Maximum entropy spectral analysis

James G. Berryman

ABSTRACT

A review of the maximum entropy spectral analysis (MESA) method for time series is presented. Then, empirical evidence based on maximum entropy spectra of real seismic data is shown to suggest that $M = 2N/\ln 2N$ is a reasonable *a priori* choice of the operator length M for discrete time series of length N. Various data examples support this conclusion.

INTRODUCTION

When analyzing seismic traces, it is often useful to know what frequencies are present in the data. Filtering and smoothing of data should be done with knowledge of the frequency content. In the standard approach to spectral analysis, the Fourier transform of the trace (amplitude spectrum) is computed. This approach is quite reliable for long data sequences (1000 or more data points) and is satisfactory for somewhat shorter sequences. Unfortunately, this technique becomes unreliable for very short time samples due to the increased importance of end effects: (a) the resolution of true peaks in the spectrum becomes poor and (b) spurious peaks may be introduced because of the implicit (and incorrect) assumption often made that the known data sequence is repeated periodically in time.

A different approach to spectral anaysis was introduced into the geophysical literature by **?**. His idea was to obtain an estimate of the power spectrum (square of the amplitude spectrum) by maximizing the spectral entropy with the known autocorrelation values as constraints. In principle, this approach should give a power spectrum that is consistent with the available information, but maximally noncommittal with regard to the unavailable information. It turns out that the resulting mathematical problem can be solved exactly using linear matrix theory. In fact, the method requires computation of the minimum phase deconvolution operator [also known as the "prediction error filter" (**?**)], which has received much attention in the geophysical literature. The power spectrum is then given by the square inverse of the operator's Fourier transform. Burg's method is known as maximum entropy spectral analysis (MESA) and is closely related both to deconvolution and to autoregressive analysis of stationary random time series.

The method of computing the spectrum in MESA can be easily understood in terms of filter theory. If we apply a prediction error filter to an input time series, the output will be a white spectrum. It is well-known that the spectrum of the output is the spectrum of the input times the spectrum of the filter. Since a white spectrum is constant, an estimate of the input spectrum is given by the inverse of the spectrum of the prediction error filter.

MESA has one principal advantage over the standard Fourier transform method of spectral analysis: resolution of peaks in the power spectrum is enhanced for short data sequences. MESA has two principal disadvantages: (a) computation time is increased (substantially for long data sequences) and (b) the best choice of for the operator length is not known (poor choices can give misleading results for short data samples). A possible solution to this second problem is discussed in the section on **Choosing the Operator Length**.

At least two other approaches to spectral analysis are possible. (a) The maximum likelihood method or MLM **?** has been shown by **?** to be the inverse of the arithmetic average of inverse maximum entropy spectra of increasing operator length. Thus, MLM weights the strongest peaks of MESA the least and cannot give very good resolution. (b) Using the terms of stochastic theory (**?**), the ordinary power spectrum assumes that the underlying process is a moving average (MA) process. Using MESA can be viewed as being equivalent to assuming the process is autoregressive (AR). In fact, a discretely sampled geophysical time series is most likely to be a combination of the two, namely an autoregressive-moving-average (ARMA) process. It is possible to estimate the spectrum under the ARMA assumption; however, a substantial increase in computation time is required (over MESA), while the resolution of peaks should remain nearly the same.

A brief discussion of the theory and practice of MESA has appeared previously in **?**. An expanded version of this account is given in the following pages. The work presented here leads to the conclusion that for short time series MESA may well be a useful tool, and that MESA is probably the best available alternative to standard methods for such short data processing problems.

THE VARIATIONAL PRINCIPLE

Given a discrete (possibly complex) time series $\{X_1, \ldots, X_N\}$ of N values with sampling interval Δt (and Nyquist frequency $W = 1/2\Delta t$), we wish to compute an estimate of the power spectrum $P(f)$, where f is the frequency. It is well known that

$$P(f) = \lim_{N\to\infty} \frac{1}{N} \left| \sum_{n=1}^{N} X_n \exp\left(i2\pi f n \Delta t\right) \right|^2 = \sum_{n=-\infty}^{\infty} R_n \exp\left(i2\pi f n\Delta t\right), \tag{1}$$

where the autocorrelation function R is defined by (for $n \geq 0$)

$$R_n = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N-n} X_i^* X_{i+n} = R_{-n}^*. \tag{2}$$

Now suppose that we use the finite sequence $\{X_i\}$ to estimate the first M autocorrelation values $R_0, \ldots, R_{M-1}$. (Methods of obtaining these estimates are discussed in the section on **Computing the Prediction Error Filter**.) Then, **?** has shown that maximizing the average entropy (see Appendix A for a derivation)

$$h = \frac{1}{4W} \int_{-W}^{W} \ln \left[2WP(f)\right] df, \tag{3}$$

subject to the constraint that (1) is equivalent to extrapolating the autocorrelation R_n for $|n| \geq M$ in the most random possible manner.

Doing the math, we find that

$$\frac{\delta h}{\delta R_n} = \frac{1}{4W} \int_{-W}^{W} P^{-1}(f) \exp\left(i2\pi f n \Delta t\right) df = \left\{ \begin{array}{ccc} \lambda_n/2 & \text{for} & |n| < M \\ 0 & \text{for} & |n| \geq M. \end{array} \right. \tag{4}$$

The λ's are Lagrange multipliers to be determined. That the variation of h with respect to R_n for $|n| \geq M$ should be zero is the essence of the variational principle. The value of h is then stationary with respect to changes in the R_n's, which are unknown. We can infer from Equation (4) that

$$P^{-1}(f) = \sum_{n=-(M-1)}^{M-1} \lambda_n \exp\left(-i2\pi f n \Delta t\right). \tag{5}$$

Making the Z-transform to $Z = \exp\left(-i2\pi f \Delta t\right)$, Equation (5) becomes a polynomial of the complex parameter Z:

$$P^{-1}(f) = \sum \lambda_n Z^n. \tag{6}$$

Since P is necessarily real and nonnegative, Equation (6) can be uniquely factored as

$$P^{-1}(f) = 2WE_M^{-1} \left[\sum_m a_m Z^m\right] \left[\sum_n a_n^* Z^{-n}\right] = 2WE_M^{-1} \left|\sum_n a_n Z^n\right|^2, \tag{7}$$

with $a_0 = 1$. The first sum in (7) has all of its zeroes outside the unit circle (minimum phase) and the second sum has its zeroes inside the unit circle (maximum phase).

Fourier transforming Equation (1), we find that

$$R_n = \int_{-W}^{W} P(f) \exp\left(-i2\pi f n \Delta t\right) df. \tag{8}$$

Substituting (7) into (8), we find (after a few more transformations) that R_n is given by the contour (complex) integral

$$R_n = \frac{E_M}{2\pi i} \oint_{|Z|=1} \frac{Z^{n-1}}{|\sum a_m Z^m|^2} dZ. \tag{9}$$

The integrand of (9) can have simple poles inside the contour of integration at $Z = 0$ and at any zero of the maximum phase factor. The poles for $Z \neq 0$ can be eliminated by taking a linear combination of Equation (9) "for various values of n." Using the Cauchy integral theorem, we find that

$$\sum_j a_j^* R_{n-j} = \frac{E_M}{2\pi i} \oint_{|Z|=1} \frac{Z^{n-1}}{\sum a_m Z^m} dZ = \begin{cases} E_m & \text{for} \quad n = 0 \\ 0 & \text{for} \quad n > 0, \end{cases} \tag{10}$$

since $a_0 = 1$. Equation (10) and its complex conjugate for the a_m are exactly the standard equations for the maximum and minimum phase spike deconvolution operators $\{a_m^*\}$ and $\{a_m\}$, respectively.

Notice that, if we define the $N \times N$ matrix T_{N-1} as the equidiagonal matrix of autocorrelation values whose elements are given by

$$[T_{N-1}]_{ij} \equiv R_{i-j}, \tag{11}$$

then Equation (10) may be seen as a problem of inverting the matrix T to find the vector $\left\{a_{N-1}^*, \ldots, 1\right\}$. Equation (10) can be solved using the well-known Levinson algorithm for inverting a Toeplitz matrix (**?**). Therefore, a power spectral estimate can be computed by using (10) to find the a_n's, and (7) to compute the spectrum.

