necessary to choose the data taper and the number of sections. There is no universally right answer for either of these; the choice depends largely on what the spectrum looks like, which makes the analysis an iterative process informed by personal judgement. If the dynamic range of the spectrum is large, then you need the good bias protection offered by, for example, the 4π prolate data taper. If the dynamic range is lower, you might decide to retain better resolution near the central lobe in frequency

* Welch, Peter D. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified, periodograms, *IEEE Trans. Audio Electroacoustics* , **AU-15** , 70-73.

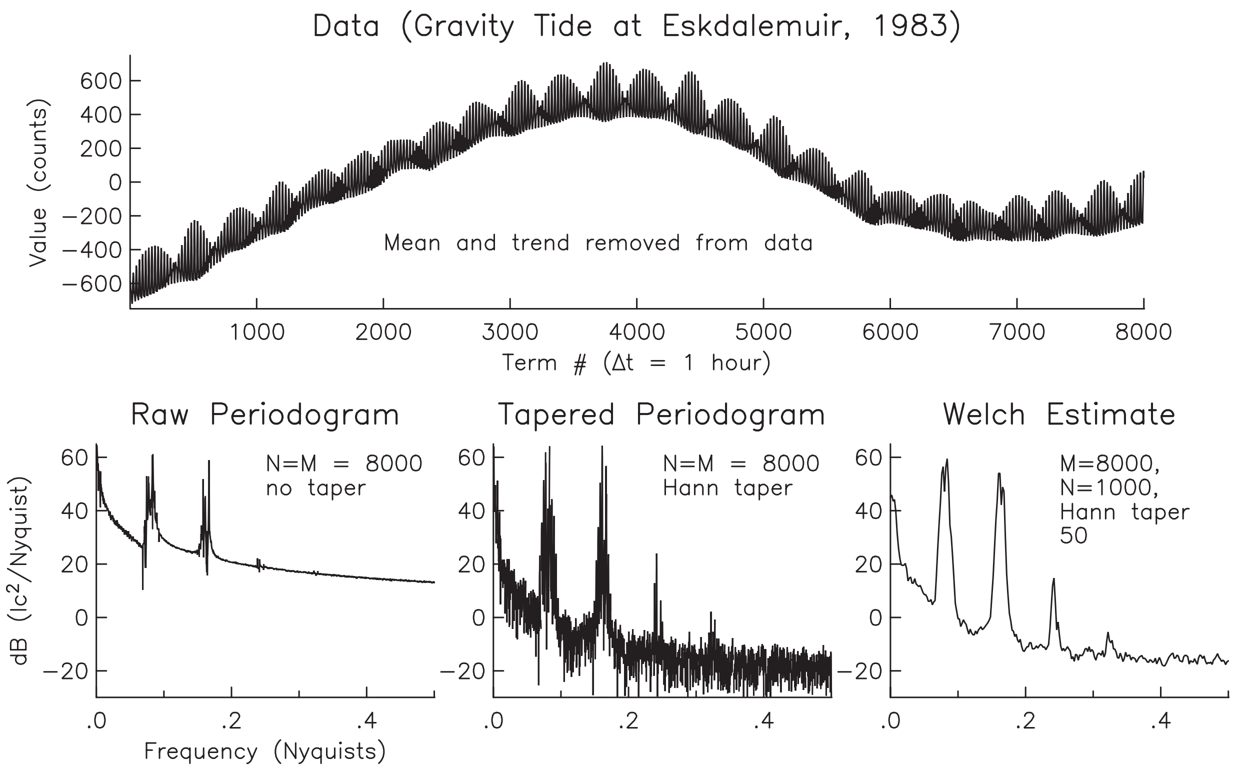

Figure 13-1: A data series (of gravity, from the Project IDA gravimeter at Eskdalemuir in Scotland), and three estimates of its power spectral density. The raw periodogram is so biased as to be useless. Tapering removes the bias but the estimates are so variable as to still not be useful. Welch's method gives a good estimate of the spectrum, at the price of losing resolution (shown by the loss of many separate peaks in the tidal bands).

of the data taper. The Hann taper is a common choice in this case. The choice of M, the number of data sections, depends on the kind of resolution/variance tradeoff you want to make. Too many sections will ultimately give poor frequency resolution, too few will give high variance in the spectral estimate.

