

E2.1

A Tutorial Overview of Modern Spectral Estimation

S. Lawrence Marple Jr.

ARCO Power Technologies, Inc.
1250 Twenty-Fourth Street, N.W., Suite 850
Washington, D.C. 20037
U.S.A.

ABSTRACT A summary of several modern spectral estimation methods is presented in this tutorial. Most of the methods may be explained in the context of parametric time series modeling. A few methods involve nonparametric treatment. Techniques discussed include classical spectral estimation, autoregressive (maximum entropy), ARMA, Prony, maximum likelihood, Pisarenko, and MUSIC methods. Many of the techniques have fast computational algorithms, making these methods viable for real time applications. The tutorial concludes with a commentary concerning current spectral estimation research.

1. INTRODUCTION

Extensive research in the last decade has led to a proliferation of new digital spectral estimation techniques as alternatives to the classical periodogram estimate. The acceleration of research has been promoted by the promise of improved performance, such as higher frequency and spatial resolution or increased signal detectability. The promise seen in the research environment has not always been fulfilled in practical applications, particularly when real signals, rather than computer-generated synthetic signals, are used. The basis for the disappointment can be traced to the testing methods found in the literature. More commentary on this issue is presented in Section 8, following a tutorial presentation of several key spectral estimation methods.

The unifying approach employed in this tutorial is to view each spectral estimation technique as based on the fitting of measured data to an assumed model. The variations in performance among the various spectral estimates may often be attributed to how well the assumed model matches the process under analysis. Different models may yield similar results, but some models may require fewer model parameters than other models and be therefore more efficient in their representation of the process.

Estimation of the power spectral density (PSD), or simply the spectrum, of discretely sampled deterministic and stochastic processes is usually based on procedures employing the Fast Fourier Transform (FFT). This approach to spectral analysis is computationally efficient and produces reasonable results for a large class of signal processes. In spite of these advantages, there are several inherent performance limitations of the FFT approach. The most prominent limitation is that of frequency resolution, i.e., the ability to distinguish the spectral responses of two or more signals. The frequency resolution in Hertz is roughly the reciprocal of the time interval in seconds over which sampled data is available. A second limitation is due to the implicit windowing of the data that occurs when processing with the FFT. Windowing manifests itself as "leakage" in the spectral domain, i.e., energy in the main lobe of a spectral response "leaks" into the sidelobes, obscuring and distorting other nearby spectral responses that are present.

These two performance limitations of the FFT approach are particularly troublesome when analyzing short data records. Short data records occur frequently in practice because many measured processes are brief in duration or have slowly time-varying spectra that may be considered constant only for short record lengths. In radar, for example, only a few data samples are available from each received radar pulse. In sonar, the motion of targets results in a slowly time-varying spectral response due to Doppler effects.

In an attempt to alleviate the inherent limitations of the FFT approach, many alternative spectral estimation procedures have been proposed within the last decade. The apparent improvement in resolution provided by these techniques have fostered their popularity, even though classical FFT-based spectral estimation has been shown to often provide better performance at very low signal-to-noise ratios. Even in those cases where improved spectral fidelity is achieved by use of an alternative spectral estimation procedure, the computational requirements of that alternative method may be significantly higher than the FFT processing required to compute the periodogram. This makes some modern spectral estimators unattractive for real-time implementation.

The nonparametric, classical, one-dimensional time (or space) series spectral estimation methods are presented in Section 3. Parametric spectral estimation techniques are covered in Section 4, including many of the so-called modern spectral estimation methods. The maximum likelihood spectral estimate, Pisarenko's method, and the MUSIC method are described in Section 5. A short discussion of multidimensional maximum entropy spectral estimation may be found in Section 6. Additional spectral estimation methods are presented in greater detail in reference [1].

2. SPECTRAL DENSITY BASICS

Traditional spectrum estimation, as currently implemented using the FFT, is characterized by many tradeoffs in an effort to produce statistically reliable spectral estimates. There are tradeoffs in windowing, time-domain averaging, and frequency-domain averaging of sampled data obtained from random processes in order to balance the need to reduce sidelobes, to perform effective ensemble averaging, and to ensure adequate spectral resolution. The spectrum analysis of a random process is, in concept, not obtained directly from the process $x(t)$ itself, but is based on knowledge of the autocovariance function assuming a zero mean process

$$R_{xx}(\tau) = E[x(t+\tau)x^*(t)] \quad (2.1)$$

The Wiener-Khinchin theorem relates $R_{xx}(\tau)$ via the Fourier transform to $P(f)$, the power spectral density (PSD),

$$P(f) = \int_{-\infty}^{\infty} R_{xx}(\tau) \exp(-j2\pi f\tau) d\tau \quad (2.2)$$

As a practical matter, one does not usually know the statistical autocovariance function. Thus, an additional assumption often made is that the random process is ergodic in the first and second moments. This property permits the substitution of time averages for ensemble averages. For an ergodic process, the statistical autocovariance function may then be equated to

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t+\tau)x^*(t) dt \quad (2.3)$$

It is possible to show, with the use of (2.3), that (2.2) may be equivalently expressed as

$$P(f) = \lim_{T \rightarrow \infty} E \left\{ \frac{1}{2T} \left| \int_{-T}^T x(t) \exp(-j2\pi ft) dt \right|^2 \right\} \quad (2.4)$$

The expectation operator is required since the ergodic property of $R_{xx}(\tau)$ does not necessarily imply that the Fourier transform of the process $x(t)$ is also ergodic; this means that the limit in (2.4) without the expectation operation will not converge in any statistical sense.