One gap in the analysis should be filled before we proceed. That the variational principle is a stationary principle (*i.e.*, $\delta h = 0$) is obvious. That it is truly a maximum principle however requires some proof. First note that the average entropy h computed from substituting (7) into (3) is exactly

$$h = \frac{1}{2} \ln E_M. \tag{12}$$

This fact can be proven by writing (3) as

$$\begin{aligned} 2h = \ln E_M \quad & + \tfrac{M-1}{2\pi i} \oint \ln Z \times \tfrac{dZ}{Z} \\ & - \tfrac{1}{2\pi i} \oint \ln \left(\textstyle\sum_n a_n Z^n\right) \tfrac{dZ}{Z} \\ & - \tfrac{1}{2\pi i} \oint \ln \left(\textstyle\sum_n a_n^* Z^{M-n-1}\right) \tfrac{dZ}{Z}. \end{aligned} \tag{13}$$

The first integral in (13) vanishes identically as is shown in Appendix B. The second integral vanishes because its argument is analytic for all $|Z| < 1$ except for $Z = 0$, and the residue there is $\ln a_0 = 0$. The third integral can be rewritten as

$$\frac{1}{2\pi i} \oint \ln \left(\sum a_n^* Z^{M-n-1}\right) \frac{dZ}{Z} = \ln a_0 + \sum_{n=1}^{M-1} \frac{1}{2\pi i} \oint \ln (Z - Z_n) \frac{dZ}{Z}, \tag{14}$$

where the Z's are the $M - 1$ zeroes of the maximum phase factor ($|Z| < 1$). Each of the integrals on the right side of (14) vanishes because of the identities proven in Appendix B.

For small deviations from the constraining values of R_n, and from the values of R_n computed from (8) once $P_M(f)$ is known, we can expand h in a Taylor series:

$$h = \frac{1}{2}\ln E_M + \sum_{n=-(M-1)}^{M-1} \lambda_n r_n - \sum_{m,n=-\infty}^{\infty} H_{mn} r_m r_n^*. \tag{15}$$

The r_n's are small deviations in the R_n's. The λ_n's are defined by (4). The matrix elements of H are given by

$$H_{mn} = -\frac{\delta^2 h}{\delta R_m \delta R_n^*} = \frac{1}{4W}\int_{-W}^{W} \frac{Z^{n-m}}{P^2(f)} df, \tag{16}$$

with $Z = \exp(-i2\pi f\Delta t)$. H is obviously Hermitian and is seen to be positive definite because

$$\sum_{mn} H_{mn} v_m v_n^* = \frac{1}{4W}\int_{-W}^{W} \frac{|\sum v_m Z^{-m}|^2}{P^2(f)} df \geq 0, \tag{17}$$

where $\{v_n\}$ is an arbitrary complex vector and the equality in (17) holds only when $\{v_n\}$ is identically zero.

The result (17) is sufficient to prove that h is not only stationary, but actually a maximum.

The analysis given in this section has at least two weak points: (a) For real data, we never measure the autocorrelation function directly. Rather, a finite time series is obtained and an autocorrelation estimate is computed. Given the autocorrelation estimate, an estimate of the minimum phase operator must then be inferred. A discussion of various estimates of the autocorrelation is given in the next section on **Computing the Prediction Error Filter**, along with a method of estimating the prediction error filter without computing an autocorrelation estimate. (b) Even assuming we could compute the "best" estimate of the autocorrelation, that estimate is still subject to random error. The probability of error increases as we compute values of R_n with greater lag n. Since there is a one-to-one correspondence between the R_n's and the a_n's, the length of the operator can strongly affect the accuracy of the estimated MESA power spectrum. A method of estimating the optimum operator length for a given sample length N will be discussed in the subsequent section on **Choosing the Operator Length**.

COMPUTING THE PREDICTION ERROR FILTER

When the autocorrelation values $R_0, \ldots, R_{M-1}$ are known, Equation (10) is a linear set of M equations for the M unknown a_M^*'s. On the other hand, if a prediction error filter $\{a_m\}$ and prediction error E_m are known, Equation (10) together with (2) forms a linear set of equations that could be solved for the R_n's. Thus, there exists a one-to-one correspondence between the prediction error filter and the autocorrelation function. This relationship is exploited by **?** in his algorithm for computing the minimum phase operator.

The autocorrelation function defined by (2) requires an infinite series, yet it can only be estimated from a series of finite length N. Given the data set $\{X_1, \ldots, X_N\}$, a reasonable estimate of R_n for large N is given by

$$R_n = \frac{1}{N} \sum_{m=1}^{N-n} X_m^* X_{m+n} \quad \text{for} \quad 0 \leq n \leq N-1. \tag{18}$$

This estimate has at least two shortcomings: (a) Conceptually, the autocorrelation should be an arithmetic average of the $N-n$ lag products in (18). The true arithmetic average is (for $n \geq 0$)

$$R'_n = \frac{1}{N-n} \sum_{m=1}^{N-n} X_m^* X_{m+n} \simeq \frac{N}{N-n} R_n. \tag{19}$$

Equation (19) might be used as the autocorrealtion estimate instead of (18). Unfortunately, this is seldom possible because the Hermitian Toeplitz matrix T defined in (11) is not always nonnegative definite when the definition (19) is used (**?**). A stable operator $\{a_m\}$ cannot be found if T is not nonnegative definite. We conclude that (19) is not a satisfactory estimate of R_n. (b) Suppose for the moment that the matrix T computed using (19) happens to be positive definite. Then each estimated R'_n is being computed from only $N-n$ measurements of the n-lag product, whereas R_0 is estimated from N measurements of the zero-lag product. From measurement theory, it is clear that the uncertainty increases approximately as $(N-n)^{-\frac{1}{2}}$. In fact, this increase in the uncertainty of R_n is unavoidable regardless what choice of estimate for R_n we use as long as N remains finite. One might try to alleviate this problem by using periodic boundary conditions, so that

$$R_n = \frac{1}{N} \sum_{m=1}^{N} X_m^* X_{m+n} \tag{20}$$

and

$$X_{m+N} \equiv X_m. \tag{21}$$

However, this approach merely trades one problem for another one. The periodic assumption introduces spurious peaks into the spectrum by making unfounded assumptions about time series behavior off the ends of the data. Although nevertheless a fairly common approach, this method really cannot improve the accuracy of the computed R_n's for seismic traces having typical lengths.

We conclude that, if an autocorrelation must be computed, then Equation (18) should be used. However, **?** has observed that, in order to compute the maximum entropy spectrum, all that is required is an estimate of the minimum phase deconvolution operator. If this estimate can be computed without first estimating the autocorrelation values, then so much the better.

Suppose an estimate of the operator length M is known ($a_0 = 1$ for $M = 1$). How can the operator length be increased from M to $M+1$? Note that by definition the forward prediction error is given by

$$f_{i+M}(M) = \sum_{j=0}^{M-1} a_j(M) X_{i+M-j} \quad \text{for} \quad 1 \le i \le N-M, \tag{22}$$

and the backward prediction error is

$$b_i(M) = \sum_{j=0}^{M-1} a_j^*(M) X_{i+j} \quad \text{for} \quad 1 \le i \le N-M. \tag{23}$$

Similarly, we have

$$f_{i+M+1}(M+1) = \sum_{j=0}^{M} a_j(M+1) X_{i+M+1-j}, \tag{24}$$

and

$$b_i(M+1) = \sum_{j=0}^{M} a_j^*(M+1) X_{i+j}, \tag{25}$$

which are the linear combinations of (22) and (23) given by

$$f_{i+M+1}(M+1) = f_{i+M+1}(M) + C_{M+1} b_i(M) \tag{26}$$

and

$$b_i(M+1) = b_i(M) + C_{M+1}^* f_{i+M+1}(M). \tag{27}$$

Assuming the value of C_{M+1} is known, (22)–(27) can be used to show that the recursion formulas for the a's are:

$$\begin{array}{l} a_0(M+1) = 1, \\ a_i(M+1) = a_i(M) + C_{M+1} a_{M-i}^*(M) \quad \text{for} \quad 1 \le i \le M-1, \quad \text{and} \\ a_M(M+1) = C_{M+1}. \end{array} \tag{28}$$

Equation (28) is exactly the recursion relation for a minimum phase operator when $|C_i| < 1$ for all $i \le M+1$. Thus, estimating the a's reduces to estimating the C's. A criterion for choosing C_{M+1} is still required.

