Editor: Denis Donnelly, donnelly@siena.edu

The Fast Fourier Transform for Experimentalists Part III: Classical Spectral Analysis

By Bert Rust and Denis Donnelly

EACH ARTICLE IN THIS CONTINUING SERIES ON THE FAST FOURIER TRANSFORM (FFT) IS DESIGNED TO ILLUMINATE NEW FEATURES OF THE WIDE-RANGING APPLICABILITY OF THIS TRANSFORM. THIS SEGMENT DEALS WITH SOME ASPECTS OF THE spectrum estimation problem. Before we begin, here's a short refresher about two elements we introduced previously, windowing[1] and convolution.[2] As we noted in those installments, a convolution is an integral that expresses the amount of overlap of one function as it is shifted over another. The result is a blending of the two functions. Closely related to the convolution process are the processes of cross-correlation and autocorrelation. Computing the cross-correlation differs only slightly from the convolution; it's useful for finding the degree of similarity in signal patterns from two different data streams and in determining the lead or lag between such similar signals. Autocorrelation is also related to the convolution; it's described later. Windowing, used in extracting or smoothing data, is typically executed by multiplying time-domain data or its autocorrelation function by the window function. A disadvantage of windowing is that it alters or restricts the data, which, of course, has consequences for the spectral estimate. In this installment, we continue our discussion, building on these concepts with a more general approach to computing spectrum estimates via the FFT.

Spectrum Estimation's Central Problem

The periodogram, invented by Arthur Schuster in 1898,[3] was the first formal estimator for a time series's frequency spectrum, but many others have emerged in the ensuing century. Almost all use the FFT in their calculations, but they differ in their assumptions about the missing data; that is, the data outside the observation window. These assumptions have profound effects on the spectral estimates. Let t be time, f be frequency, and $x(t)$ a real function on the interval $-\infty < t < \infty$. The continuous Fourier transform (CFT) of $x(t)$ is defined by

$$X(f) = \int_{-\infty}^{\infty} x(t)\exp(-2\pi i f t)\,dt, \quad -\infty \le f \le \infty, \tag{1}$$

where $i \equiv \sqrt{-1}$. If we knew $x(t)$ perfectly and could compute Equation 1, then we could compute an energy spectral density function

$$E(f) = |X(f)|^2, \quad -\infty \le f \le \infty, \tag{2}$$

and a *power spectral density function* (PSD) by

$$P(f) = \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \left| \int_{-\mathcal{T}}^{\mathcal{T}} x(t)\exp(-2\pi i f t)\,dt \right|^2, \quad -\infty \le f \le \infty. \tag{3}$$

But we have only a discrete, real time series

$$x_j = x(t_j), \text{ with } t_j = j\Delta t, \quad j = 0, 1, \ldots, N-1, \tag{4}$$

defined on a finite time interval of length $N\Delta t$. We saw in Part I[1] that sampling $x(t)$ with sample spacing Δt confined our spectral estimates to the Nyquist band $0 \le f \le 1/2\Delta t$. We used the FFT algorithm to compute the discrete Fourier transform (DFT)

$$X_k = \sum_{j=0}^{N-1} x_j \exp\left(-2\pi i \frac{j}{N} k\right), \quad k = 0, 1, \ldots, N/2, \tag{5}$$

which approximates the CFT $X(f)$ at the Fourier frequencies

$$f_k = \frac{k}{N\Delta t}, \; k = 0, 1, \ldots, N/2. \tag{6}$$

We then computed periodogram estimates of both the PSD and the amplitude spectrum by

$$P(f_k) = \frac{1}{N}|X_k|^2, \; k = 0, 1, \ldots, N/2,$$

$$A(f_k) = \frac{2}{N}|X_k|, \; k = 0, 1, \ldots, N/2. \tag{7}$$

We also saw that we could approximate

the CFT and the frequency spectrum on a denser frequency mesh simply by appending zeroes to the time series. This practice, called zero padding, is just an explicit assertion of an implicit assumption of the periodogram method—namely, that the time series is zero outside the observation window. Frequency spectrum estimation is a classic underdetermined problem because we need to estimate the spectrum at an infinite number of frequencies using only a finite amount of data. This problem has many solutions, differing mainly in what they assume about the missing data.