Attempting to estimate $P(f)$ with finite data sets using (2.4), without taking into consideration the expectation operation and the limit operation, can lead to meaningless spectral estimates if no statistical averaging is performed; i.e., the variance of the PSD estimate will not tend to zero as T increases without bound.

3. CLASSICAL METHODS

Two spectral estimation techniques based on Fourier transform operations have evolved [2]. The PSD estimate based on the indirect approach via an autocorrelation estimate was popularized by Blackman and Tukey. The other PSD estimate, based on the direct approach via an FFT operation on the data, is the one typically referred to as the periodogram.

With a finite data sequence, only a finite number of discrete autocorrelation function values, or lags, may be estimated. Blackman and Tukey proposed the spectral estimate

$$\hat{P}_{BT}(f) = \Delta t \sum_{m=-M}^M \hat{R}_{xx}(m) \exp(-j2\pi f m \Delta t) \quad (3.1)$$

based on the available biased autocovariance lag estimates $\hat{R}_{xx}(m)$,

$$\hat{R}_{xx}(m) = \frac{1}{N} \sum_{k=0}^{N-m} x_{k+m} x_k^* \quad (3.2)$$

where $-1/(2\Delta t) \leq f \leq 1/(2\Delta t)$ and $\hat{\cdot}$ denotes an estimate. This spectral estimate is the discrete-time version of the Wiener-Khinchine expression (2.2).

The direct method of spectrum analysis is the modern version of Schuster's periodogram. A sampled data version of expression (2.4), for which measured data is available only for samples x_0, \dots, x_{N-1} , is

$$\hat{P}_{PER}(f) = \frac{1}{N\Delta t} \left| \sum_{n=0}^{N-1} x_n \exp(-j2\pi f n \Delta t) \right|^2 \quad (3.3)$$

also defined for the frequency interval $-1/(2\Delta t) \leq f \leq 1/(2\Delta t)$. Note that the expectation operation in (2.4) has been ignored for the moment. Use of the FFT will permit evaluation of (3.3) at the discrete set of N equally spaced frequencies $f_m = m\Delta f$ Hz, for $m = 0, 1, \dots, N-1$ and $\Delta f = 1/N\Delta t$.

$$\hat{P}_m = \hat{P}_{PER}(f_m) = \frac{1}{N\Delta t} |X_m|^2 \quad (3.4)$$

where X_m is the discrete Fourier transform (DFT) at frequency index m . \hat{P}_m is identical to the energy spectral density [squared modulus of the Fourier transform of a time function], except for the division by the time interval of $N\Delta t$ seconds required to make \hat{P}_m a power spectral density.

In order to reduce the bias of the Blackman-Tukey spectral estimate due to the implied rectangular window over a finite number of lag estimates, various windows other than rectangular are often applied to the lag estimates in order to suppress the sidelobes, and therefore reduce the bias. In order to emulate ensemble averaging needed to make the periodogram a statistically consistent spectral estimate, two basic approaches have evolved. The averaging of modified periodograms is one such technique; it breaks the original time sequence into segments, windows each segment (in order to reduce the bias), calculates a periodogram of each modified segment, and then averages the segment periodograms. A particular version of this averaging method using the FFT algorithm was proposed by Welch (see his paper in [3]). A second method is to compute the periodogram as in (3.4), and then average over adjacent frequency bins (i.e., low pass filter the periodogram). More details on windowing and classic spectral estimation computation may be found in reference [4]. The two methods of periodogram averaging may not be possible to perform in situations where only short data sequences are available.

In situations that involve a time series process with time-varying characteristics that can only be considered relatively constant for only short time intervals (a short time frame), the short-time Fourier transform (STFT) is often used as a time-frequency representation to describe such signals. Speech and sonar data are two signal types that involve time-varying characteristics. The STFT is defined as

$$X(nL, f) = \sum_{m=-\infty}^{\infty} x(m\Delta t) w[(nL-m)\Delta t] \exp(-j2\pi f m \Delta t) \quad (3.5)$$

where $w(n\Delta t)$ is a frame analysis window and L is an integer which denotes the separation in time between adjacent short-time frames. Thus, for a fixed n , $X(nL, f)$ represents the Fourier transform at time nL of windowed data samples. Taking the modulus of $X(nL, f)$, as in (3.3), at each L produces a short-time spectral estimate [see ref. 5].

The performance of classical spectral estimates at a given frequency f may be characterized by the stability-time-bandwidth product inequality

$$\Delta S \Delta T \Delta f > 1 \quad (3.6)$$

where ΔT is the time interval over which data has been measured, Δf is the resolution in Hertz, and ΔS is the stability factor, defined as the ratio of the spectral estimate variance over the spectral estimate mean. In order to have a stable spectral estimate for a fixed data set of ΔT seconds duration, ΔS must be made small. However, expression (3.6) indicates this can only be achieved by giving up resolution (accepting a larger value for Δf). Thus, spectral estimation involves a tradeoff between statistical stability and resolution.