? suggests that one reasonable procedure is to choose the C_{M+1} that minimizes the total power of the prediction errors. Setting

$$\frac{d}{dC_{M+1}^*} \sum_{i=1}^{N-M-1} \left[|f_{i+M+1}(M+1)|^2 + |b_i(M+1)|^2 \right] = 0, \tag{29}$$

the estimate becomes

$$C_{M+1} = -\frac{2 \sum b_i^*(M) f_{i+M}(M)}{\sum \left[|f_{i+M}(M)|^2 + |b_i(M)|^2 \right]}. \tag{30}$$

Substituting (30) for C_{M+1} into the total power, it is not difficult to show that

$$\begin{aligned} 0 \quad &\leq \tfrac{1}{2}\sum\left[|f_{i+M+1}(M+1)|^2 + |b_i(M+1)|^2\right] \\ &= \tfrac{1}{2}\sum\left[|f_{i+M}(M)|^2 + |b_i(M)|^2\right](1-|C_{M+1}|^2) \\ &= E_M\,(1-|C_{M+1}|^2) = E_{M+1}, \end{aligned} \tag{31}$$

where $E_0 = R_0$. Equation (31) guarantees that $|C_{M+1}| \leq 1$, as is required for a to be minimum phase.

Finally, the algorithm for computing the set $\{a_m\}$ is this: (a) Compute the C's using Equation (30). (b) Store the C's until the desired operator length M is attained. (c) Compute the a's from the C's using the recursion (28). This algorithm (simplified for real data) is the one used in the maximum entropy processor for MESA that I developed.

It is important to notice before proceeding further that the Burg algorithm has been constructed to remove the first difficulty discussed earlier in computing R_n. All the information $\{X_n\}$ has been used; the operator is minimum phase; but no explicit averaging of lag products was required. On the other hand, this algorithm does nothing to alleviate the second problem we discussed. It is still inherent in the finite time series problem that the numbers we compute become less reliable as the operator length increases.

A major difficulty in applying MESA is that there is no built-in mechanism for choosing the operator length. From the derivation of (5), it is clear the operator length should be N if the first N autocorrelation values are known precisely and unknown otherwise. However, the autocorrelation function has (normally) been estimated from the time series data and its estimated values are inaccurate for n close to N. How to choose a practical operator length M satisfying $1 < M < N$ is therefore the subject of the next section.

CHOOSING THE OPERATOR LENGTH

Numerous procedures for choosing the operator length have been discussed in the literature. In this section of the paper, first a discussion of the general principles behind the operator length optimization is given. Then several of the most prominent practical methods in use are critically reviewed. Finally, a new criterion is derived which is easier to apply and believed to be more appropriate for use in MESA.

A mean square error criterion

One theoretically sound procedure for choosing a truncation point is based on a mean square error criterion (?). Suppose that $P_M(f)$ is our M-th estimate of the true spectrum $P(f)$. Then we might wish to mimimize the square error

$$E[P_M(f) - P(f)]^2 = Var[P_M(f)] + B^2[P_M(f)], \tag{32}$$

where E is expectation, Var is variance

$$Var(P_M) = E(P_M^2) - E^2(P_M), \tag{33}$$

and B is bias

$$B[P_M(f)] = P(f) - E[P_M(f)]. \tag{34}$$

In general, as M increases, the bias decreases while the variance increases. Thus, (32) will have a minimum for some value of M.

This criterion is not of practical value unless it is possible to obtain reasonably good estimates of the variance and bias of P_M. This problem is not easily solved, but satisfactory approximate solutions can probably be found. However, this approach will NOT be pursued here.

This notation has been introduced to help the reader understand why one should *expect* such an optimum operator length to exist. Spectral estimates are nearly always designed to decrease the bias as M increases. (MESA is clearly designed this way.) However, when the bias *is* small, the variance is a measure of the ragged oscillations $P_M(f)$ makes around $P(f)$. Since most people prefer to study a smooth spectrum, a balance between variance and bias is our goal.

The common criteria

A number of fairly simple criteria are commonly discussed in the literature. Some of these will be reviewed here.

? suggest monitoring the magnitude of C_M to determine the operator length empirically. Their criterion is to choose that M value for which C_M first satisfies $|C_M| << 1$. The argument is that C_M computed from (30) is "a partial correlation coefficient," measuring the correlation between the forward and backward prediction errors. When $|C_M| \simeq 1$, the correlation is high. When $|C_M| << 1$, the correlation is low — presumably because most of the predictable information in the data has been removed by the filter. However, they point out that this procedure fails to produce reliable results for series not purely autoregressive in character. Numerical studies of the author on real seismic traces have shown the fluctuations in $|C_M|$ to be too great for this approach to give a reliable criterion.

? review a number of possible approaches. The two which are probably easiest to apply are the F-test and the relative error coefficient test: (a) The F-test monitors E_M and checks whether the change in going from E_M to E_{M+1} is statistically significant according to some predetermined criterion. This method is limited by computer round off error for large data samples. It is also limited for small data samples because the predetermined criterion of statistically significant change may very well be met for all $M \leq N$. (b) The relative error coefficient test amounts to finding the minimum of the modified prediction error

$$E'_M = \frac{N}{N-M} E_M. \tag{35}$$

The prediction error is modified in this manner to account for the decrease in the degrees of statistical freedom for the time series as the operator length increases. Clearly, E_M decreases whereas the multiplicative factor increases as M increases. E'_M will therefore exhibit a minimum. A number of such minima can (and do) occur in practice. The parameter E'_M is easily monitored while computing the C_M's using the Burg algorithm. Results obtained using this approach have been found satisfactory for moderate to large data samples. For small N, the variations in both factors in (35) can be dramatic and the results become less reliable.

? review a number of alternatives and conclude that the final prediction error (FPE) criterion of **?** is an objective basis for choosing the operator length. This criterion monitors

$$(FPE)_M = \frac{N+M}{N-M} E_M. \tag{36}$$

Like (35), this expression has a minimum since E_M decreases monotonically while the multiplicative factor increases monotonically with M. In fact, (35) and (36) have very similar behavior, the principal difference being that "when E_M is sufficiently smoothly varying," the minimum of (36) always occurs for smaller M values than that of (35). For short time series with sharp spectral lines, **?** found that FPE did not give a clear minimum. Both (35) and (36) suffer from this same ambiguity. For data samples of length $20 \leq N \leq 40$ in their work, they found that $M = N/2$ was a satisfactory choice. This choice is also confirmed for short time series by the work of **?**.

Although each of these criteria has its merits, none of them is really satisfactory for a data sample of arbitrary length. Furthermore, none of them has been derived in the spirit of MESA, *i.e.*, with no assumptions about the data off the ends of the sample. Much has been said about the application of optimum criteria from autoregressive analysis to MESA (**?**). But an important point should be made: The fact that an autoregressive process has the maximum entropy is interesting but irrelevant. The spectrum of an *arbitrary* time series (whether MA, AR, or ARMA of any order) can be *estimated* using MESA. But, making *any* assumption about the nature of the stochastic process that generated the series is contrary to the spirit of MESA.

Thus, it seems that the choice of operator length should be made without assumptions concerning the nature of the stochastic process involved. The argument in the next subsection is based only on information theory, and measurement theory. It is believed to free of these inconsistencies.