Before considering other solutions to this problem, let's reconsider one of the examples from Part I[1] (specifically, Figure 1b), but make it more realistic by simulating some random measurement errors. More precisely, we take $N = 32$, $\Delta t = 0.22$, and consider the time series

$$t_j = j\Delta t, j = 0, 1, 2, \ldots, N-1,$$

$$x_j = x(t_j) = \sin[2\pi f_0(t_j + 0.25)] + \in_j, \quad (8)$$

with $f_0 = 0.5$, and each $\in_j$ a random number drawn independently from a normal distribution with mean zero and standard deviation $\sigma = 0.25$. This new time series is plotted together with the original uncorrupted series in Figure 1a. Both series were zero padded to length 1,024 (992 zeroes appended) to obtain the periodogram estimates given in Figure 1b. It's remarkable how well the two spectra agree, even though the noise's standard deviation was 25 percent of the signal's amplitude.

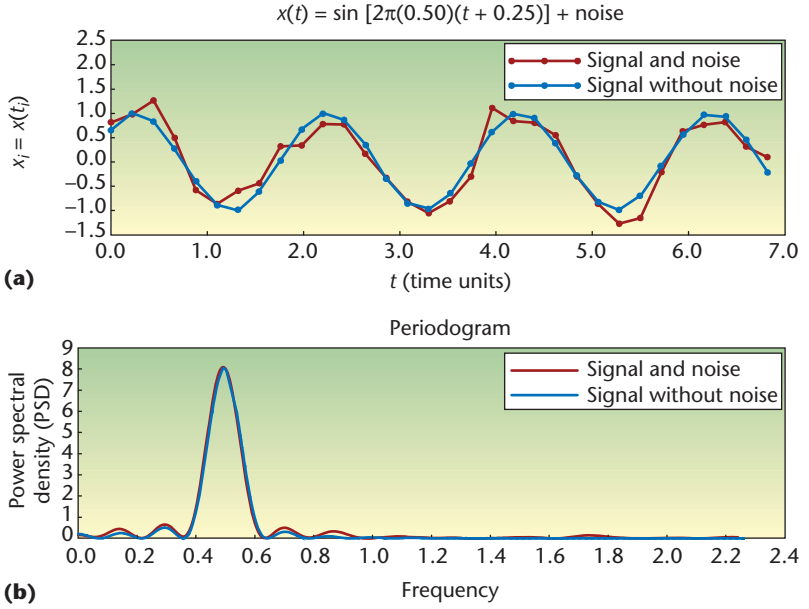

Figure 1. Original and new time series as defined by Equation 8. (a) The noise-corrupted time series and the uncorrupted series originally used in Part I's Figure 1b. The noise is independently, identically distributed *n*(0, 0.25). (b) Periodograms of the two times series plotted in (a). For the noise-corrupted series, the peak is centered on frequency $\hat{f}_0 = 0.493$.

The Autocorrelation Function

After the periodogram, the next frequency spectrum estimators to emerge were Richard Blackman and John Tukey's *correlogram* estimators.[4] They're based on the *autocorrelation theorem* (sometimes called Wiener's theorem), which states that if $X(f)$ is the CFT of $x(t)$, then $|X(f)|^2$ is the CFT of the *autocorrelation function* (ACF) of $x(t)$. Norbert Wiener defined the latter function as[5]

$$\rho(\tau) = \lim_{\mathcal{T}\to\infty} \frac{1}{2\mathcal{T}} \int_{-\mathcal{T}}^{\mathcal{T}} x^*(t)x(t+\tau)dt\,, \quad -\infty < \tau < \infty, \quad (9)$$

in which the variable τ is called the *lag* (the time interval for the correlation of $x(t)$ with itself), and $x^*(t)$ is the complex conjugate of $x(t)$. Thus, if we could access $x(t)$, we could compute the PSD in two ways: either by Equation 3 or by

$$P(f) = \int_{-\infty}^{\infty} \rho(\tau)\exp(-2\pi i f\tau)d\tau\,. \quad (10)$$

But again, we have access to only a noisy time series $x_0, x_1, \ldots, x_{N-1}$, so to use the second method, we need estimates for $\rho(\tau)$ evaluated at the discrete lag values

$$\tau_m = m\Delta t, \; m = 0, 1, \ldots, N-1. \quad (11)$$

Because we're working with a real time series, and $\rho(\tau_{-m}) = \rho(\tau_m)$, we don't need to worry about evaluating $\rho(\tau)$ at negative lags.

