2

Spectral Estimation

Steven Kay
University of Rhode Island

2.0 INTRODUCTION

The problem of spectral estimation is that of determining the distribution in frequency of the power of a random process. Questions such as "Does most of the power of the signal reside at low or high frequencies?" or "Are there resonances in the spectrum?" are often answered as a result of a spectral analysis. As one might expect, spectral analysis finds wide use in such diverse fields as radar, sonar, speech, biomedicine, economics, geophysics, and others in which signals of unknown or questionable origin are of interest.

Estimation of the power spectral density is usually complicated by the lack of a sufficiently long duration data record on which to base the spectral estimate. The shortness of the record may be due to a genuine lack of data, as in the seismic patterns of an erupting volcano, or due to an artificially imposed restriction necessary to ensure that the spectral characteristics of a signal do not change over the duration of the data record, as in speech processing. For reliable spectral estimates we would wish for large amounts of data, but for many practical cases of interest, such as the two just mentioned, the data set is limited. The principles of statistical inference allow us to make the most of the available data. Tradeoffs can be expected in spectral estimation; the paramount one is bias versus variance. As we will see, if the spectral estimator yields good estimates on the average (low bias), then we can expect much variability from one data realization to the next (high variance). On the other hand, if we choose a spectral estimator with low variability, then on the average the spectral estimate may be poor. The only way out of this dilemma is to increase the data record length.

Spectral estimators may be classified as either nonparametric or parametric. The nonparametric ones such as the periodogram, Blackman-Tukey, and minimum variance spectral estimators require no assumptions about the data other than wide-sense

Steven Kay is with the Electrical Engineering Department, University of Rhode Island, Kingston, RI 02881.

stationarity. The parametric spectral estimators, on the other hand, are based on rational transfer function or time series models of the data. Hence, their application is more restrictive. The time series models are the autoregressive, moving average, and autoregressive moving average types. The advantage of the parametric spectral estimator is that when applicable it yields a more accurate spectral estimate. Without having to increase the data record length, we can simultaneously reduce the bias and the variance over the nonparametric spectral estimator. Of course, the improvement is due to the use of *a priori* knowledge afforded by the modeling assumption. Such adjectives as *high resolution* [1], *maximum entropy* [2], and *linear prediction* [3] are all synonymous with *parametric*.

2.1 DEFINITIONS

It will be assumed that $x(n)$ is a real discrete-time random process that is wide-sense stationary. To be wide-sense stationary, the mean of $x(n)$ for any n must be the same and the autocorrelation function must depend only on the lag between samples. Mathematically, it is assumed that

$$E[x(n)] = m_x \tag{2.1}$$

$$E[x(n)x(n+k)] = r_x(k) \tag{2.2}$$

where m_x is the mean and $r_x(k)$ is the autocorrelation function evaluated at lag k, both of which are independent of n. For convenience we will further assume that $m_x = 0$. This assumption is not restrictive in that we may equivalently consider $y(n) = x(n) - m_x$ with the result that $m_y = 0$.

The power spectral density (PSD) of $x(n)$ is defined as

$$P_x(\omega) = \sum_{k=-\infty}^{\infty} r_x(k) \exp(-j\omega k) \tag{2.3}$$

over the range $-\pi \le \omega \le \pi$. Equation (2.3) is sometimes called the Wiener-Khinchin theorem. The interpretation of $P_x(\omega)d\omega$ is as the average power of $x(n)$, which resides in the frequency band from ω to $\omega + d\omega$. Although this interpretation may not be evident from Eq. (2.3), an alternative but equivalent definition of the PSD will more clearly illustrate this property. The PSD may also be defined as

$$P_x(\omega) = \lim_{M\to\infty} E\left[\frac{1}{2M+1}\left|\sum_{n=-M}^{M} x(n)\exp(-j\omega n)\right|^2\right] \tag{2.4}$$

Eq. (2.4) says that the PSD at frequency ω is found by first taking the magnitude squared of the Fourier transform of $x(n)$ and then dividing by the data record length to yield power. Since the power will be a random variable (a different value will be obtained for each realization of $x(n)$), the expected value is taken. Finally, since the random process is in general of infinite duration, a limiting operation is required.