The conventional Blackman-Tukey and periodogram approaches to spectral estimation have the following advantages: 1) computationally efficient if only a few lags are needed (BT) or if the FFT is used (periodogram), 2) PSD estimate directly proportional to the power for sinusoidal processes, and 3) a good model for some applications (the model is a sum of harmonically-related sinusoids). The disadvantages of these techniques are: 1) suppression of weak signal main-lobe responses by strong signal sidelobes, 2) frequency resolution limited by the available data record duration, independent of the characteristics of the data or its SNR, 3) introduction of distortion in the spectrum due to sidelobe leakage, 4) need for some sort of pseudo ensemble averaging to obtain statistically consistent periodogram spectra, and 5) the appearance of negative PSD values with the BT approach when some autocovariance sequence estimates are used.

4. PARAMETRIC METHODS

Often one has more knowledge about the process from which the data samples are taken, or at least is able to make a more reasonable assumption other than to assume the data is zero outside the window. Use of *a priori* information (or assumptions) may permit selection of an exact model for the process that generated the data samples, or at least a model that is a good approximation to the actual underlying process. It is then usually possible to obtain a better spectral estimate based on the model by determining the parameters of the model from the observations. The lack of a true model of the process under measurement does not mean one cannot try the parametric methods; use of these models may still yield reasonable results. One major motivation for the current interest in the modelling, or parametric, approach to spectral estimation is the higher frequency resolution achievable with these modern techniques over that achievable with the traditional techniques previously discussed. The degree of improvement in resolution and spectral "fidelity", if any, will be determined by the ability to fit an assumed model with a few parameters to the measured data.

Many deterministic and stochastic discrete-time processes encountered in practice are well approximated by a rational transfer function model. In this model, an input driving sequence (n_n) and the output sequence (x_n) that is related to the data are related by the linear difference equation,

$$x_n = \sum_{l=0}^p b_l n_{n-l} - \sum_{k=1}^q a_k x_{n-k} \quad (4.1)$$

This most general linear model is termed an autoregressive-moving average (ARMA) model. The interest in these models stems from their relationship to linear filters with rational transfer functions.

The system function $H(z)$ between the input n_n and output x_n for the ARMA process is the rational function

$$H(z) = \frac{B(z)}{A(z)} \quad (4.2)$$

where

$$A(z) = z - \text{transform of AR branch} = \sum_{m=0}^p a_m z^{-m} \quad (4.3)$$

$$B(z) = z - \text{transform of MA branch} = \sum_{m=0}^q b_m z^{-m}$$

The PSD of the output process of an ARMA filter driven by a white noise process is then

$$P_{ARMA}(f) = P_x(f) = \sigma^2 |B(f)/A(f)|^2 \quad (4.4)$$

where $A(f) = A(\exp[j2\pi f \Delta t])$ and $B(f) = B(\exp[j2\pi f \Delta t])$. Specification of the parameters (a_k) (termed the autoregressive coefficients), the parameters (b_k) (termed the moving-average coefficients), and σ^2 is equivalent to specifying the spectrum of the process (x_n). Without loss of generality, one can assume $a_0 = 1$ and $b_0 = 1$ since any filter gain can be incorporated into σ^2 .

4.1. Autoregressive Spectral Estimation

If all the (b_k), except $b_0 = 1$, are zero, then

$$x_n = - \sum_{k=1}^p a_k x_{n-k} + n_n \quad (4.5)$$

and the process is strictly an autoregression of order p driven by white noise process n_n . The process is termed AR in that the sequence x_n is a linear regression on itself with n_n representing the error. With this model, the present value of the process is expressed as a weighted sum of past values plus a noise term of variance σ^2 . The PSD is

$$P_{AR}(f) = \frac{\sigma^2}{|A(f)|^2} = \frac{\sigma^2}{\left| 1 + \sum_{k=1}^p a_k \exp(-j2\pi f k \Delta t) \right|^2} \quad (4.6)$$

This model is sometimes termed an all-pole model due to fact that the only frequency dependence of the spectrum is in the denominator. To estimate the PSD one need only estimate ($a_1, a_2, \dots, a_p, \sigma^2$).

Known autocovariance case

A relationship between the AR parameters and the autocovariance function can be developed (see the chapter by Kailath for details of this development). This relationship is known as the Yule-Walker normal equations. In matrix form, they are compactly expressed as

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & \dots & R_{xx}(-p) \\ R_{xx}(1) & R_{xx}(0) & \dots & R_{xx}(-p+1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(p) & R_{xx}(p-1) & \dots & R_{xx}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (4.7)$$

To determine the AR parameters and σ^2 , one must then solve (4.7) with the $p+1$ estimated autocovariance lags $R_{xx}(0), \dots, R_{xx}(p)$ and use the fact that $R_{xx}(-m) = R_{xx}^*(m)$. A computationally efficient algorithm known as the Levinson recursion can solve (4.7) with order p^2 operations (Zohar, [6]).