Because $\rho(\tau)$ is a limit of the average value of $x^*(t)x(t + \tau)$ on the interval $[-\mathcal{T}, \mathcal{T}]$, the obvious estimator is the sequence of average values

$$\hat{\rho}_m = \hat{\rho}(m\Delta t) = \frac{1}{N-m}\sum_{n=0}^{N-m-1} x_n x_{n+m}\,, \quad m = 0, 1, \ldots, N-1. \quad (12)$$

This sequence is sometimes called the *unbiased estimator of* $\rho(\tau)$ because its expected value is the true value—that is, $\varepsilon\{\hat{\rho}(m\Delta t)\} = \rho(m\Delta t)$. But the data are noisy, and for successively larger values of m, the average $\hat{\rho}_m$ is based on fewer and fewer terms, so the variance grows and, for large m, the estimator becomes unstable. Therefore, it's common practice to use the biased estimator

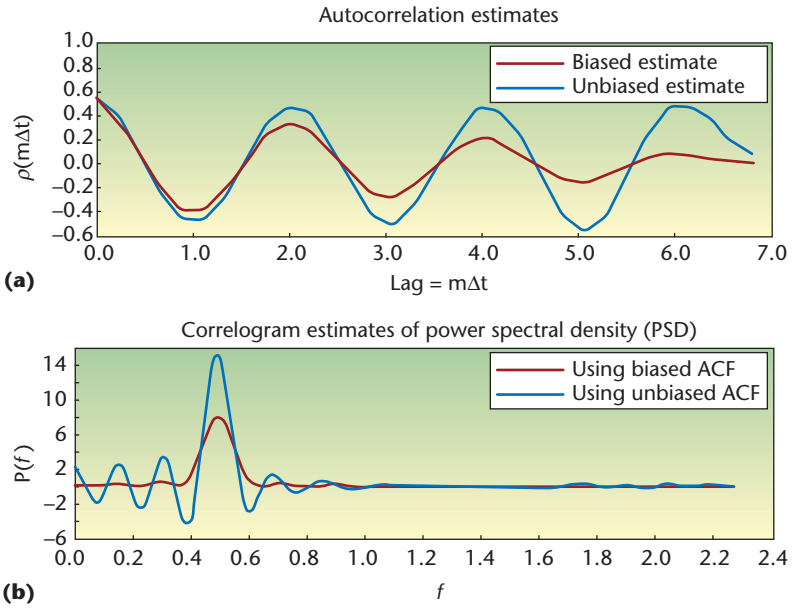

Figure 2. Autocorrelation and correlogram estimates for the noisy time series defined by Equation 8. (a) Biased and unbiased estimates of the autocorrelation function (ACF); (b) correlogram estimates obtained from the ACF estimates in (a).

$$\hat{\rho}_m = \hat{\rho}(m\Delta t) = \frac{1}{N}\sum_{n=0}^{N-m-1} x_n x_{n+m}, \quad m = 0, 1, \ldots, N-1, \tag{13}$$

which damps those instabilities and has a smaller total error (bias + variance) than does the unbiased estimator. (*Bias* is the difference between the estimator's expected value and the true value of the quantity being estimated.) Figure 2a gives plots of both estimates for the times series that Equation 8 defines.