Eq. (2.4) is identical to Eq. (2.3) as we will now show. Starting with Eq. (2.4), we have

$$\begin{aligned} P_x(\omega) &= \lim_{M\to\infty} E\left[\frac{1}{2M+1}\sum_{n=-M}^{M}\sum_{m=-M}^{M} x(n)x(m)\exp[-j\omega(n-m)]\right] \\ &= \lim_{M\to\infty} \frac{1}{2M+1}\sum_{n=-M}^{M}\sum_{m=-M}^{M} r_x(n-m)\exp[-j\omega(n-m)] \end{aligned} \tag{2.5}$$

But

$$\sum_{n=-M}^{M}\sum_{m=-M}^{M} f(n-m) = \sum_{k=-2M}^{2M} (2M+1-|k|)f(k) \tag{2.6}$$

which may be verified by considering $f(n-m)$ as the (n, m) element of a matrix of dimension $(2M+1) \times (2M+1)$. Applying Eq. (2.6) to Eq. (2.5), we obtain

$$\begin{aligned} P_x(\omega) &= \lim_{M\to\infty} \frac{1}{2M+1}\sum_{k=-2M}^{2M} (2M+1-|k|)r_x(k)\exp(-j\omega k) \\ &= \lim_{M\to\infty} \sum_{k=-2M}^{2M}\left(1-\frac{|k|}{2M+1}\right)r_x(k)\exp(-j\omega k) \\ &= \sum_{k=-\infty}^{\infty} r_x(k)\exp(-j\omega k) \end{aligned}$$

The last step assumes that $r_x(k) \to 0$ as $k \to \infty$ at a sufficiently rapid rate, which is violated for random processes with nonzero means or sinusoidal components. Excluding these processes, Eq. (2.4) is equivalent to Eq. (2.3). Note that Eq. (2.3) is able to accommodate nonzero means and sinusoidal components but only via the use of Dirac delta functions. In practice, the data records will of necessity be finite and so this slight discrepancy is a moot point. Equations (2.3) and (2.4) will provide starting points for the class of nonparametric spectral estimators.

2.2 THE PROBLEM OF SPECTRAL ESTIMATION

The problem of PSD estimation or, more succinctly, spectral estimation is the following. Given a *finite* segment of a realization of a random process, i.e., $\{x(0), x(1), \ldots, x(N-1)\}$, estimate the PSD $P_x(\omega)$ for $|\omega| \le \pi$. Because $r_x(-k) = r_x(k)$ the PSD will be an even function, so we really need to estimate it only over the frequency interval $0 \le \omega \le \pi$. It is clear from the definition of the PSD given by Eq. (2.4) that the stated problem is difficult if not impossible. To find the PSD requires knowledge of $x(n)$ for all n as well as all possible realizations so that the expectation operation can be applied. Therefore, we should not expect perfect estimates. Another viewpoint is that, according to Eq. (2.3), spectral estimation is equivalent to autocorrelation estimation. Given N samples of $x(n)$ we must estimate an infinite number of autocorrelation lags, again an impossible task. One way out of this dilemma is to assume *a priori* that the autocorrelation function has certain properties. These properties would allow us to determine exactly the autocorrelation function for $k \geq p + 1$ if we knew the values for $k = 0, 1, \ldots, p$. (Recall that $r_x(-k) = r_x(k)$.) As an example, we might assume that

$$r_x(k) = -a_1 r_x(k-1) - \cdots - a_p r_x(k-p) \qquad \text{for } k \geq p + 1 \tag{2.7}$$

so that given the initial conditions of the difference equation $\{r_x(1), r_x(2), \ldots, r_x(p)\}$ and the coefficients $\{a_1, a_2, \ldots, a_p\}$ we could generate the remaining lags using Eq. (2.7). Add $r_x(0)$ to this information and we now have enough information to determine $P_x(\omega)$. The salient feature of this approach is that a model has been assumed for the autocorrelation function or, equivalently, for the PSD. The general spectral estimation problem in which we must estimate a continuous function has been reduced to a parameter estimation problem in which we estimate a finite set of parameters. For the example of Eq. (2.7) we will see in Section 2.5.2 that the a_k coefficients are easily determined from $r_x(k)$ for $0 \leq k \leq p$. Hence, knowledge of $r_x(k)$ for $0 \leq k \leq p$ is equivalent to knowledge of the PSD. Basic concepts in statistics tell us that if $N \gg p + 1$, our estimates of $r_x(k)$ for $0 \leq k \leq p$ and hence of the PSD will be good. This modeling approach forms the basis of the parametric techniques of Section 2.5.