An information theory criterion

Suppose we have found an estimate of the prediction error filter of length M using the autocorrelation estimates $R_0, \ldots, R_{M-1}$. In order to increase the operator length to $M+1$, additional information is needed: namely, R_M. A quantitative measure of the information in the operator is easily obtained from the average entropy, which we know is given by $h'_m = \frac{1}{2}\ln E_M$. Using (31), notice that

$$h'_{M+1} = \frac{1}{2}\ln E_M + \frac{1}{2}\ln\left(1 - |C_{M+1}|^2\right) \le h'_M. \tag{37}$$

Thus, the entropy *decreases* as the operator length *increases.* The bound information (Brillouin, 1956) I_M in the power spectrum is therefore given by

$$I'_M = -h'_M = -\frac{1}{2}\ln E_M, \tag{38}$$

which obviously [since, from (37), we have $-h'_M \le -h'_{M+1}$] increases monotonically with M as it should.

If the autocorrelation values $R_0, \ldots, R_{n-1}$ were known precisely, bound information would continue to increase by using all the estimates and letting $M \to N-1$. But the R's are not precisely known. The finite number of measurements used to compute the R estimates means that only $N-n$ measurements of R_n were made, whereas N measurements of R_0 were made. The quality of information contained in R_0 is correspondingly higher than that in R_n. A quantitative measure of this change is therefore required.

For the moment, take Equation (19) as our estimate of the autocorrelation. Then, assuming that the X_i's are normally distributed, **?** shows that

$$Var(X_i^* X_{i+n} - R_n) = R_0^2 + R_n^2. \tag{39}$$

Since $|R_0| \ge |R_n|$, for all n, and generally $|R_0| >> |R_n|$, for large n, the variance (39) can be approximated by the constant R_0^2. Viewing R_n as a measured quantity (which in fact it usually is not) and using standard arguments from measurement theory, we find that

$$R_n = \frac{1}{N-n}\sum_m X_m^* X_{m+n} \pm 0.67\frac{R_0}{\sqrt{N-n}}, \tag{40}$$

with fifty percent confidence if R_n is also normally distributed.

The probable error in R_n increases like $(N-n)^{-1/2}$ as $n \to N-1$. We imagine that the factor $(N-n)^{-1/2}$ is proportional to the probability p_n that an operator computed from $R_0, \ldots, R_N$ is a worse estimate of the true opertor than was the operator computed using only $R_0, \ldots, R_{n-1}$. Since we know empirically that the estimate worsens as $M \to N$ with probability one, the $P(n)$ are normalized by writing:

$$P(n) = \alpha(N-n)^{-\frac{1}{2}} \tag{41}$$

and

$$1 = \sum_{n=0}^{N-1} P_n \simeq \alpha\int_0^N (N-n)^{-\frac{1}{2}} dn = 2\sqrt{N}\alpha. \tag{42}$$

Equation (42) determines the value of α, for the data that are available.

The average entropy of measurement error associated with an operator of length M is

$$\begin{aligned} h''_M &= -\textstyle\sum_{n=0}^{M-1} P(n)\ln P(n) \\ &\simeq -\textstyle\int_0^M P(n)\ln P(n)dn. \end{aligned} \tag{43}$$

The second line of (43) is valid for large N. The value of h''_M increases as M increases in agreement with our intuition. It is well known that the largest average entropy for N probabilities is $\ln N$. Letting $M \to N$ in (43), we find

$$h''_N = \ln\left(\frac{2N}{e}\right) < \ln N, \tag{44}$$

which is consistent.

Combining (38) and (43), the average information in the power spectrum can be quantitatively estimated using the expression

$$I_M = -\left(h'_M + h''_M\right) = -\frac{1}{2}\ln E_M + \int_0^M P(n)\ln P(n)dn. \tag{45}$$

The first term increases while the second term decreases with increasing M. A maximum will occur for some value $1 < M < N$. The spectrum with the maximum information is the optimum spectrum; the value of M that maximizes (45) is the value we are seeking.

The values of (45) can be monitored continuously while the operator is being computed. However, an approximate analytic solution for the maximum can be found without making very restrictive assumptions on the behavior of E_M. Numerical studies of the author on real seismic data have shown that E_M can be represented approximately by

$$E_M \propto M^{-\beta}, \tag{46}$$

where β is a slowly varying function of M. Generally, β is in the range $2 \ge \beta \ge \frac{1}{2}$, with $\beta \simeq 2$ for small M and $\beta \to \frac{1}{2}$ for large M. Leaving β arbitrary for the moment, substituting (46) into (45), and finding the stationary point, we have

$$\frac{\beta}{2}M^{-1} = -P(M)\ln P(M). \tag{47}$$

Using (42) for α, Equation (47) can be solved graphically for M. The solution for $\beta = 2$ is plotted as the solid line in Figure 1.

An analytic bound on M can be obtained from (47) by noting that the right-hand side of (47) increases with M, so its minimum value occurs when $M = 0$. Thus, M has the very simple bound:

$$M \le \beta\frac{N}{\ln 2N}. \tag{48}$$

Since we have stated already that $\beta \le 2$ in general, a useful bound on M for all N appears to be

$$M \le \frac{2N}{\ln 2N}. \tag{49}$$

Figure 1 compares the values of M obtained from (47), from (48), and from $M = N/2$. The value $\beta = 2$ is chosen because of the empirical evidence mentioned above and also because

$$|h'_M| \leq \frac{\beta}{2} \ln M \leq \ln N \tag{50}$$

is valid for all $M \leq N$ only for $\beta \leq 2$. The comparison with $M = N/2$ is of interest because various authors (including this one) have often found this value to be satisfactory for small N. The derivation given above is strictly valid only for large N. But the estimate (49) interpolates well between these extremes as is seen in Figure 1.

Because the correspondence between $P(n)$ and $(N-n)^{-\frac{1}{2}}$ has been established by this heuristic argument, the results of this section of the paper should not be interpreted as rigorous estimates of the optimum operator length. Nevertheless, I believe that (47) and (49) are reasonable estimates of the operator length. The derivation was not founded on any assumptions about the type of stochastic process generating the time series. Hence, these estimates are definitely *not* intended to be an estimate of the order of some underlying autoregressive process. Rather, (49) is an upper bound on the operator length that will extract the most reliable information for a data sample of length N. For example, suppose the time series $\{X_1, \ldots, X_N\}$ is a representation of an autoregressive series of order $L \leq M$. Then computing the operator of length L should give the most efficient estimate of the spectrum; but computing the additional $(M - L)$ terms should do little to alter that spectrum. Next, suppose the time series is a representation of an AR series of order $L > M$. The arguments above indicate that we probably cannot obtain a really good estimate of the operator (or the spectrum), because our data sample is simply too small. The best we can hope to do is to compute the operator of length M. In either case, when additional information about the underlying stochastic process is lacking, the best operational decision that can be made appears to be choosing M according to Equations (47) or (49).

EXAMPLES

One good way to study various choices of operator length is to use unexpanded checkshot data. Two traces of this type are shown in Figure 2. By choosing a large window, we obtain a good estimate of the spectrum of the pulse. Then, by choosing smaller and smaller windows which pinch down on the pulse, we should expect the positions of the major peaks to remain unchanged while resolution becomes more difficult.