An alternative representation of (4.6) is

$$P_{AR}(f) = \sum_{n=-\infty}^{\infty} r_{xx}(n) \exp(-j2\pi f n \Delta t) \quad (4.8)$$

where

$$r_{xx}(n) = \begin{cases} R_{xx}(n), & \text{for } |n| \leq p \\ - \sum_{k=1}^p a_k r_{xx}(n-k), & \text{for } |n| > p \end{cases} \quad (4.9)$$

Thus, the autoregressive model produces one of many possible extensions of the autocovariance sequence that one might derive from the available autocovariance lag values. From this, it is easy to see that the AR PSD preserves the known lags and recursively extends the lags beyond the window of known lags. The AR PSD

output of a p pole and q zero filter excited by white noise, i.e.,

$$x_n = - \sum_{k=1}^p a_k x_{n-k} + \sum_{k=0}^q b_k w_{n-k} \quad (4.24)$$

where $R_{xx}(k) = \sigma^2$ for $k=0$, else it is zero, and $b_0 = 1$. The poles of the filter are assumed to be within the unit circle of the z -plane. The zeros of the filter may lie anywhere in the z -plane.

Once the parameters of the ARMA(p,q) model are identified, the spectral estimate is obtained as

$$P_x(f) = \frac{1}{H} \frac{|\exp(j2\pi f)|^2 P_w(f)}{\left| 1 + \sum_{k=1}^p a_k \exp(-j2\pi f k \Delta t) \right|^2} \quad (4.25)$$

Many ARMA parameter estimation techniques have been formulated theoretically, which usually involve many matrix computations and/or iterative optimization techniques. These approaches are normally not practical for realtime processing. Suboptimum techniques have therefore been developed to make the computational load more manageable. These techniques are usually based on a least squares criterion and require solutions of linear equations. These methods generally estimate the AR and MA parameters separately rather than jointly as required for optimal parameter estimation. The AR parameters can first be estimated independently of the MA parameters if one uses the so-called high-order (or modified) Yule-Walker equations. A final point in favor of the suboptimal linear approaches is that iterative optimization techniques are not guaranteed to converge or may converge to the wrong solution.

Assuming the autocovariance lags of a process known to be an ARMA(p,q) process are available, then the extended Yule-Walker equations that yield the AR parameters only is given by

$$\begin{bmatrix} R_{xx}(q) & R_{xx}(q-1) & \dots & R_{xx}(q-p+1) \\ R_{xx}(q+1) & R_{xx}(q) & \dots & R_{xx}(q-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(q+p-1) & R_{xx}(q+p-2) & \dots & R_{xx}(q) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_{xx}(q+1) \\ R_{xx}(q+2) \\ \vdots \\ R_{xx}(q+p) \end{bmatrix} \quad (4.26)$$

A fast algorithm for solving (4.26) has been developed by Zohar [6]. Once the AR parameter estimates have been found, the MA parameters may be found by filtering the data with the all-zero filter $A(z)$, where

$$A(z) = 1 + \sum_{k=1}^q a_k z^{-k}$$

to yield a purely MA process. The usual techniques to find the MA parameters may then be used (see reference [1] for details).

A second technique for estimating the ARMA parameters utilizes the identity

$$\frac{B(z)}{A(z)} = \frac{1}{C(z)}$$

where

$$C(z) = 1 + \sum_{k=1}^p c_k z^{-k}$$

to equate an ARMA model to an infinite order AR model. The $\{c_k\}$ may thus be estimated using a high order AR fit with AR estimation algorithms only, and then relating them as above to the ARMA parameters [7].

A third technique based upon least squares input-output identification has been proposed by many researchers. The normal equations exhibit a nonlinear character due to the unknown cross covariance between the input and output. If n_n is unobservable, the $R_{nx}(k)$ cannot be estimated. If, however, n_n were known or could be roughly estimated, so that $R_{nx}(k)$ could be estimated, then the ARMA parameters could be found as the solution of a set of linear equations. In practice, n_n is estimated from x_n in a boot-strap approach, for example, with a lattice filter configuration [8].

4.3 Prony's Method

Prony's method, a technique for modeling data of equally spaced samples by a linear combination of exponentials, is not a spectral estimation technique in the usual sense, but a spectral interpretation is provided in this section. The original

procedure by Baron de Prony exactly fitted an exponential curve having p exponential terms (each term with two parameters — an amplitude A_i and an exponent α_i where $A_i \exp[\alpha_i t]$) to $2p$ data measurements. For the case where only an *approximate* fit with p exponentials to a data set of N samples is desired, such that $N > 2p$, a least squares estimation procedure is used. This procedure is called the extended Prony method.

The model assumed in the extended Prony method is a set of p exponentials of arbitrary amplitude, phase, frequency, and damping factor. The discrete-time function

$$\hat{x}_n = \sum_{m=1}^p b_m z_m^n, \quad \text{for } n=0, \dots, N-1 \quad (4.27)$$

is the model to be used for approximating the measured data x_0, \dots, x_{N-1} . For generality, b_m and z_m are assumed complex and

$$b_m = A_m \exp(j\theta_m) \quad (4.28)$$

$$z_m = \exp[(\alpha_m + j2\pi f_m)\Delta t]$$

where A_m is the amplitude, θ_m is the phase in radians, α_m is a damping factor, f_m is the oscillation frequency in Hertz, and Δt represents the sample interval in seconds. Finding $\{A_m, \theta_m, \alpha_m, f_m\}$ and p that minimize the squared error

$$E = \sum_{n=0}^{N-1} |x_n - \hat{x}_n|^2 \quad (4.29)$$

is a difficult nonlinear least squares problem. An alternative suboptimum solution that does not minimize (4.29), yet still provides satisfactory results, is based on Prony's technique. Prony's method solves two sequential sets of linear equations with an intermediate polynomial rooting step that concentrates the nonlinear aspects of the problem.