A word of caution concerning the parametric techniques is in order. Since they rely on a model, it is critical that the model be correct. Any departure from the model will result in a systematic or bias error in the spectral estimate. As an example, if we model the PSD by a Gaussian function

$$P_x(\omega) = A \exp[-\tfrac{1}{2}(\omega - \omega_0)^2/\sigma^2], \qquad |\omega| \leq \pi$$

then we need only estimate $\{A, \omega_0, \sigma\}$ to estimate the PSD. If N is large, our estimates of these parameters presumably will not vary greatly from realization to realization. Hence, the spectral estimate will have low variability. If, however, the true PSD does not fit the model, the spectral estimator will be highly biased, an example of which is shown in Fig. 2.1. Clearly, the accuracy of the model is of utmost importance.

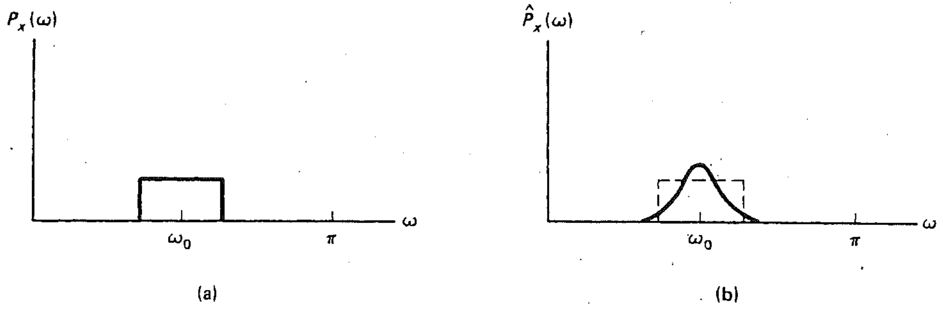


Figure 2.1 Example of difficulty of modeling approach using Gaussian model. (a) True power spectral density; (b) spectral estimate.

Since spectral estimators are random functions, it is necessary to base them on sound statistical techniques of estimation. Furthermore, we will need to describe the accuracy of a spectral estimate in statistical terms. To accomplish these goals, the next section briefly reviews the basic concepts of statistics employed in spectral estimation.

2.3 REVIEW OF STATISTICS

2.3.1 Properties of Estimators

The field of statistics deals with inferring information from random data. The theory of *statistical estimation* is particularly important for spectral estimation and hence we will restrict our review to this area. To illustrate the concepts, we will consider the problem of the estimation of the mean m_x of a wide-sense stationary Gaussian random process. We assume that the PSD $P_x(\omega)$ is white

$$P_x(\omega) = \sigma_x^2 + m_x^2 \delta_c(\omega), \qquad |\omega| \leq \pi \tag{2.8}$$

so that the variance is σ_x^2. Here $\delta_c(\omega)$ is the Dirac delta function. The observed data are $\{x(0), x(1), \ldots, x(N-1)\}$. We wish to find an estimator of m_x, which we will denote by $\hat{m}_x$. To be useful the estimator should be a function of only the observed data. A reasonable estimator might be the sample mean

$$\hat{m}_x = \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{2.9}$$

Note that $\hat{m}_x$ is a random variable since it is a function of several random variables. As such its value will change for each realization of the observed data. The *rule* of assigning a value to $\hat{m}_x$ is called the *estimator* of m_x, while the *actual value* of $\hat{m}_x$ computed for a realization of data is called the *estimate*. A desirable property of an estimator is that on the average it should yield the correct value or

$$E[\hat{m}_x] = m_x \tag{2.10}$$

The sample mean has this property since

$$E[\hat{m}_x] = E\left[\frac{1}{N}\sum_{n=0}^{N-1} x(n)\right] = \frac{1}{N}\sum_{n=0}^{N-1} E[x(n)] = m_x$$