The ACF we have just described is sometimes called the *engineering autocorrelation* to distinguish it from the *statistical autocorrelation*, which is defined by

$$\hat{r}_m = \frac{\frac{1}{N}\sum_{n=0}^{N-m-1}(x_n - \bar{x})(x_{n+m} - \bar{x})}{\frac{1}{N}\sum_{n=0}^{N-1}(x_n - \bar{x})^2},$$

where $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x_n$. (14)

The individual $\hat{r}_m$ are true correlation coefficients because they satisfy

$$-1 \le \hat{r}_m \le 1, \; m = 0, 1, \ldots, N-1. \tag{15}$$

Correlogram PSD Estimators

Once we've established the ACF estimate, we can use the FFT to calculate the discrete estimate to the PSD. More precisely, the ACF estimate is zero padded to have M lags, which gives $M/2 + 1$ frequencies in the PSD estimate, which we can then compute by approximating Equation 10 with

$$\hat{P}_k = \hat{P}(f_k) = \sum_{j=0}^{M-1} \hat{\rho}_j \exp\left(-2\pi i \frac{j}{M} k\right), \quad k = 0, 1, \ldots, M/2. \tag{16}$$

Zero padding in this case is an explicit expression of the implicit assumption that the ACF is zero for all lag values $\tau > (N-1)\Delta t$. We must assume that because we don't know the data outside the observation window. Assuming some nonzero extension for the ACF would amount to an implicit assumption about the missing observed data.

Figure 2b plots the correlograms corresponding to the biased and unbiased ACF estimates, shown in Figure 2a. The negative sidelobes for the unbiased correlogram show dramatically why most analysts choose the biased estimate even though its central peak is broader. The reason for this broadening, and for the damped sidelobes, is that the biased ACF, Equation 13, can also be computed by multiplying the unbiased ACF, Equation 12, by the triangular (Bartlett) tapering window

$$w_k = 1 - \frac{k}{N}, \quad k = 0, 1, 2, \ldots, N-1. \tag{17}$$

Recall that we observed the same sort of peak broadening and sidelobe suppression in Part I's Figure 10 when we multiplied the observed data by a Blackman window before computing the periodogram.

Notice that the biased correlogram estimate plotted in Figure 2b is identical to the periodogram estimate plotted in Figure 1b. The equality of these two estimates, computed in very different ways, constitutes a finite dimensional analogue of Wiener's theorem for the continuous PSD.

Figure 2b's two PSD correlograms aren't the only members of the class of correlogram estimates. We can obtain other variations by truncating the ACF estimate at lags $\tau < (N-1)\Delta t$ and by smoothing the truncated (or untruncated) estimate with one of the tapering windows defined in Part I's Equation 11. Most of those windows were originally developed for the correlogram method; they were then retroactively applied to the periodogram method when the latter was resurrected in the mid 1960s. In those days, people often used very severe truncations, with the estimates being

set to zero at 90 percent or more of the lags. Not only did this alleviate the variance instability problem, but it also reduced the computing time—an important consideration before the invention of the FFT algorithm, and when computers were much slower than today.

The effect of truncating the biased ACF estimate is shown in Figure 3, where $m_{\max}$ is the largest index for which the nonzero ACF estimate is retained. More precisely,

$$\hat{\rho}_m = \frac{1}{N-m}\sum_{n=0}^{N-m-1} x_n x_{n+m},$$
$$m = 0,1,\ldots,m_{\max},$$
$$\hat{\rho}_m = 0,\ m = m_{\max}+1,\ldots,N-1. \quad (18)$$

It's clear that smaller values of $m_{\max}$ produce more pronounced sidelobes and broader central peaks than larger values. The peak broadening is accompanied by a compensating decrease in height to keep the area under the curve invariant. PSD is measured in units of power-per-unit-frequency interval, so the peak's area indicates its associated power.

Figure 4 shows the effect of tapering the truncated ACF estimates used in Figure 3 with a Hamming window

$$w_m = 0.538 + 0.462\cos\left(\frac{m\pi}{m_{\max}}\right),$$
$$m = 0, 1, 2, \ldots, m_{\max}. \quad (19)$$

The sidelobes are suppressed by the tapering, but the central peaks are further broadened. This loss in resolution is the price we must pay to smooth the sidelobes and eliminate their negative excursions.

