Editors: David Winch, winch@TaosNet.com
Denis Donnelly, donnelly@siena.edu

The Fast Fourier Transform for Experimentalists, Part IV: Autoregressive Spectral Analysis

By Bert Rust and Denis Donnelly

IT'S RARE THAT WE HAVE ONLY ONE WAY IN WHICH TO APPROACH A PARTICULAR TOPIC—FORTUNATELY, SPECTRUM ESTIMATION ISN'T ONE OF THOSE RARE CASES. IN THE MOST recent article of this series,[1] we considered the periodogram and correlogram estimators for the power spectral density (PSD) function. However, they are only two of several possibilities.

In this installment, we consider two additional kinds of spectrum estimates: autoregressive (AR) estimates and the maximum entropy (ME) method. In the first approach, we assume that an AR process generates the time series, which means we can compute the PSD of the time series from estimates of the AR parameters. The second approach is a special case of the first, but it uses a different method for estimating the AR parameters. Specifically, it chooses them to make the PSD's inverse transform compatible with the measured time series, while remaining maximally noncommittal about the data outside the observational window.

Autoregressive Time-Series Models

Both the periodogram and correlogram estimates make rather unrealistic assumptions about the data outside the observational window. Moreover, when they use tapering windows or truncation of the autocorrelation function (ACF), they change the observed data. The years since the early 1970s have seen the development of a new class of PSD estimators that are based on the idea of fitting a parametric time-series model to the observed data. This enables us to use estimates of the parameters in the theoretical expression of the model's PSD to get an estimate of the observed series' PSD. If the model is a good representation of the process that generated the data, then hopefully it will give a more realistic extrapolation for the missing data.

The class of models used most often assumes that the data are generated by an AR process in which each new data point is formed from a linear combination of the preceding data plus a random shock. The basic idea is that a system's future states depend in a deterministic way on previous states, but at each time step, a random perturbation drives the system forward. We can write the AR models of orders 1, 2, and 3 as

$$\begin{aligned} &AR(1): x_n = -a_1 x_{n-1} + u_n, && n = 1, 2, \ldots, N-1, \\ &AR(2): x_n = -a_1 x_{n-1} - a_2 x_{n-2} + u_n, && n = 2, 3, \ldots, N-1, \\ &AR(3): x_n = -a_1 x_{n-1} - a_2 x_{n-2} - \\ &\qquad a_3 x_{n-3} + u_n, , && n = 3, 4, \ldots, N-1 \end{aligned} \quad (1)$$

where a_1, a_2, and a_3 are the AR parameters (whose values must be determined to make the model fit the data), and u_n is the random shock at time step n. We assume the random shocks to be samples from a zero-mean distribution whose variance remains constant in time. The choice of negative signs for the parameters is a universal convention adopted for notational convenience in derivations that we won't give here.

Autoregressive Spectral Estimates

In general, for any integer $p < N-1$, the $AR(p)$ model is

$$x_n = -\sum_{k=1}^{p} a_k x_{n-k} + u_n, \; n = p, \; p+1, \ldots, N-1. \quad (2)$$

We can show that the PSD function for this model is

$$P_{AR}(f) = \frac{\rho_w}{\left|1 + \sum_{j=1}^{p} a_j \exp(-2\pi i f \, j\Delta t)\right|^2}, -\frac{1}{2\Delta t} \le f \le \frac{1}{2\Delta t}, \quad (3)$$

where ρ_w is another adjustable parameter that we can estimate along with $a_1, a_2, \ldots, a_p$ by solving the $(p+1) \times (p+1)$ linear system of equations,

$$\begin{bmatrix} \rho_0 & \rho_{-1} & \rho_{-2} & \cdots & \rho_{-p} \\ \rho_1 & \rho_0 & \rho_{-1} & \cdots & \rho_{-p+1} \\ \rho_2 & \rho_1 & \rho_0 & \cdots & \rho_{-p+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_p & \rho_{p-1} & \rho_{p-2} & \cdots & \rho_0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \cdots \\ a_p \end{bmatrix} = \begin{bmatrix} \rho_w \\ 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \quad (4)$$