Such an estimator is said to be *unbiased*. If this is not the case, then the bias of the estimator is defined to be

$$B[\hat{m}_x] = m_x - E[\hat{m}_x] \tag{2.11}$$

and is referred to as a *bias error*. An estimator may be unbiased but fluctuate wildly from realization to realization. For a reliable estimate we would also like the variance of the estimator to be small. We will denote the variance of an estimator by $\mathrm{VAR}[\hat{m}_x]$. The variance of the sample mean estimator is easily found. Noting that $x(n)$ is a white noise process, we see that samples of $x(n)$ are uncorrelated. This is easily verified by taking the inverse Fourier transform of $P_x(\omega)$, which is given by Eq. (2.8), to yield the autocorrelation function

$$r_x(k) = \sigma_x^2 \delta(k) + m_x^2$$

Since $r_x(k) - m_x^2 = 0$ for $m \neq 0$, the data samples are uncorrelated. It follows that

$$\mathrm{VAR}[\hat{m}_x] = \frac{1}{N^2}\sum_{n=0}^{N-1} \mathrm{VAR}[x(n)] = \frac{1}{N^2}(N\sigma_x^2) = \frac{1}{N}\sigma_x^2 \tag{2.12}$$

It is comforting that the variance approaches zero as the number of data samples tends to infinity. In a sense, then, the estimator is consistent in that

$$\lim_{N\to\infty} \hat{m}_x = m_x \tag{2.13}$$

and we should probably always require our estimators to have this property. Looking at it in another way we may rewrite Eq. (2.13) as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x(n) = E[x(m)] \tag{2.14}$$

The estimator $\hat{m}_x$ may be thought of as a temporal mean since we are averaging in time the samples of one realization, while $E[x(m)]$ represents an ensemble mean. The ensemble mean is computed by effectively averaging over many realizations at the identical instant of time. Since the temporal average as $N \to \infty$ equals the ensemble average, we refer to $x(n)$ as being *ergodic* in the mean. For arbitrary processes, ergodicity in the mean is attained if

$$\lim_{k\to\infty} r_x(k) = m_x^2 \tag{2.15}$$

Consistency of the sample mean estimator requires that the samples of the process be uncorrelated if separated by large enough lags. Otherwise, not enough temporal averaging will be accomplished to cancel out the random fluctuations of the samples.

It frequently occurs in practice that estimators are biased. Furthermore, the bias may depend on the parameter to be estimated. (If the bias did not depend on the unknown parameter, then in most cases we could remove the bias. As an example, consider the mean estimator $\sum_{n=0}^{N-1} x(n)$.) In such cases it is better to describe the performance of the estimator using the mean-square error (MSE)

$$\text{MSE}[\hat{m}_x] = E[(\hat{m}_x - m_x)^2] \tag{2.16}$$

In the case of unbiased estimators, we strive to use an estimator having a low variance, but here an estimator with a low MSE is desirable. The MSE is easily related to the variance since

$$\begin{aligned}\text{MSE}[\hat{m}_x] &= E[[(\hat{m}_x - E[\hat{m}_x]) + (E[\hat{m}_x] - m_x)]^2] \\ &= E[(\hat{m}_x - E[\hat{m}_x])^2 + 2(\hat{m}_x - E[\hat{m}_x])(E[\hat{m}_x] - m_x) + (E[\hat{m}_x] - m_x)^2] \\ &= \text{VAR}[\hat{m}_x] + 2[E[\hat{m}_x] - E[\hat{m}_x]][E[\hat{m}_x] - m_x] + B^2[\hat{m}_x]\end{aligned}$$

$$\text{MSE}[\hat{m}_x] = \text{VAR}[\hat{m}_x] + B^2[\hat{m}_x] \tag{2.17}$$

It is apparent from Eq. (2.17) that if an estimator is unbiased, then the MSE is identical to the variance. The question arises as to whether we can trade bias for variance in an attempt to yield a smaller MSE. For example, if $\hat{m}_x$ is an unbiased estimator and we define a new estimator as

$$\tilde{m}_x = a\hat{m}_x$$

then from Eq. (2.11),

$$B[\tilde{m}_x] = m_x - am_x = (1 - a)m_x \tag{2.18}$$

and also

$$\mathrm{VAR}[\check{m}_x] = a^2\,\mathrm{VAR}[\hat{m}_x] \tag{2.19}$$

Clearly, for $|a| < 1$ the variance is decreased but the bias is increased over that for $\hat{m}_x$. Also, from Eqs. (2.17)–(2.19),

$$\mathrm{MSE}[\check{m}_x] = a^2\,\mathrm{VAR}[\hat{m}_x] + (1 - a)^2\, m_x^2 \tag{2.20}$$