Tapering the biased ACF estimates with the Hamming window amounts to twice tapering the unbiased estimates; we can obtain the former from the latter by tapering them with the Bartlett window, Equation 17. Figure 5 shows the effect of a single tapering of the unbaised estimates with the Hamming window, Equation 19. Note that the sidelobes are not completely suppressed, but they're not as

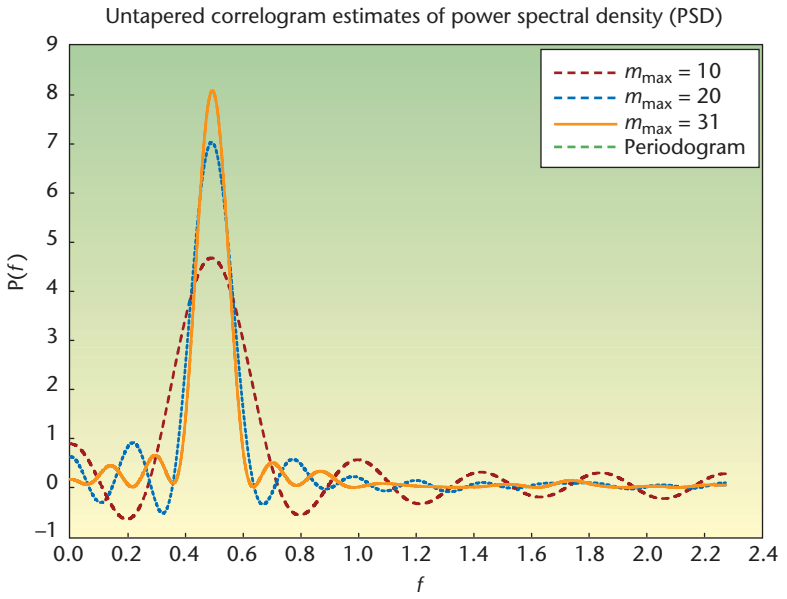

Figure 3. Three correlogram estimates for Equation 8 computed from the biased autocorrelation function (ACF) estimator in Equation 13. The periodogram, although plotted, doesn't show up as a separate curve because it's identical to the m_{max} = 31 correlogram.

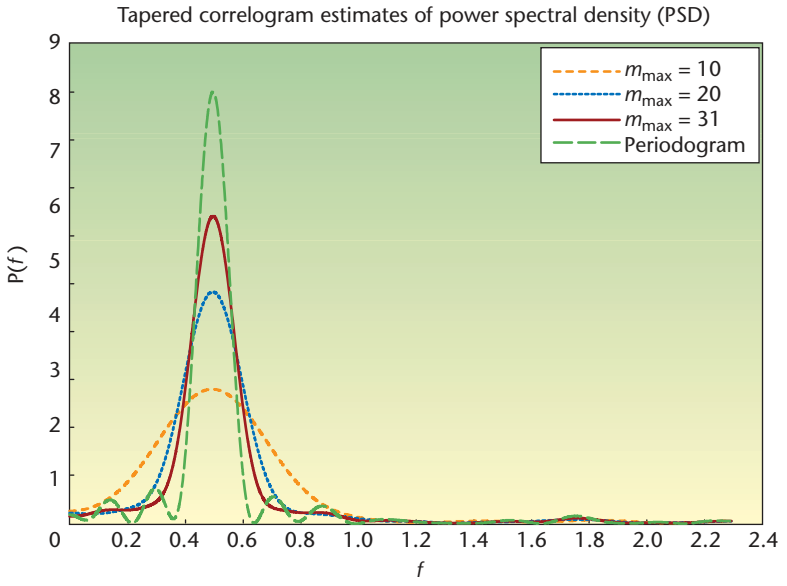

Figure 4. Three correlogram estimates for the time series generated by Equation 8. We computed the estimates by tapering three truncations of the biased estimator in Equation 13 with a Hamming window. The periodogram was also plotted for comparison. Although it has sidelobes, its central peak is sharper than those of the correlograms.