Copublished by the IEEE CS and the AIP 1521-9615/05/$20.00 © 2005 IEEE

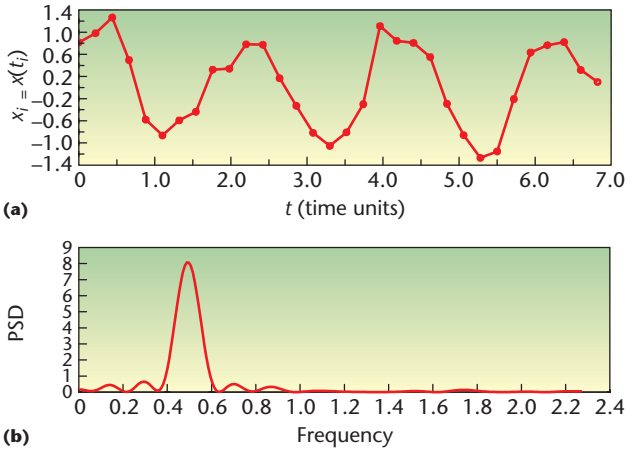

Figure 1. The time series generated by Equation 10 and its periodogram. The discrete points in the upper plot are joined by straight-line segments to emphasize the time series nature of the data. The time series was zero padded to length M = 1,024 to compute the periodogram in the lower plot.

which are sometimes called the *Yule-Walker equations*. The ρ-values in the matrix are just the autocorrelations $\rho_k = \rho(\tau_k) = \rho(k\Delta t)$ that we defined in the last issue[1] with

$$\rho(\tau) = \lim_{\mathcal{T}\to\infty} \frac{1}{2\mathcal{T}} \int_{-\mathcal{T}}^{\mathcal{T}} x^*(t)x(t+\tau)dt, -\infty < \tau < \infty \qquad (5)$$

where x^* is the complex conjugate of $x(t)$. We're working with real data, so $\rho_{-k} = \rho_k$, which means that the matrix is symmetric and positive definite. Note that the element in row i and column j is just $\rho_{(i-j)}$, which makes it a *Toeplitz matrix*. Norman Levinson[2] exploited this special structure to devise a recursive algorithm that solves the system in times proportional to $(p + 1)^2$ rather than the $(p + 1)^3$ required by a general linear equations solver.

We can summarize the steps required to compute an autoregressive spectral estimate as follows:

1. Choose an autoregressive order $p \le N - 1$.
2. Compute ACF estimates $\hat{\rho}_0, \hat{\rho}_1, \ldots, \hat{\rho}_p$ using the biased estimator

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

$$m = 0, 1, \ldots, N-1. \qquad (6)$$

3. Substitute $\hat{\rho}_0, \hat{\rho}_1, \ldots, \hat{\rho}_p$ into the matrix in Equation 4 and use the Levinson algorithm to compute estimates $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p$ and $\hat{\rho}_w$.
4. Substitute $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p$ and $\hat{\rho}_w$ into Equation 3 to compute the PSD estimate $\hat{P}_{AR}(f)$ on any desired frequency mesh.

It's absolutely necessary to use the biased ACF estimator in step 2. Using the unbiased estimator produces an unstable linear system (see Equation 4) with a matrix that numerically isn't positive definite.

It's easy to do the calculations in the final step by using the fast Fourier transform (FFT) algorithm to compute the denominator in Equation 3. If we define $\hat{a}_0 \equiv 1$, then

$$1 + \sum_{j=1}^{p} \hat{a}_j \exp(-2\pi \mathbf{i} f\, j\Delta t) = \sum_{j=0}^{p} \hat{a}_j \exp(-2\pi \mathbf{i} f\, j\Delta t). \qquad (7)$$

Suppose we want to evaluate $P_{AR}(f)$ at $(M/2 + 1)$ equally spaced frequencies

$$f_k = \frac{k}{M\Delta t}, \; k = 0, 1, \ldots, M/2, \qquad (8)$$

where $M > p$. Then,

$$\sum_{j=0}^{p} \hat{a}_j \exp(-2\pi \mathbf{i} f_k\, j\Delta t) = \sum_{j=0}^{p} \hat{a}_j \exp\left(-2\pi \mathbf{i} \frac{j}{M} k\right), \qquad (9)$$

and we can compute these values quite quickly by zero padding the sequence $\hat{a}_0, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p$ to have M terms and applying the FFT algorithm.