which is minimized over a by the optimal choice

$$a_{\mathrm{OPT}} = \frac{m_x^2}{m_x^2 - \mathrm{VAR}[\hat{m}_x]}. \tag{2.21}$$

If $\hat{m}_x$ is the sample mean estimator, then

$$a_{\mathrm{OPT}} = \frac{m_x^2}{m_x^2 - \sigma_x^2/N}$$

Unfortunately, the optimal value of a depends on m_x, which is exactly what we are attempting to estimate. It is interesting to note that if we knew *a priori* that m_x was close to zero, we could choose a close to zero to reduce the variance but still not increase the bias significantly. This type of operation is termed a *bias-variance tradeoff*, and we frequently encounter it in spectral estimation.

The final performance measure of an estimator, which will prove useful, is the confidence interval. Instead of describing the performance of an estimator by its mean and variance, it is more complete to state the probability density function (PDF) of the estimator. For our example, $\hat{m}_x$ is Gaussian since it is a sum of jointly Gaussian random variables. Hence,

$$\hat{m}_x \sim N(m_x, \sigma_x^2/N) \tag{2.22}$$

where $\sim$ means "is distributed according to" and $N(\mu, \sigma^2)$ denotes a Gaussian (normal) distribution with mean μ and variance σ^2. The PDF of the estimator itself may be conveniently summarized by giving an interval on the real line where m_x will lie with high probability. For instance, we might say that m_x lies within the interval $\hat{m}_x \pm \Delta$ with probability 0.9. The interval $(\hat{m}_x - \Delta, \hat{m}_x + \Delta)$ is called a 90% confidence interval for m_x. We now derive this confidence interval.

First consider the probability statement

$$\mathrm{Prob}\left[-\alpha \le \frac{\hat{m}_x - m_x}{\sqrt{\sigma_x^2/N}} \le \alpha\right] = 0.9 \tag{2.23}$$

where "Prob" denotes probability. We must choose α so that Eq. (2.23) is true. According to Eq. (2.22),

$$\frac{\hat{m}_x - m_x}{\sqrt{\sigma_x^2/N}} \sim N(0, 1)$$

so $\alpha = 1.645$. Manipulating Eq. (2.23), we have

$$\begin{aligned} \mathrm{Prob}[-\alpha\sqrt{\sigma_x^2/N} - \hat{m}_x \le -m_x \le \alpha\sqrt{\sigma_x^2/N} - \hat{m}_x] &= 0.9 \\ \mathrm{Prob}[\hat{m}_x - \alpha\sqrt{\sigma_x^2/N} \le m_x \le \hat{m}_x + \alpha\sqrt{\sigma_x^2/N}] &= 0.9 \end{aligned} \tag{2.24}$$

The 90% confidence interval is $\hat{m}_x \pm 1.645\sigma_x/\sqrt{N}$. Although it appears from Eq. (2.24) that m_x will fall within the confidence interval with 90% probability, this is an incorrect interpretation because m_x is not a random variable. We should say that the *random* interval $\hat{m}_x \pm 1.645\sigma_x/\sqrt{N}$ will *cover* the true value of m_x with 90% probability. An illustration is given in Fig. 2.2 in which the random interval covers the true value of the mean 90% of the time.

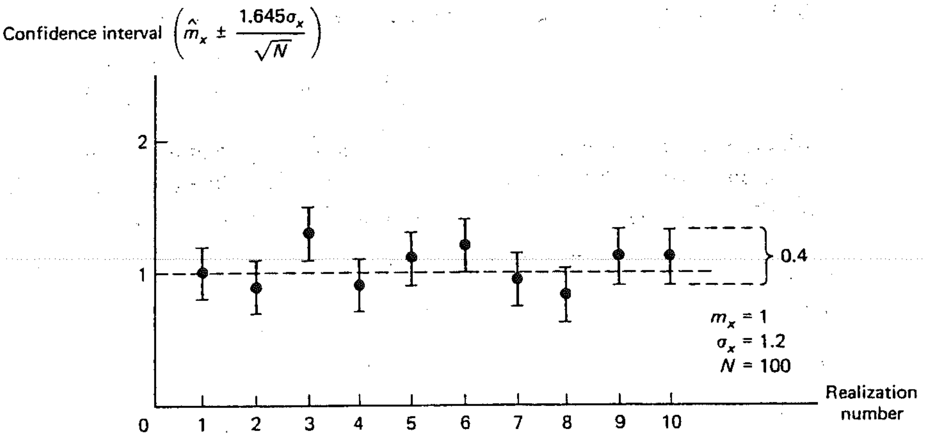


Figure 2.2 Illustration of concept of 90% confidence interval for mean.