The key to the Prony technique is to recognize that (4.27) is the homogeneous solution to a constant coefficient linear difference equation, the form of which is found as follows. Define the polynomial $\psi(z)$ as

$$\Psi(z) = \prod_{k=1}^p (z - z_k) = \sum_{i=0}^p a_i z^{p-i}, \quad a_0 = 1 \quad (4.30)$$

Thus $\psi(z)$ has the complex exponentials $\{z_k\}$ of (4.28) as its roots and complex coefficients $\{a_i\}$ when multiplied out. Based on (4.27), one way of expressing \hat{x}_{n-m} is

$$\hat{x}_{n-m} = \sum_{i=1}^p b_i z_i^{n-m} \quad (4.31)$$

for $0 \leq n-m \leq N-1$. Multiplying (4.31) by a_m and summing over the past $p+1$ products yields

$$\sum_{m=0}^p a_m \hat{x}_{n-m} = \sum_{i=1}^p b_i \sum_{m=0}^p a_m z_i^{n-m} \quad (4.32)$$

which is defined for $p \leq n \leq N-1$. If in (4.32) the substitution $z_j^{n-m} = z_j^n z_j^{-p} z_j^{p-m}$ is made, then

$$\sum_{m=0}^p a_m \hat{x}_{n-m} = \sum_{i=1}^p b_i z_i^{n-p} \sum_{m=0}^p a_m z_i^{p-m} = 0 \quad (4.33)$$

The zero result in (4.33) follows by recognizing that the final summation above is just the polynomial $\psi(z_i)$ of (4.30), evaluated at one of its roots. Expression (4.33) then yields the recursive difference equation

$$\hat{x}_n = - \sum_{m=1}^p a_m \hat{x}_{n-m} \quad (4.34)$$

defined for $p \leq n \leq N-1$. The approximation error $\epsilon_n = x_n - \hat{x}_n$ can be substituted into (4.34), and the moving average error ϵ_n defined as

$$\epsilon_n = - \sum_{m=0}^p a_m \epsilon_{n-m}, \quad \text{for } n=p, \dots, N-1 \quad (4.35)$$

so that

$$x_n = - \sum_{m=1}^p a_m x_{n-m} + \epsilon_n \quad (4.36)$$

The so-called extended Prony method is then seen to suboptimally minimize

$$\sum_{n=p}^{N-1} |\epsilon_n|^2, \quad \text{rather than the true optimum obtained by minimizing } \sum_{n=p}^{N-1} |\epsilon_n|^2.$$

Thus, the extended Prony parameter estimation procedure reduces to that of an AR/linear prediction parameter estimation, for which all the least squares approaches discussed in section 4.1, can be applied.

Once the z_i have been determined from the polynomial rooting, expression (4.27) reduces to a set of linear equations in the unknown b_m parameters, expressible in matrix form as

$$\Phi \mathbf{B} = \hat{\mathbf{X}} \quad (4.37)$$

where

$$\Phi = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_1 & z_2 & \dots & z_p \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{p-1} & z_2^{p-1} & \dots & z_p^{p-1} \end{bmatrix}$$

$$\mathbf{B} = [b_1 \dots b_p]^T$$

$$\hat{\mathbf{X}} = [\hat{x}_0 \dots \hat{x}_{N-1}]^T$$

Note that Φ is a Van der Monde matrix form. A least squares minimization of $\Sigma(x - \hat{x})^2$ yields the well-known solution

$$\mathbf{B} = [\Phi^H \Phi]^{-1} \Phi^H \mathbf{X} \quad (4.38)$$

Here, the H denotes the complex conjugate transpose operation. Determining the a_i parameters by a least squares estimation, rooting the polynomial, and solving for the b_j parameters (or residues) constitutes the entire extended Prony procedure.

It is possible to "define" a Prony spectrum by simply taking the modulus of the Fourier transform of (4.27) to yield

$$\hat{P}_{\text{PRONY}}(f) = |\hat{x}(f)|^2 \quad (4.39)$$

where

$$\hat{x}(f) = \sum_{m=1}^p A_m \exp(j\theta_m) \frac{2\alpha_m}{[\alpha_m^2 + (2\pi(f - f_m))^2]} \quad (4.40)$$

Other kinds of Prony "spectrums" could have been defined, but this seems to make the most sense as we shall see in the applications section.