We now discuss briefly the variance of the estimates from this method (following Welch). We take the estimator to be

$$\hat{h}(f) \;=\; \frac{1}{M}\sum_{j=1}^{M} \hat{h}_j(f)$$

Each estimate $\hat{h}_j$ is a direct spectral estimate based on some subset of the data $X_0, X_1, X_2, \;\ldots\;, X_{N-1}$ and we assume that some appropriate data taper w_k has been applied to each section in order to reduce bias in the spectral estimate. Each section is displaced by D samples with respect to the origin of the previous one, and is L points long. If we do not have overlapping sections (i.e., if $D = L$), we have M independent estimates of $\hat{h}$, each with 2 degrees of freedom: the variance in the spectral estimates is reduced by a factor of M by the section averaging process:

$$Var\,[\hat{h}(f)] \;\approx\; \frac{1}{M}\,\left[\mathcal{E}[\hat{h}(f)]\right]^2$$

However, usually we do overlap the tapered data sections. If each periodogram estimator is

$$\hat{h}_j \;=\; \frac{1}{L}\mathcal{F}[|w_n x_n|^2]$$

and the sections overlap, then even for a locally white spectrum there will be covariance in the spectral estimates: they can no longer be regarded as independent. Let

$$d_j \;=\; Cov\,\left[\hat{h}_k(f_n)\hat{h}_{k+j}(f_n)\right]$$

and

$$\hat{h}(f_n) \;=\; \frac{1}{M}\sum_{i=1}^{M}\hat{h}_i(f_n) \qquad \text{for} \quad n \;=\; 0,\, 1,\, \ldots L/2$$

Then it can be shown that

$$Var[\hat{h}(f_n)] \;=\; \frac{1}{M}\left\{d_0 \;+\; 2\sum_{j=1}^{M-1}\;\frac{M-j}{M}d_j\right\}$$

Now if

$$\rho_j \;=\; Correlation\;\left\{\hat{h}_k(f_n), h_{k+j}(f_n)\right\} \;=\; \frac{d_j}{d_0}$$

then this becomes

$$\begin{aligned} Var[\hat{h}(f_n)] \;&=\; \frac{d_0}{M}\left[1\;+2\sum_{j=1}^{M-1}\;\frac{M-j}{M}\rho_j\right] \\ &=\; \frac{Var[h_k(f_n)]}{M}\left[1\;+\;2\sum_{j=1}^{M-1}\frac{M-j}{M}\rho_j\right] \end{aligned}$$

If $X(j)$ is a sample from a Gaussian process and $h(f)$ is flat over the band sampled by our estimator, then we had

$$Var[\hat{h}_k(f)] \;=\; h^2(f_n)$$

It is also possible to show that

$$\rho_j \;=\; \left[\sum_{k=0}^{L-1} w_k w_{k+jD}\right]^2\;\left[\sum_{k=0}^{L-1} w_k^2\right]^{-2}$$

which is to say that the correlation depends only on the shape of the data taper for a locally white spectrum. Then,

$$Var[\hat{h}(f_n)] \; = \; \frac{\hat{h}^2(f_n)}{M} \; \left[1 \; + \; 2 \; \sum_{j=1}^{M-1} \; \rho_j \right]$$

and so when the spectrum is flat the variance is controlled by the shape and degree of overlap of the window. If the spectrum is locally complex, this estimate will be invalid. In the non-overlapping case $\rho_j \; = \; 0$. Therefore the *computationally* most efficient means of acquiring any desired variance is to have non-overlapping segments. It can be shown that for data windows shaped like $1 - t^2$ or $1 - |t|$ (data interval $-1 \leq t \leq 1$) and 50% overlap the variance is