Two Examples

If we choose the AR order p properly, the peaks in the $AR(p)$ spectrum will be sharper than those in the periodogram or correlogram estimates. There is no clear-cut prescription for choosing p, but a fairly wide range of values will usually give acceptable results. To illustrate the effect of the choice of p, let's revisit an example time series used in the last issue.[1] Again, we'll take $N = 32$, $\Delta t = 0.22$, and consider the time series generated by

$$t_j = j\Delta t, \quad j = 0, 1, 2, \ldots, N-1,$$
$$x_j = x(t_j) = \sin[2\pi f_0(t_j + 0.25)] + \varepsilon_j, \qquad (10)$$

with $f_0 = 0.5$, and each ε_j a random number drawn independently from a normal distribution with mean zero and standard deviation $\sigma = 0.25$. Figure 1 plots the time series and its periodogram, and Figure 2 gives three different $AR(p)$ spectra for the time series, together with the periodogram for comparison. Table 1 gives the locations of the peak centers. Both the $AR(16)$ and $AR(24)$ estimates give better results than the periodogram, but for real-world problems, it's best to try several orders in the range $N/2 \le p \le 3N/4$ and compare them to make the final choice. Our own experience has indicated that the best choice usually has $p \approx 2N/3$.

To better illustrate the AR methods' power, let's reconsider another time series originally introduced in Part I of our series (specifically, Figure 2a).[3] We generated it by summing two sine waves, with amplitudes $A_1 = A_2 = 1.0$, frequencies $f_1 = 1.0$ and $f_2 = 1.3$, and phases $\phi_1 = \phi_2 = 0$, at $N = 16$ equally spaced time points with $\Delta t = 0.125$. Again, we add random noise to make the problem more realistic, and write

$$t_j = j\Delta t, \quad j = 0, 1, \ldots, N-1,$$
$$x_j = \sin[2\pi f_1 t_j] + \sin[2\pi f_2 t_j] + \varepsilon_j, \tag{11}$$

with the ε_j chosen independently from a normal distribution; the mean is 0 and standard deviation $\sigma = 0.25$. This is the same error distribution used in the preceding example, but the samples used here differ from any used there. The top graph of Figure 3 gives plots of the noisy and noise-free time series, and the bottom graph gives their periodograms. Figure 4 gives plots of the PSD's periodogram and *AR*(12) estimates. The latter clearly indicates the presence of two peaks, although it doesn't completely resolve them. The two maxima occur at frequencies very near the true values used to generate the time series. It's remarkable that the *AR*(12) estimate could obtain such good agreement with the true values using only 16 noise-corrupted data points.

The Maximum Entropy Approach

John Parker Burg invented the ME method in the late 1960s; he exhibited its strengths and advantages in oral presentations at geophysics conferences, but he didn't publish the mathematical derivations that defined and justified it until his PhD thesis[4] appeared in 1975. This lack of published documentation produced a great deal of independent work by other researchers who were trying to understand and extend the method. In fact, the ME method was one of the chief motivators for the development of the AR methods and can be classified as an AR method, although Burg didn't use AR models in its development.

Rather, Burg started with the definition for PSD, that is,

$$P(f) = \int_{-\infty}^{\infty} \rho(\tau) \exp(-2\pi i f \tau) d\tau, \tag{12}$$

but sought a function $P_e(f)$, defined on the Nyquist band $-1/(2\Delta t) \le f \le 1/(2\Delta t)$, which satisfied three guiding principles:

1. The inverse Fourier transform of $P_e(f)$ should return the autocorrelation function unchanged by any filtering or tapering operations:

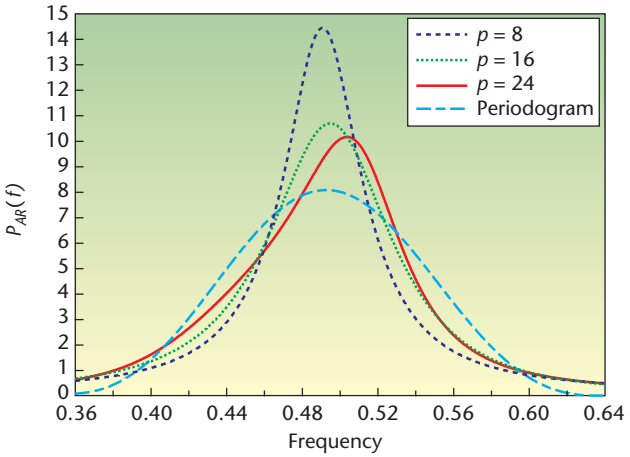

Figure 2. *AR*(*p*) power spectral density (PSD) estimates. For *p* = 8, 16, and 24, and the periodogram for the time series generated by Equation 10, the plot doesn't cover the whole Nyquist band 0 ≤ f ≤ 2.273, but rather only the frequency range spanned by the central peak in the periodogram. Using the whole Nyquist range renders the *AR*(*p*) peaks so narrow that it's difficult to distinguish between them.

Table 1. Peak centers.

Estimate	Periodogram	*AR*(8)	*AR*(16)	*AR*(24)
$\hat{f}_0$	0.493	0.491	0.495	0.504

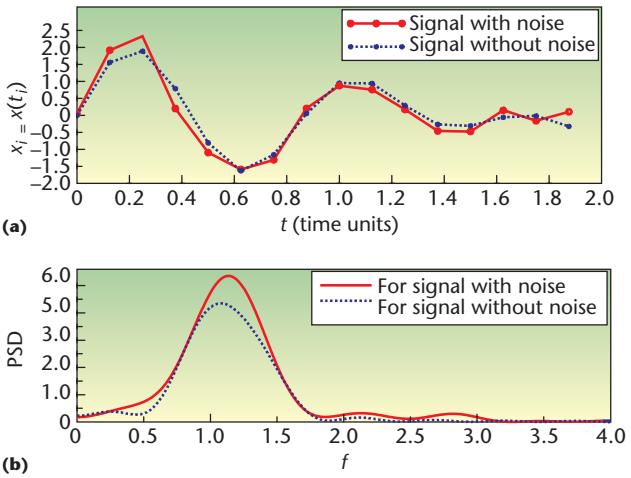

Figure 3. Time series. (a) The noise-corrupted time series generated by Equation 11; the noise is independently and identically distributed *n*(0, 0.25). (b) Periodograms of the two time series plotted in part (a). In neither case was the periodogram method able to resolve two separate peaks. For the noisy spectrum, the unresolved lump peaks at frequency $\hat{f}$ = 1.136.

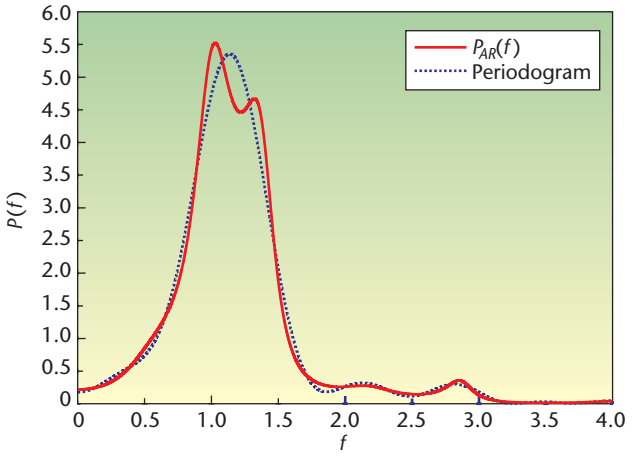

Figure 4. Power spectral density (PSD). The *AR*(12) and the untapered periodogram estimates of the PSD for time series generated by Equation 11. The two maxima in the *AR*(12) spectrum occur at frequencies $\hat{f}_1 = 1.027$ and $\hat{f}_2 = 1.321$, which are very near the true values $f_1 = 1.00$ and $f_2 = 1.30$.

$$\rho_m = \rho(m\Delta t) = \int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} P_e(f)\exp(2\pi \mathbf{i} f m\Delta t)df,$$

$$m = 0, 1, \ldots, N-1. \quad (13)$$

2. $P_e(f)$ should correspond to the most random or unpredictable time series whose autocorrelation function agrees with the known values.
3. $P_e(f) > 0$ on the interval $-1/(2\Delta t) \le f \le 1/(2\Delta t)$.