5. NONPARAMETRIC METHODS

5.1. Maximum Likelihood Spectral Estimation

The maximum likelihood spectral estimate (MLSE) falls into the category of a nonparametric technique in the sense that no model parameters are explicitly computed. The original concept was developed by Capon for frequency-wavenumber analysis [9]. A filter model analogy will be used to describe this method. The MLSE was originally developed for seismic array frequency-wavenumber analysis. In this method, one estimates the PSD by effectively measuring the power out of a set of narrow-band filters. MLSE is actually a misnomer in that the spectral estimate is not necessarily a true maximum likelihood estimate of the PSD, it may more appropriately be termed the Capon spectral estimate after its inventor. The name MLSE will be retained here only for historic reasons. The difference between MLSE and conventional BT/periodogram spectral estimation is that the shape of the narrow-band filters in MLSE are, in general, different for each frequency, whereas they are fixed with the BT/periodogram procedures. The filters adapt to the process second order statistics for which a PSD estimate is sought. In particular, the filters are finite impulse response (FIR) types with p weights (taps).

$$\mathbf{A} = [a_0 \ a_1 \ \dots \ a_{p-1}]^T \quad (5.1)$$

The coefficients are chosen so that at a frequency under consideration, f_s , the frequency response of the filter is unity (i.e., an input sinusoid at that frequency would be undistorted at the filter output) and the variance of the output process is minimized. Thus the filter should adjust itself to reject components of the spectrum not near f_s so that the output power is due mainly to frequency components close to f_s . To obtain the filter, one minimizes the output variance σ^2 , given by

$$\sigma^2 = \mathbf{A}^H \mathbf{R}_{xx} \mathbf{A} \quad (5.2)$$

subject to the unity frequency response constraint (so that the sinusoid of frequency f_s is filtered without distortion),

$$\mathbf{E}^H \mathbf{A} = 1 \quad (5.3)$$

where \mathbf{R}_{xx} is the autocovariance matrix of x_n , and \mathbf{E} is the vector

$$\mathbf{E} = [1 \ \exp(j2\pi f_0 \Delta t) \ \dots \ \exp(j2\pi(p-1)f_0 \Delta t)]^T$$

The solution for the filter weights is easily shown to be

$$\mathbf{A}_{\text{OPT}} = \frac{\mathbf{R}_{yy}^{-1} \mathbf{E}}{\mathbf{E}^H \mathbf{R}_{yy}^{-1} \mathbf{E}} \quad (5.4)$$

and the minimum output variance is then

$$\sigma_{\text{MIN}}^2 = \frac{\Delta t}{\mathbf{E}^H \mathbf{R}_{yy}^{-1} \mathbf{E}} \quad (5.5)$$

It is seen that the frequency response of the optimum filter is unity at $f = f_0$ and that the filter characteristics change as a function of the underlying autocovariance

function. Since the minimum output variance is due to frequency components near f_0 , then $\sigma_{\text{MIN}}^2 \Delta t$ can be interpreted as a PSD estimate. Thus, the MLSE PSD is defined as

$$\hat{P}_{\text{ML}}(f_0) = \frac{1}{\mathbf{E}^H \mathbf{R}_{yy}^{-1} \mathbf{E}} \quad (5.6)$$

To compute the spectral estimate, one only needs an estimate of the autocovariance matrix.

The MLSE and AR PSD have been related analytically as follows (see Burg's paper in reference [3])

$$\frac{1}{\hat{P}_{\text{ML}}(f)} = \frac{1}{p} \sum_{m=1}^p \frac{1}{\hat{P}_{\text{AR}}^{(m)}(f)} \quad (5.7)$$

where $\hat{P}_{\text{AR}}^{(m)}(f)$ is the AR PSD for an m th order model and $\hat{P}_{\text{ML}}(f)$ is the MLSE PSD, both based upon a known autocovariance matrix of order p .

5.2. Pisarenko Harmonic Decomposition

In general, a $2p$ -th order difference equation of real coefficients of the form

$$x_n = - \sum_{m=1}^{2p} a_m x_{n-m} \quad (5.8)$$

can represent a deterministic process consisting of p real sinusoids of the form $\sin(2\pi f_i \Delta t)$. In this case, the $\{a_m\}$ are coefficients of the symmetric polynomial

$$z^{2p} + a_1 z^{2p-1} + \dots + a_{2p} z + 1 = 0 \quad (5.9)$$

$$\sum_{i=1}^p (z - z_i)(z - z_i^*) = 0$$

with unit modulus roots that occur in complex conjugate pairs of the form $z_i = \exp(j2\pi f_i \Delta t)$, where the f_i are arbitrary frequencies such that $|f_i| \leq 1/2\Delta t$, and $i = 1, \dots, p$. For sinusoids in additive white noise w_n , the observed process is

$$y_n = x_n + w_n = - \sum_{m=1}^{2p} a_m x_{n-m} + w_n \quad (5.10)$$

where $E(w_n w_{n+k}) = \sigma^2 \delta_k$, $E(w_n) = 0$, and $E(x_n w_n) = 0$ since the noise is assumed to be uncorrelated with the sinusoids. Substituting $x_{n-m} = y_n - w_{n-m}$ into (5.10), it is possible to express (5.10) as

$$\sum_{m=0}^{2p} a_m y_{n-m} = \sum_{m=0}^{2p} a_m w_{n-m} \quad (5.11)$$

where $a_0 = 1$ by definition. Expression (5.11) has the structure of an ARMA(p, p) model. However, this ARMA has a special symmetry in which the AR parameters are identical to the MA parameters in the model.