$$Var[\hat{h}(f_n)] \; \sim \; \frac{11}{9} \; \cdot \; \frac{h^2(f_n)}{M}$$

So the variance is inflated by the overlap if the number of segments remains the same, but an overall reduction in variance is achieved for fixed record length since overlapping segments increases the value of M. This expression is also approximately valid for a Hann taper, and it is this, with a 50% overlap, which would be our suggested first approach to estimating the power spectral density of any data series of reasonable length.

13.8. Prewhitening

We noted in Section 13.1 the advantage of "prewhitening" the data series; the reason for this may be clearer now that we have discussed bias in the periodogram– the "whiter" the spectrum to be estimated, the less we need to be concerned about bias. In any case some kind of modification of the series is almost always needed, since most data series are non-stationary in some way, the simplest being a constant slope (non-stationary mean). The standard solution is to fit the "trend" (mean, best fitting line or other low-frequency component) using least squares fitting, remove this trend, and perform the spectrum estimation on the residual. Alternatively, the trend can be removed by first-differencing the series (or performing more elaborate prewhitening), computing the spectrum, and then correcting it for the frequency response. For first-differencing, this will always give an infinite value at zero frequency– but this cannot be estimated very well by these procedures anyway.

13.9. Frequency Averaging

Section averaging of the periodogram estimate is not the only means of obtaining a consistent estimate of the power spectral density using the periodogram. Instead of averaging over time sections we could smooth the resulting periodogram in the frequency domain, choosing an interval around each frequency of interest and averaging the estimates for neighbouring frequencies together. This is also called **band averaging** . A more complicated approach, would be to weight these estimates in some way according to their distance from the nominal frequency of interest

$$\hat{h}_X(f_0) \; = \; \sum_{j \, = \, -M}^{M} w_j \hat{h}_X(f_0 - \frac{\Delta j}{M})$$

The weight function w_j would be symmetric and chosen to provide a local average in frequency of the periodogram estimates. Not surprisingly, the smoothing results in a loss of resolution, and could introduce bias even if this has been kept small in the initial estimate (and we should be sure that it is).

13.10. Autocovariance Windowing

The approach to spectral estimation through smoothing of the periodogram led to a method of spectral estimation much used in the past: taking the Fourier transform of the estimated autocovariance function. This is a tempting course of action, given that we have defined the power spectral density in terms of the Fourier transform of the autocovariance: but as with the periodogram estimate, that some relationship holds does not always make it the basis for a good estimator. We have, from equation (3), that the periodogram estimate can be written as the Fourier transform of

$$\left(1 - \frac{|k|}{N}\right) \hat{R}_X(k)$$

where $\hat{R}_X(k)$ is the unbiased estimate of the autocovariance function. The part in parentheses downweights this function for large k, and makes the estimate better– though still not good, since it is just the periodogram. A more general technique is to reduce the variance of the spectral estimate by “windowing” the autocovariance function with a window function $w(k)$, called the **lag window** , which downweights values of $\hat{R}(k)$ for high k. This estimate is

$$\hat{h}_X(f) = \sum_{k=1-N}^{N-1} \left(1 - \frac{|k|}{N}\right) \hat{R}_X(k) w(k) e^{-2\pi i k f}$$

Now, by the convolution theorem, multiplying the autocovariance function by the lag window is the same as convolving the periodogram with the Fourier transform of this window. Thus autocovariance windowing can be regarded as the same operation as band averaging. The Fourier transform of the lag window is usually known as the spectral window; neither should be confused with the data taper discussed in section 13.3. The older (and even more recent) literature has lengthy discussions of different lag windows and their properties; we believe however that there is little if any advantage to this method over either Welch’s method or multitaper analysis.