The first condition merely states that the measured data shouldn't be changed in any way in computing $P_e(f)$. The second is a statement about what is to be assumed about the data outside the observational window. Essentially, it says that those assumptions should be minimized.

To measure a time series' randomness or unpredictability, Burg used the information theoretic concept of *entropy*. A random process

$$\ldots x(-2\Delta t), x(-\Delta t), x(0), x(\Delta t), x(2\Delta t), \ldots \quad (14)$$

is said to be *band limited* if its PSD function is zero everywhere outside its Nyquist band. If $P(f)$ is such a PSD function, then the time series' entropy rate (entropy per sample) is given by

$$h\{(P(f)\} = \int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} \ln[P(f)]df. \quad (15)$$

Burg's idea was to maximize this quantity, subject to the constraints imposed by Equation 13. More precisely, he sought to impose the constraint at lags 0, Δt, $2\Delta t$, …, $p\Delta t$ with $p < N$ and then choose from the set of all nonnegative functions $P(f)$ that satisfy those $p + 1$ constraints the particular one that minimizes the entropy rate (Equation 15). We can write the problem formally as

$$h\{P_e(f)\} = \max_{P(f)}\left\{\int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} \ln[P(f)]df \;\middle|\; \begin{array}{c} P(f) > 0, \\ \int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} P(f)\exp(2\pi \mathbf{i} f m\Delta t)df = \rho_m, \\ m = 0,1,\ldots,p \end{array}\right\}. \quad (16)$$

We need techniques from the calculus of variations to solve it; we can show that

$$P_e(f) = \frac{\rho_e}{\left|1 + \sum_{j=1}^{p} a_j \exp(-2\pi \mathbf{i} f j\Delta t)\right|^2}, \; -\frac{1}{2\Delta t} \le f \le \frac{1}{2\Delta t}, \quad (17)$$

where $a_1, a_2, \ldots, a_p$ and ρ_e are parameters satisfying

$$\begin{bmatrix} \rho_0 & \rho_1 & \rho_2 & \cdots & \rho_p \\ \rho_1 & \rho_0 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_2 & \rho_1 & \rho_0 & \cdots & \rho_{p-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_p & \rho_{p-1} & \rho_{p-2} & \cdots & \rho_0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \cdots \\ a_p \end{bmatrix} = \begin{bmatrix} \rho_e \\ 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}. \quad (18)$$

Equation 17 is the same as Equation 3, and, because we're working with real data for which $\rho_{-k} = \rho_k$, Equation 18 is the same as Equation 4. Thus the maximum entropy method is correctly classified as an AR method, even though Burg used different methods to estimate the autocorrelations and parameters in Equation 18.

Forward and Backward Prediction Filters

Burg regarded the vector $(1 \;\; a_1 \;\; a_2 \;\; \ldots \;\; a_p)^T$ as a prediction filter, which he applied to the data $x_0, x_1, \ldots, x_{N-1}$ in both the forward and reverse directions to get forward and backward predictions $\hat{x}_n^f$, $\hat{x}_n^b$ and their corresponding prediction errors e_n^f, e_n^b:

$$\hat{x}_n^f = -\sum_{k=1}^{p} a_k x_{n-k}, \quad e_n^f = x_n - \hat{x}_n^f, \quad n = p, p+1, \ldots, N-1,$$

$$\hat{x}_n^b = -\sum_{k=1}^{p} a_k x_{n+k}, \quad e_n^b = x_n - \hat{x}_n^b, \quad n = 0, 1, \ldots, N-p-1. \quad (19)$$