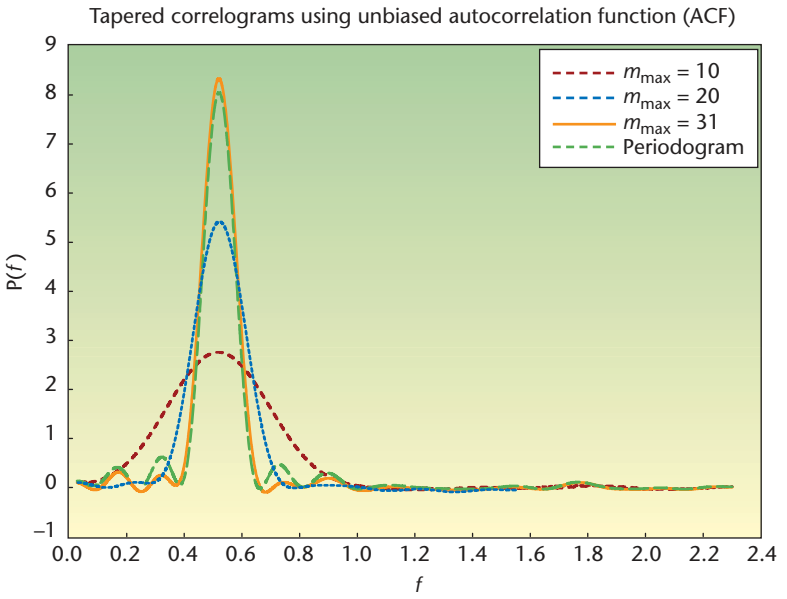

Figure 5. Three correlogram estimates for the time series generated by Equation 8. We computed the estimates by tapering three truncations of the unbiased estimator in Equation 12. We also plotted the periodogram for comparison; again, it has a sharper peak but larger sidelobes .

pronounced as in Figure 3, in which the tapering used the Bartlett window. However, the central peaks are also slightly broader here. This is yet another example of the trade-off between resolution and sidelobe suppression.

This particular example contains only a single-sinusoid, so it doesn't suggest any advantage for the tapering and truncation procedures, but they weren't developed to analyze a time series with such a simple structure. Their advantages are said to be best realized when the signal being analyzed contains two or more sinusoids with frequencies so closely spaced that sidelobes from two adjacent peaks might combine and reinforce one another to give a spurious peak in the spectrum. But of course, if two adjacent frequencies are close enough, then the broadening of both peaks might cause them to merge into an unresolved lump.

Much ink has been used in debating the relative merits of the various truncation and windowing strategies, but none of them have proven to be advantageous, so correlogram estimates are beginning to fall out of favor. For the past 30 years or so, most researchers have concentrated on autoregressive spectral estimates, which, as we shall see in Part 4, give better resolution because they make better assumptions about the the data outside the window of observation. CiSE

References

1. D. Donnelly and B. Rust, "The Fast Fourier Transform for Experimentalists, Part I: Concepts," *Computing in Science & Eng.*, vol. 7, no. 2, 2005, pp. 80–88.
2. D. Donelly and B. Rust, "The Fast Fourier Transform for Experimentalists, Part II: Convolutions," *Computing in Science & Eng.*, vol. 7, no. 3, 2005, pp. 92–95.
3. A. Schuster, "On the Investigation of Hidden Periodicities with Application to a Supposed Twenty-Six-Day Period of Meteorological Phenomena," *Terrestrial Magnetism*, vol. 3, no. 1, 1898, pp. 13–41.
4. R.B. Blackman and J.W. Tukey, *The Measurement of Power Spectra*, Dover Publications, 1959.
5. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT Press, 1949.

Denis Donnelly is a professor of physics at Siena College. His research interests include computer modeling and electronics. Donnelly received a PhD in physics from the University of Michigan. He is a member of the American Physical Society, the American Association of Physics Teachers, and the American Association for the Advancement of Science. Contact him at donnelly@siena.edu.

Bert Rust is a mathematician at the US National Institute for Standards and Technology. His research interests include ill-posed problems, time-series modeling, nonlinear regression, and observational cosmology. Rust received a PhD in astronomy from the University of Illinois. He is a member of SIAM and the American Astronomical Society. Contact him at bwr@nist.gov.