The first column of Figure 3 gives the ordinary power spectrum of Figure 2(a) computed using five different length windows. In seconds, the windows are from top to bottom (0.0,2.0), (0.1,0.5), (0.15,0.45), (0.15,0.25), and (0.175,0.225). No taper is included in the Fourier transform of the trace. We see that the resolution is quite

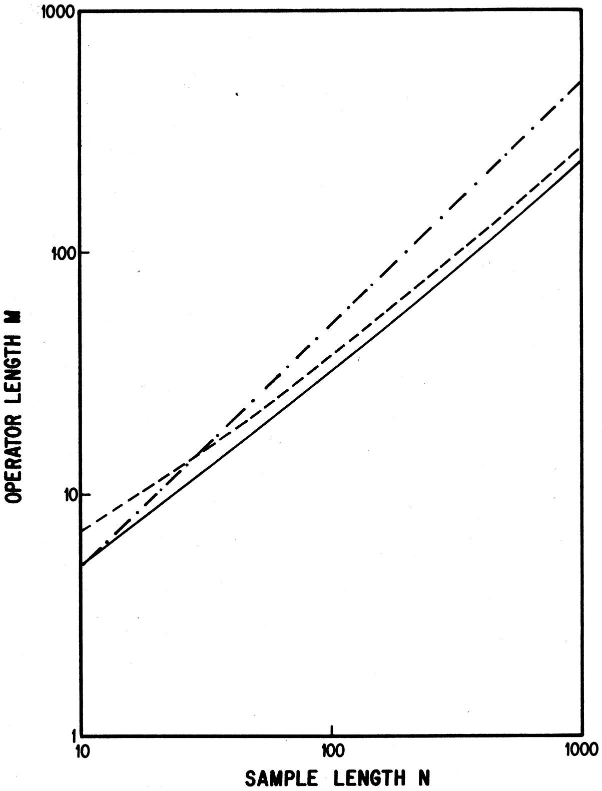


Figure 1: Operator length M as a function of data sample length N for three different operator length estimates. The solid line is the solution of Equation (47). The dash line is $M = 2n/\ln(2N)$. The dot-dash line is $M = N/2$. [**NR**]

good for the two second window, but the resolution gets progressively worse until it is almost nonexistent for the 50 ms window.

In contrast, the first column of Figure 4 gives the maximum entropy spectrum of Figure 2(a) with the operator length $M = 2N/\ln 2N$. We see that the two second window MESA spectrum is essentially the same as the ordinary spectrum. However, as the number of data points decreases from 500 (for a two second window) to 13 (for a 50 ms window), we see that MESA is still able to resolve the peaks at 10 Hz and 30 Hz.

The first column of Figures 5–7 give examples of the results obtained from MESA for the trace of Figure 2(a) with other choices of operator length. Choosing $M = N/2$ in Figure 5 gives acceptable results for all but the smallest window where the 10 Hz peak has moved towards 20 Hz. Also, the computation time was increased for the longest three windows. Choosing $M = N$ in Figure 6 demonstrates the fact that choosing a longer operator does *not* lead to improved results. Here the single peak at 10 Hz has been split into two spurious peaks for the shortest two windows. The longest three windows give very spiky spectra and (although they properly indicate where the spectrum lies) they do not give useful power spectra. This Figure shows how the variance in (32) can dominate the bias and produce useless power spectra. Finally, we choose $M = N/\ln N$ in Figure 7 to demonstrate the effect of choosing a slightly different functional form for M. The spectrum of Figure 4 mimics that of Figure 3 better than Figure 7 in all cases except possibly for the 100 ms window where Figure 7 gives a stronger peak at 30 Hz. For the smallest window, Figure 7 does not have its peak at 10 Hz as it should.

Figure 2(b) is also an unexpanded checkshot trace which is translated in time from that in Figure 2(a). The second columns of Figures 3–7 were computed as before, but the input trace was Figure 2(b).

As a final note, we wish to point out that it has been observed in general that $M = 2N/\ln 2N$ is in fact an upper bound on the operator lengths one would obtain from either (35) or (36). For example, using the trace of Figure 2, both the relative error coefficient test and the FPE have a series of minima for $120 < M < 135$ with an absolute minimum at $M = 132$ for both criteria. For comparison, we find $M = 145$ for $N = 500$ using (49) and $M = 127$ using (47). It is encouraging that the arguments of the subsection on **An Information Theory Criterion** give estimates for the operator length so close to those of Equations (35) and (36) without the added complication of monitoring a performance parameter.

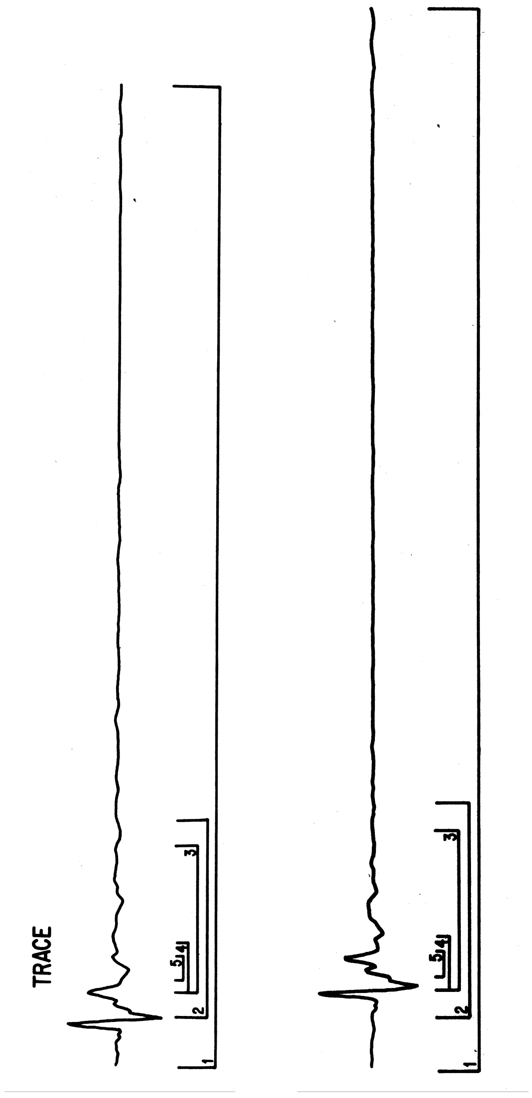
DISCUSSION

As with many procedures in seismic data enhancement and analysis, computation of power spectral estimates with MESA can be more art than science. Considerable insight into the stochastic processes involved and experience with choices of operator length is required before MESA can be considered a standard processing tool. The work summarized here was intended to eliminate some of the uncertainty in applying MESA by producing a reasonable upper bound [Equation (49)] on the operator length. In many cases, this upper bound will itself be a good choice for the operator length, since it often gives values comparable to those of other methods without requiring performance parameter monitoring.

In conclusion, I recommend that anyone wishing to obtain high resolution power spectra for seismic traces should examine both maximum entropy spectra and ordinary spectra. When both methods give peaks approximately at the same frequencies, we may be confident that the maximum entropy method is giving higher resolution of true modes in the spectrum. If the MESA peaks are not reasonably close to the peaks obtained by the better understood Fourier transform method, we should give additional and careful consideration to the proper choice of the operator length for the MESA spectrum.

ACKNOWLEDGMENTS

I thank Robert H. Stolt for helpful discussions and Jerry A. Ware for encouragement. I also thank Bob Clapp for help with the preparation of the final figures.



(a) (b)

Figure 2: Two checkshot traces: (a): The first two seconds of an unexpanded checkshot trace. The windows indicated in the Figure are repectively: $(0.0, 2.0)$, $(0.1, 0.5)$, $(0.15, 0.45)$, $(0.15, 0.25)$, and $(0.175, 0.225)$ in seconds. (b) Same as previous case for a slightly different trace. [**NR**]