He reasoned that he could get the best estimates for $a_1, a_2, \ldots, a_p$ by minimizing the sum of squares of the predictions errors, for example,

$$\sum_{n=p}^{N-1} \left|e_n^f\right|^2 + \sum_{n=0}^{N-p-1} \left|e_n^b\right|^2 . \qquad (20)$$

He was able to devise a recursive algorithm that gave estimates not only for $a_1, a_2, \ldots, a_p$, but also, at the same time, for ρ_e and for the autocorrelations $\rho_0, \rho_1, \ldots, \rho_p$. The details are complicated, so we won't give them here.[4] It's remarkable that the recursion generates a new estimator for the elements of the matrix in Equation 18 at the same time it's solving the system of equations!

Choosing the Order *p*

Like the other AR methods, the ME method requires the choice of an order $p < N$. Figure 5 exhibits the results of choosing a low, an intermediate, and a high order for the time series generated by Equation 10. The same plots are repeated using a logarithmic scaling in Figure 6. Table 2 gives the peak locations. The *ME*(3) spectrum gave the best estimate $\hat{f}_0$, but its peak is almost as broad as the periodogram. Increasing p produces sharper peaks, but the locations display a noticeable downward bias. The *ME*(14) estimate is fairly representative of the orders in the range $4 \le p \le 25$. At $p = 26$ the peak splits into two, with the dominant one giving a better $\hat{f}_0$ than any of the sharp single peaks for $p = 4, 5, \ldots, 25$. The same splitting occurs for orders $p = 27, 28, 29$, and 30, with the dominant peak becoming sharper and sharper but remaining at $\hat{f}_0 = 0.492$. These spurious splittings aren't caused by errors in the data. In fact, they occur much more readily for artificially generated time series without added noise, but the *ME*(26) spectrum clearly demonstrates that they also occur in noisy data, so great care must be exercised in interpreting high-order ME spectra. One of the ME method's strengths is its ability to resolve closely spaced peaks, but in using it for that purpose, always remember the possibility of a spurious splitting of a single peak.

Researchers have proposed several criteria for choosing the optimal order for the ME method (and for the other AR methods), but none of them work all of the time. In fact, it's easier to find a time series that confounds a given criterion than it is to develop it. Many authors[5,6] recommend $p \le N/2$, but higher order methods often give better results.Figure 7 shows the result of using a relatively high p for the time series generated by Equation 11. The very narrow spurious peak at $\hat{f} = 2.901$ is a typical occurrence when we use high values for p. Such peaks can usually be easily identified because they're so much sharper than the peaks correspond-

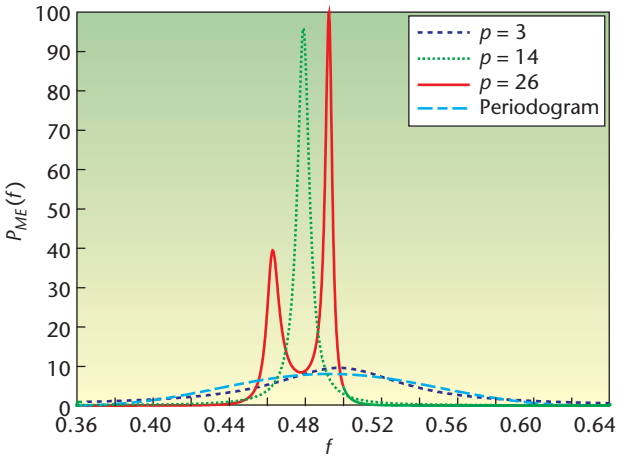

Figure 5. Maximum entropy power spectral density (PSD) estimates. For orders $p = 3$, 14, and 26, and the periodogram for the time series generated by Equation 10, we see plots along the same frequency range used for the *AR*(*p*) spectra in Figure 2. The ME peaks are even sharper than the *AR*(*p*) peaks, so they must be taller to preserve the area subtended.

Table 2. Peak locations.