If the autocovariance function of y_n is known, the ARMA parameters can be found as the solution to an eigenequation, as is now shown. An equivalent matrix expression for (5.11) is

$$\mathbf{Y}^T \mathbf{A} = \mathbf{W}^T \mathbf{A} \quad (5.12)$$

where

$$\mathbf{Y}^T = [y_n \ y_{n-1} \ \dots \ y_{n-2p}]$$

$$\mathbf{A}^T = [1 \ a_1 \ \dots \ a_{2p-1} \ a_{2p}] \quad a_{2p-1} = a_1$$

$$\mathbf{W}^T = [w_n \ w_{n-1} \ \dots \ w_{n-2p}]$$

Premultiplying both sides of (5.12) by the vector \mathbf{Y} and taking the expectation yields

$$\mathbf{E}[\mathbf{Y} \mathbf{Y}^T] \mathbf{A} = \mathbf{E}[\mathbf{Y} \mathbf{W}^T] \mathbf{A} \quad (5.13)$$

Defining

$$\mathbf{X}^T = [x_n \ \dots \ x_{n-2p}]$$

then

$$\mathbf{E}[\mathbf{Y} \mathbf{Y}^T] = \mathbf{R}_{yy} = \begin{bmatrix} R_{yy}(0) & \dots & R_{yy}(-2p) \\ \vdots & \ddots & \vdots \\ R_{yy}(2p) & \dots & R_{yy}(0) \end{bmatrix} \quad (5.14)$$

$$\mathbf{E}[\mathbf{Y} \mathbf{W}^T] = \mathbf{E}[(\mathbf{X} + \mathbf{W}) \mathbf{W}^T] = \mathbf{E}[\mathbf{W} \mathbf{W}^T] = \sigma^2 \mathbf{I}$$

Here \mathbf{R}_{yy} is the Toeplitz autocovariance matrix for the observed process and \mathbf{I} is the identity matrix. The fact that $\mathbf{E}[\mathbf{X} \mathbf{W}^T] = 0$ follows from the assumption that the sinusoids are uncorrelated with the noise. Expression (5.13) is then rewritten as

$$\mathbf{R}_{yy} \mathbf{A} = \sigma^2 \mathbf{A} \quad (5.15)$$

which is an eigenequation where the noise variance (σ^2) is an eigenvalue of the autocovariance matrix \mathbf{R}_{yy} . The ARMA parameter vector \mathbf{A} is the eigenvector

associated with this eigenvalue that has been scaled so that the first and last elements are unity (A is symmetric). Equation (5.16) will yield the ARMA parameters of the Pisarenko technique. It turns out [1] that the noise variance is the smallest eigenvalue when the process is sinusoids in white noise and the autocovariance is exactly known.

Once the eigenvector A has been determined from the solution to (5.16), it remains to find the frequencies and powers to complete the harmonic decomposition (actually, nonharmonic). The frequencies are obtained by rooting the polynomial (5.9). Once the frequencies are known, the sinusoidal powers are obtained from the autocovariance lags, since

$$R_{yy}(k) = \sum_{i=1}^p P_i \cos(2\pi f_i k \Delta t) \quad (5.16)$$

for $k \neq 0$. A set of linear equations using p of the lags can then be used to find the sinusoid powers (P_i).

5.3 MUSIC Technique

A closely related eigenvector-eigenvalue decomposition of the autocovariance matrix to produce a spectral estimate is the Multiple Signal Classification (MUSIC) method [10]. If m represents the number of narrowband signal components, form a matrix S_p of the eigenvectors formed from the minimum $(p-m)$ eigenvalues of the $p \times p$ autocovariance matrix R_{xx} . The MUSIC spectral estimate is then simply

$$\hat{P}_{MUSIC}(f) = \frac{1}{E^H S_p S_p^H E} \quad (5.17)$$

where E was defined previously for the MLM spectral estimate.

This technique has also been used for high resolution beamforming. Using a uniform linear array of sensors, time samples are replaced by space samples. If the substitution

$$2\pi f = 2\pi(d/\lambda) \sin \theta$$

is made for frequency f , where d is the linear array sensor spacing and λ is the signal wavelength, then (5.17) may be used for high resolution beamforming. This is a time-space duality between uniform time samples and uniform linear array space samples.

6. MULTIDIMENSIONAL MEM [11]

The maximum entropy method (MEM) has generated an enormous amount of activity in the field of multidimensional high resolution spectral estimation. Unlike the one-dimensional case where MEM and AR were equivalent, in the M-D case the true maximum entropy estimate is distinctly different from the spectrum derived by AR modeling. In fact, the computation of the MEM spectral estimate appears to require the solution of a non-linear optimization problem. Recent research has been directed at the problem of computing the true MEM estimate. The existing approaches either attack the non-linear problem with a general optimization algorithm or assume an approximation to simplify the calculations. While the optimum MEM spectrum was shown to be the inverse of a positive multivariate polynomial, this polynomial may not be factorable as the magnitude-squared of a finite order polynomial. Since a spectrum based on a multidimensional AR model will always be of the form $|A(k)|^{-2}$, the MEM is, in the M-D case, more general than the AR spectral estimate. Note, however, the M-D AR estimator will usually not satisfy the correlation matching property (expression (4.9) in one dimension), whereas the true M-D MEM spectrum will. On the other hand, experiments with the M-D AR estimate show that it does have potential as a high-resolution M-D spectral estimator.