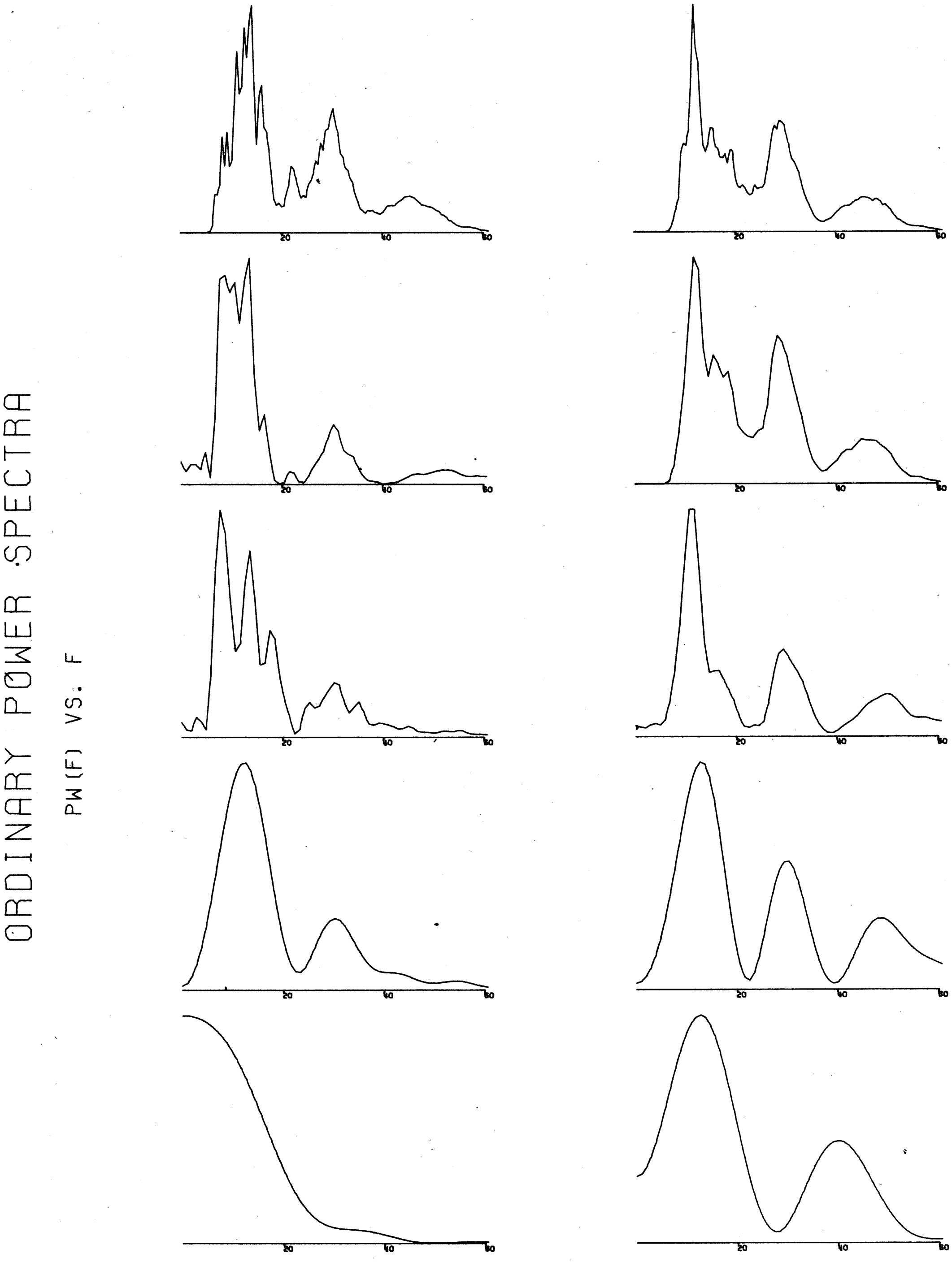


Figure 3: The ordinary power spectrum of the traces in Figure 2 as a function of frequency (0–60 Hz). Windows from top to bottom are (0.0, 2.0), (0.1, 0.5), (0.15, 0.45), (0.15, 0.25), and (0.175, 0.225) in seconds. [**NR**]

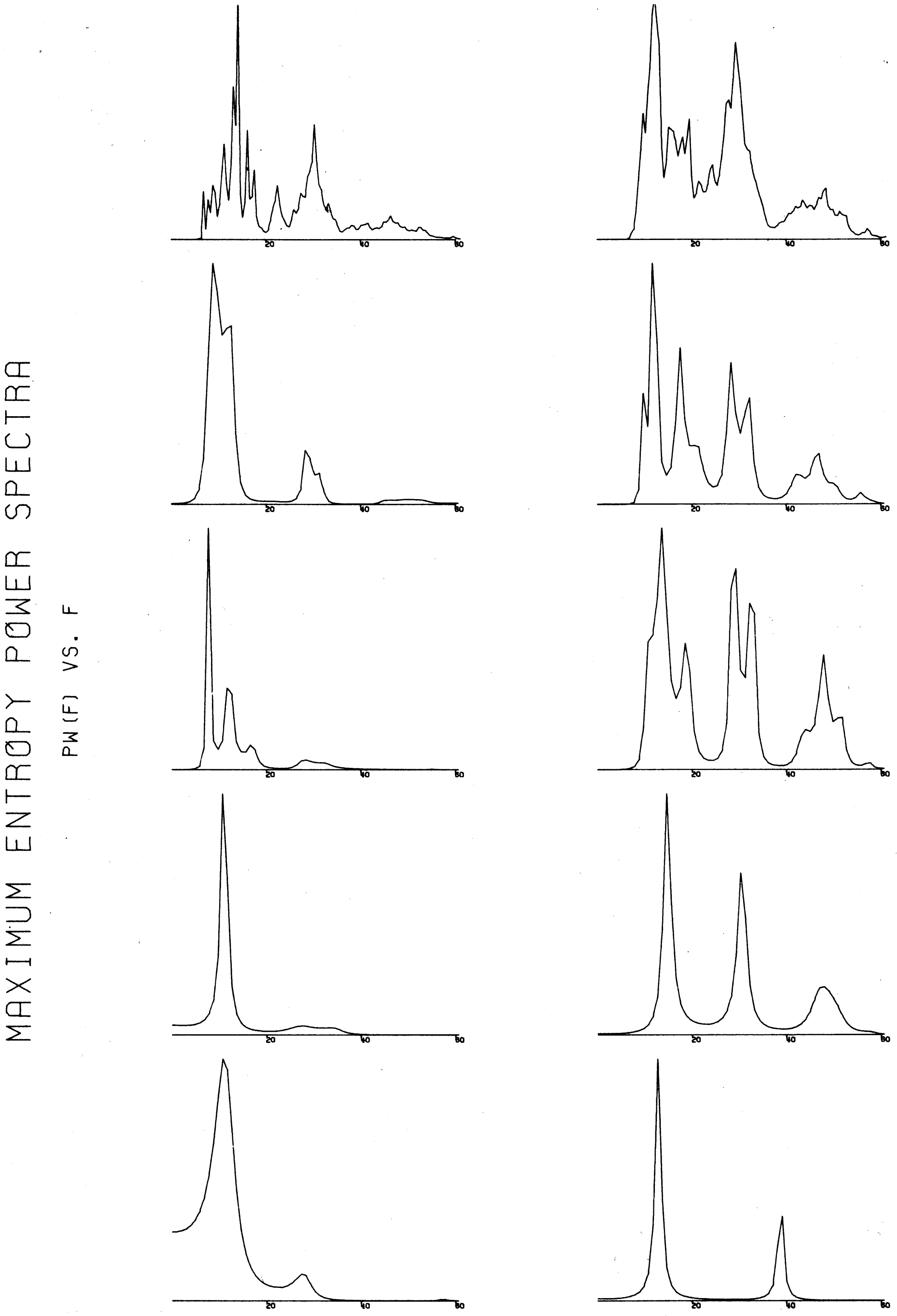


Figure 4: The maximum entropy power spectrum of the traces in Figure 2 as a function of frequency (0–60 Hz) with operator length $M = 2N/\ln(2N)$. Windows same as in Figure 3. Windows from top to bottom are $(0.0, 2.0)$, $(0.1, 0.5)$, $(0.15, 0.45)$, $(0.15, 0.25)$, and $(0.175, 0.225)$ in seconds. [**NR**]