Estimate	Periodogram	*ME*(3)	*ME*(14)	*ME*(26)
$\hat{f}_0$	0.493	0.498	0.479	0.492

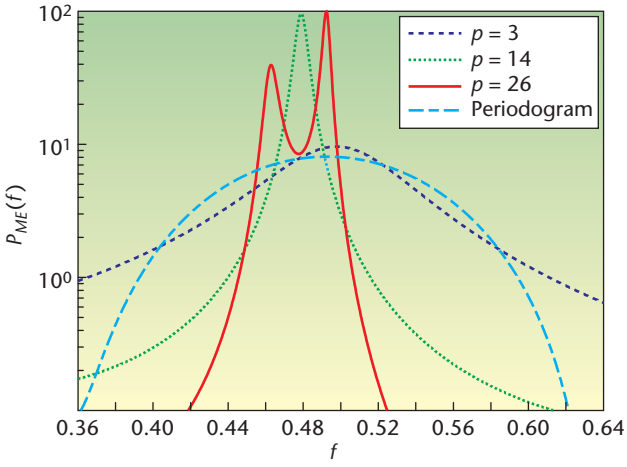

Figure 6. Another view of the plots given in Figure 5. Using the logarithm scale makes it easier to compare the ME(3) estimate with the periodogram.

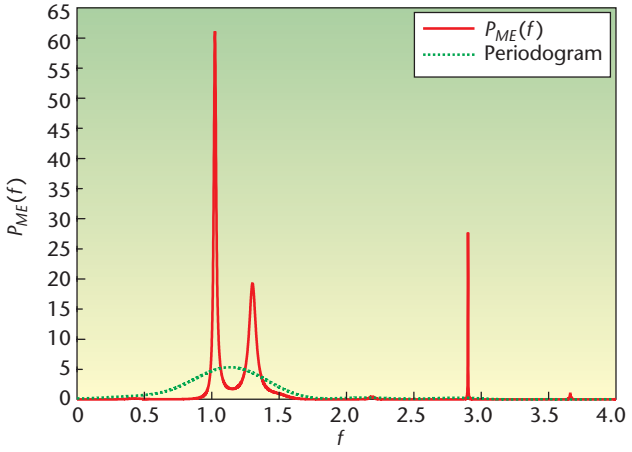

Figure 7. Maximum entropy (ME) method. In the *ME*(14) PSD estimate for the time series generated by Equation 11, the two peaks are centered at $\hat{f}_1 = 1.023$ and $\hat{f}_2 = 1.302$. These are somewhat better than the estimates from the *AR*(12) spectrum in Figure 4. The very narrow peak at $\hat{f} = 2.901$ is an artifact caused by using the very high order $p = 14$ (high relative to $N = 16$), but because it's so narrow, it doesn't indicate much power and thus can be safely ignored.

ing to real power. The one in Figure 7 is a small price to pay for the excellent resolution of the two real peaks. It's amazing that the ME method can achieve such good results using just 16 noisy data points spanning only ≈ 2.5 cycles of the higher frequency sine wave.

We've now looked at four different methods of spectrum estimation, and although we haven't exhausted the subject, we must proceed. (More details about this topic appear elsewhere.[5,6]) In the next installment, we'll take a brief look at filters and detrending before we present an analysis of a bat chirp. In the final installment, we'll discuss some statistical tests and use them to analyze atmospheric pressure differences in the Pacific Ocean that have significant environmental implications. CiSE

References

1. B. Rust and D. Donnelly, "The Fast Fourier Transform for Experimentalists, Part III: Classical Spectral Analysis," *Computing in Science & Eng.*, vol. 7, no. 5, 2005, pp. 74–78.
2. N. Levinson, "The Wiener (Root Mean Square) Error Criterion in Filter Design and Prediction," *J. Mathematical Physics*, vol. 25, 1947, pp. 261–278.
3. 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.
4. J.P. Burg, *Maximum Entropy Spectral Analysis*, PhD dissertation, Dept. of Geophysics, Stanford Univ., May 1975; http://sepwww.stanford.edu/theses/sep06/.
5. S.L. Marple Jr., *Digital Spectral Analysis with Applications*, Prentice Hall, 1987.
6. S.M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice Hall, 1988.

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 has 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.

Denis Donnelly is a professor of physics at Siena College. His research interests include computer modeling and electronics. Donnelly has 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.