7. SUMMARY OF METHODS

Figure 1 illustrates typical spectra of the spectrum estimation methods discussed in this paper. Each spectral estimate is based on the same 64-point real sample

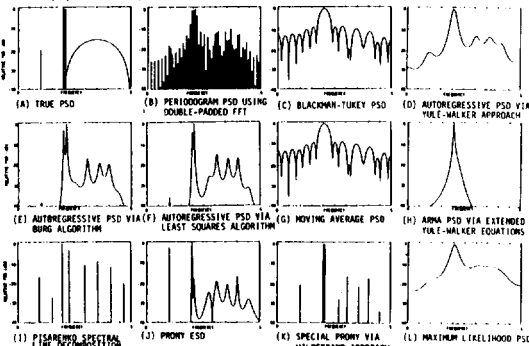


Fig. 1 Examples of various spectral estimates for the same 64-point sample sequence (from reference [1]).

sequence from a process consisting of three sinusoids and a colored noise process, obtained by filtering a white Gaussian noise process. Two of the sinusoids are very close together at .2 and .21; these are frequency values expressed as fractions of the sampling frequency (i.e., fractional frequencies range between 0, and .5). A weak sinusoid may be seen at fractional frequency .1. The colored noise process has a true spectrum of a raised cosine, as illustrated in Figure 1a. The spectral plots in Figure 1 are not intended to demonstrate relative performance of the various methods, but only to point out features of each technique.

8. COMMENTARY

Most of the effort of the spectral estimation research community to date has been directed toward new estimation algorithms, with little effort on characterizing the performance robustness of existing methods. Very often the only test case used is the two sinusoids in white noise test sequence, where the sinusoids are of equal amplitude and located in the center of the Nyquist bandwidth. Furthermore, usually only a limited number of test runs are made; running a large ensemble of test sequences is rare. When a large ensemble of sequences are run, typically only a single test signal class is selected. Thus, claims of performance in the published literature either have no meaningful statistical basis, or provide no indication of performance robustness over a wider class of signals, or both.

A thorough treatment of the methodology behind the spectral estimation techniques discussed in this article may be found in references [14] and [15].

9. CURRENT RESEARCH DIRECTIONS IN SPECTRAL ESTIMATION

Problems of current interest to spectral estimation researchers include fast algorithms and efficient algorithmic structures for real-time hardware implementation of high resolution spectral estimation techniques, improved ARMA modeling techniques, and better eigenanalysis approaches for spectral estimation and linear array beamforming. High resolution multidimensional spectral estimation continues to hold much research appeal, although it is primarily of academic rather than practical interest due to the complex formulations and solutions involved.

A topic of current research is that of bispectrum and trispectrum estimation [16],[17]. These techniques depend on third and fourth order statistical information for extracting information due to deviations from gaussian assumptions, for estimating the phase of parametric signals, and for detection and characterization of the properties of nonlinear mechanisms that generate time series.

REFERENCES

- [1] Kay, S.M. and Marple, S.L. Jr., "Spectrum Analysis - A Modern Perspective", *Proceedings of the IEEE*, vol. 69, pp. 1380-1419, November 1981.
- [2] Jenkins, G.M. and Watts, D.G., *Spectral Analysis and Its Applications*, Holden-Day, Inc. 1968.
- [3] Childers, D.G., editor, *Modern Spectrum Analysis*, IEEE Press, 1978.
- [4] Geckinli, N.C. and Yavuz, D., *Discrete Fourier Transformation and Its Applications to Power Spectral Estimation*, Elsevier Scientific Publishing Company, Amsterdam, 1983.
- [5] Rabiner, L.R. and Gold, B., *Theory and Applications of Digital Signal Processing*, Prentice-Hall, Inc., 1975.
- [6] Zohar, S., "FORTRAN Subroutines for the Solution of Toeplitz Sets of Linear Equations", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, pp. 656-658, December 1979.
- [7] Cadzow, J.A., "ARMA Modeling of Time Series", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, pp. 124-128, March 1982.
- [8] Friedlander, B., "Efficient Algorithms for ARMA Spectral Estimation", *IEEE Proc.*, vol. 130, Pt.F, pp. 195-201, April 1983.
- [9] Capon, J., "High Resolution Frequency-Wavenumber Spectrum Analysis", *Proc. IEEE*, vol. 57, pp. 1408-1418, August 1969.
- [10] Schmidt, R., "Multiple Emitter Location and Signal Parameter Estimation", *RADC Spectrum Estimation Workshop Record*, pp. 243-258, October 1979.
- [11] McClellan, J.H., "Multidimensional Spectral Estimation", *Proc. IEEE*, vol. 70, pp. 1029-1039, September 1982.
- [12] A.H. Nuttall, "Spectral analysis of a univariate process with bad data points, via maximum entropy and linear predictive techniques", NUSC Technical Report 5303, 26 March 1976.
- [13] S.L. Marple Jr. and A.H. Nuttall, "Experimental comparison of three multichannel linear prediction spectral estimators", *IEEE Proc.*, vol. 130, pt. F, pp 218-229, April 1983.
- [14] Marple, S.L. Jr., *Digital Spectral Analysis With Applications*, Prentice-Hall, Inc., 1987.
- [15] Kay, S. M., *Modern Spectral Estimation*, Prentice-Hall, 1988.
- [16] Nikias, C. L. and Raghuvver, M.R., "Bispectrum Estimation: A Digital Signal Processing Framework", *Proc. IEEE*, vol.75, pp.869-891, July 1987.
- [17] Raghuvver, M. and Chrysostomos, C., "Bispectrum Estimation: A Parametric Approach", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp.1213-1230, October 1985.