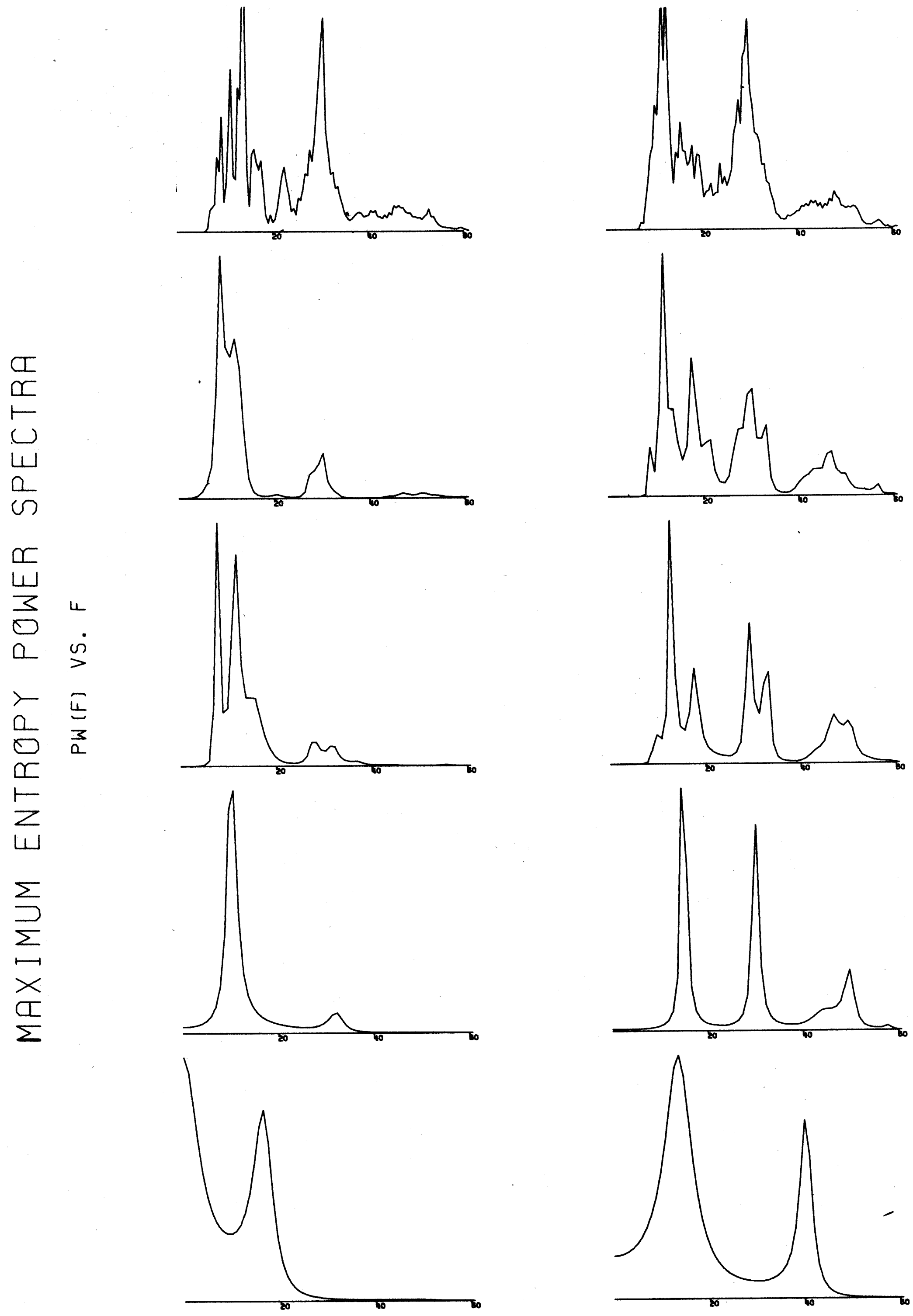


Figure 5: The maximum entropy power spectrum of the traces in Figure 2 as a function of frequency (0–60 Hz) with operator length $M = N/2$. Windows same as in Figure 3. Windows from top to bottom are $(0.0, 2.0)$, $(0.1, 0.5)$, $(0.15, 0.45)$, $(0.15, 0.25)$, and $(0.175, 0.225)$ in seconds. [**NR**]

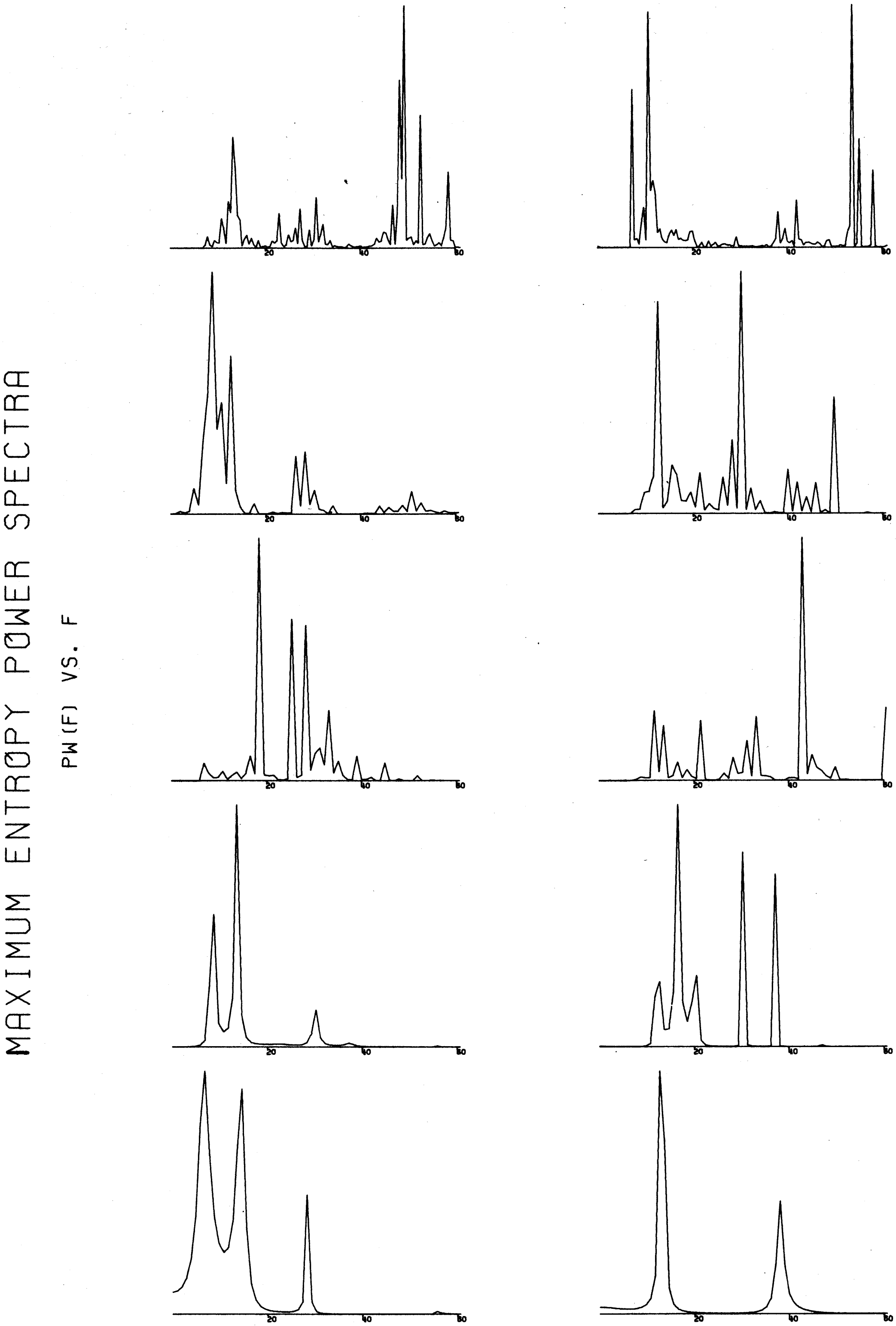


Figure 6: Same as Figure 5 with $M = N$. Windows from top to bottom are $(0.0, 2.0)$, $(0.1, 0.5)$, $(0.15, 0.45)$, $(0.15, 0.25)$, and $(0.175, 0.225)$ in seconds. [**NR**]

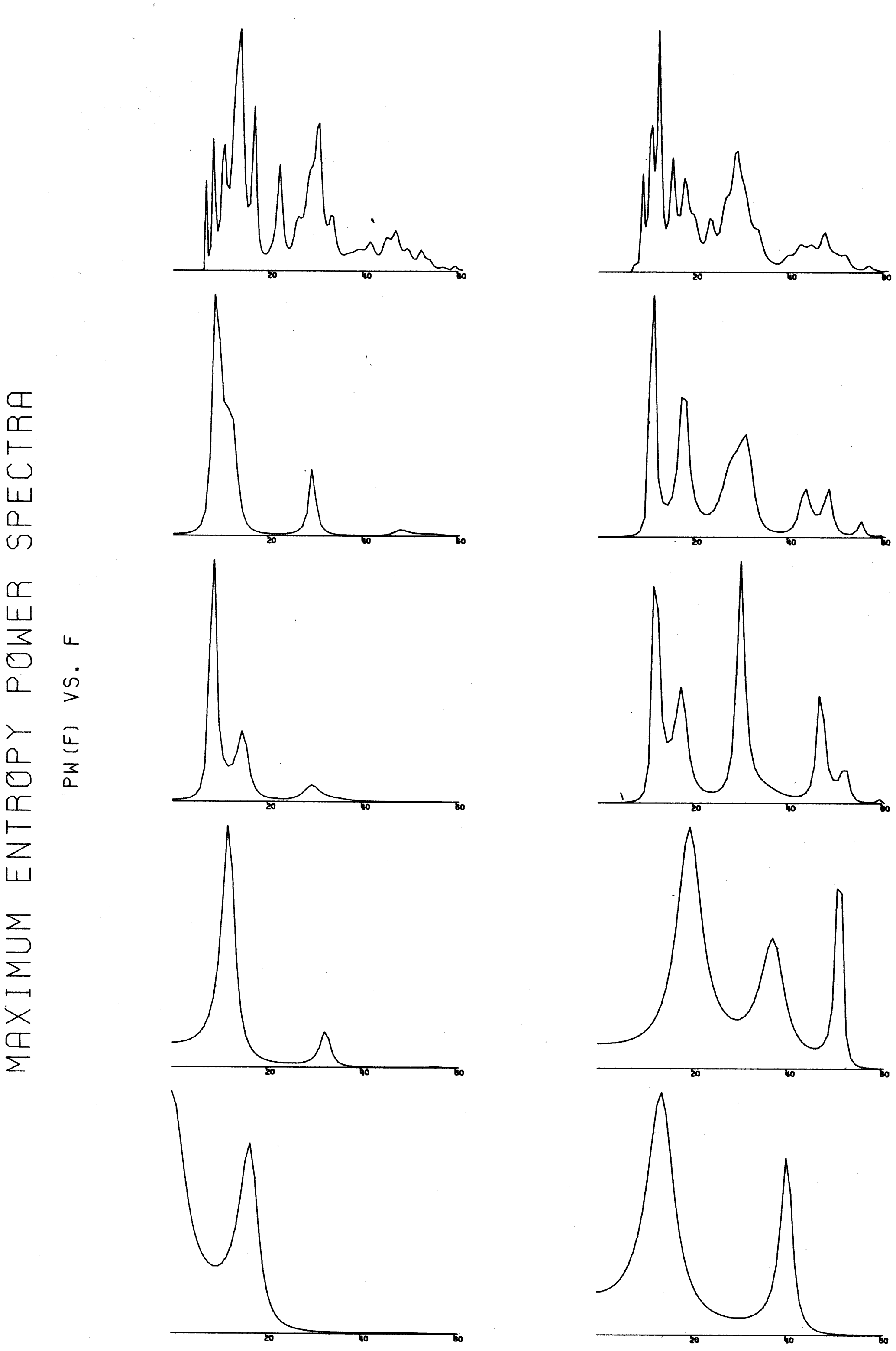


Figure 7: Same as Figure 5 with $M = N/\ln N$. Windows from top to bottom are (0.0, 2.0), (0.1, 0.5), (0.15, 0.45), (0.15, 0.25), and (0.175, 0.225) in seconds. [**NR**]

APPENDIX A

The following derivation of the relationship between entropy and power spectrum is essentially the same as that given by **?**. The derivation is included here for completeness.

The entropy of N random variables X_1, …, X_N is given by

$$H_N = -\int P(X_1,\ldots,X_N)\ln[a^{2N}P(X_1,\ldots,X_N)]d^NX = -2N\ln a - \int P\ln Pd^NX, \tag{A-1}$$

where P is the joint probability density and a is a constant with the same units as X. The power spectrum $P(f)$ computed from the autocorrelation values $R_0,\ldots,R_{N-1}$ depends only on the second-order statistics of the time series $\{X_n\}$. Therefore, the given time series cannot be distinguished from a normal (Gaussian) process.

The joint probability density for a normal process with N variables of zero mean is (using matrix notation, where X^T is the transpose of X)

$$P(X_1,\ldots,X_N) = \left[(2\pi e)^N\det T_{N-1}\right]^{-\frac{1}{2}}\exp\left(-\frac{1}{2}X^T\cdot T_{N-1}^{-1}\cdot X\right), \tag{A-2}$$

where T_{N-1} is the $N\times N$ Toeplitz matrix [see **?**] given by Equation (11) and X is the N-vector determined by $X^T=(X_1,\ldots,X_N)$. Substituting (A-2) into (A-1), we find

$$H_N = \frac{1}{2}\ln\left[(2\pi e)^N\det T_{N-1}\right] - 2N\ln a. \tag{A-3}$$

Setting the arbitrary constant $a=(2\pi e)^{1/4}$ for convenience, Equation (A-3) then becomes

$$H_N = \frac{1}{2}\ln\left(\det T_{N-1}\right). \tag{A-4}$$

Since (A-4) necessarily diverges as $N\to\infty$, a better measure of the information content of the series is the average entropy per variable given by

$$h = \lim_{N\to\infty}\frac{H_N}{N} = \lim_{N\to\infty}\ln\left(\det T_{N-1}\right)^{1/2N}. \tag{A-5}$$

The eigenvalues $\{\lambda_1,\ldots,\lambda_N\}$ of T_{N-1} are real and nonnegative since T is Hermitian and nonnegative definite. Furthermore,

$$\det T_{N-1} = \Pi_{i=1}^N\lambda_i, \tag{A-6}$$

so

$$h = \lim_{N\to\infty}\frac{1}{2N}\sum_{i=1}^N\ln\lambda_i. \tag{A-7}$$

The Szëgo theorem (**??**) states that, if F is any continuous function, then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N F(\lambda_i) = \frac{1}{2W}\int_{-W}^{W}F[2WP(f)]\,df, \tag{A-8}$$

where, as before, W is the Nyquist frequency, $P(f)$ is the power spectrum, and the λ's are the N eigenvalues of T_{N-1}.

Combining Equations (A-7) and (A-8), we find

$$h = \frac{1}{4W}\int_{-W}^{W} \ln[2WP(f)]\,df, \tag{A-9}$$

which is the sought after result.

APPENDIX B

We need to compute the integral

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left[\exp(i\theta) - Z_0\right] d\theta = \frac{P.V.}{2\pi i}\oint_{|Z|=1} \ln\left(Z - Z_0\right)\frac{dZ}{Z}, \tag{B-1}$$

where *P.V.* stands for the principal value of the contour (complex) integral when the logarithm's branch cut is taken along the negative real axis.

First, note that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left[\exp(i\theta) - Z_0\right] d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\exp(i\theta) d\theta + \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left[1 - \exp(-i\theta)Z_0\right]\, d\theta. \tag{B-2}$$

The first integral on the right is just

$$-\frac{1}{2\pi i}\int_{-\pi}^{\pi} \theta d\theta = 0, \tag{B-3}$$

since the integrand is an odd function. When $|Z_0| < 1$, the integrand of the second integral on the right can be expanded in a convergent power series. Integrating term by term, we find that

$$-\frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_{n=1}^{\infty} \frac{Z_0^n}{n} \exp\left(in\theta\right) d\theta = 0, \tag{B-4}$$

since $\exp\left(in\pi\right) - \exp\left(-in\pi\right) = \cos(n\pi) + i\sin(n\pi) - \cos(-n\pi) - i\sin(-n\pi) = 0$ (the two cosines cancel and the two sines both vanish individually for all integer values of n). Thus, we find (B-1) is identically zero for all $|Z_0| < 1$. In particular, it vanishes when $Z_0 = 0$, so

$$\frac{P.V.}{2\pi i}\oint_{|Z|=1} \ln Z \times \frac{dZ}{Z} = 0. \tag{B-5}$$

REFERENCES