2.3.2 Finding Good Estimators

Although we have discussed the properties of estimators and in particular examined the sample mean estimator, it is not yet clear how we actually find the estimator. The technique that has proven to be most useful in practice is the maximum likelihood estimator (MLE). The principle says that given the PDF $p(\mathbf{x}; \theta)$ of $\mathbf{x} = [x(0)x(1) \ldots x(N-1)]^T$, which depends on an unknown parameter θ, a good estimate of θ is found by choosing the value that maximizes $p(\mathbf{x}; \theta)$. The data sample values $\mathbf{x}_0$ are substituted for $\mathbf{x}$ so that the PDF is only a function of θ. The rationale for the approach is that by maximizing the PDF over θ we are finding the value of θ that results in the highest probability of $\mathbf{x}_0$ being observed. Since $\mathbf{x}_0$ was indeed observed, that value of θ is a reasonable one. Furthermore, it can be shown that as $N \rightarrow \infty$ the MLE is unbiased and has the smallest variance (and hence the smallest MSE) of all unbiased estimators [4]. As an example, consider the previous problem of mean estimation. To find the MLE we need to maximize $p(\mathbf{x}; m_x)$ over m_x. Since the data samples are jointly Gaussian and uncorrelated, they are also independent. Hence,

$$p(\mathbf{x}; m_x) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2\sigma_x^2}[x(n) - m_x]^2\right]$$

$$= \frac{1}{(2\pi\sigma_x^2)^{N/2}} \exp\left[-\frac{1}{2\sigma_x^2} S(m_x)\right]$$

where $S(m_x) = \sum_{n=0}^{N-1} [x(n) - m_x]^2$. Maximizing $p(\mathbf{x}; m_x)$ is equivalent to minimizing S. Hence,

$$\frac{\partial S}{\partial m_x} = -2 \sum_{n=0}^{N-1} [x(n) - m_x]$$

Setting the derivative equal to zero and replacing m_x by $\hat{m}_x$, we have

$$\sum_{n=0}^{N-1} [x(n) - \hat{m}_x] = 0 \qquad \text{or} \qquad \hat{m}_x = \frac{1}{N} \sum_{n=0}^{N-1} x(n)$$

which we have already seen to be an intuitively pleasing estimator. As a second example, consider the problem of estimating the linear trend in a time series, as illustrated in Fig. 2.3. A reasonable model for the data of Fig. 2.3 is

$$x(n) = \theta_1 n + \theta_2 + e(n), \qquad n = 0, 1, \ldots, N-1 \tag{2.25}$$

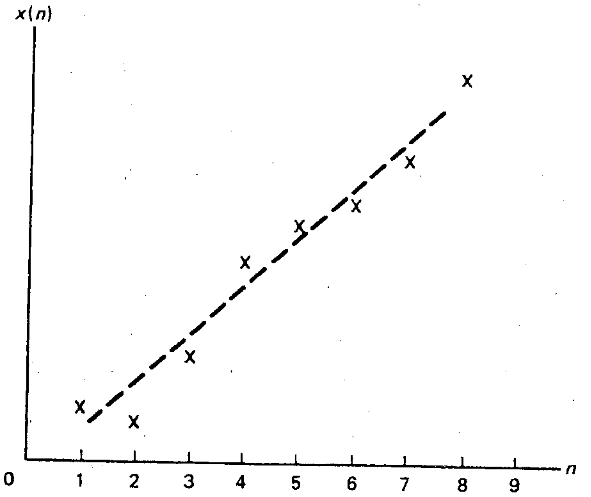


Figure 2.3 Data with apparent linear trend.

where θ_1, θ_2 are the slope and intercept, respectively, of a line. Both parameters of the line are unknown and $e(n)$ is modeled as a zero mean wide-sense stationary white Gaussian noise process with variance σ_e^2. The MLE of θ_1, θ_2 can be determined by maximizing $p(\mathbf{x}; \theta_1, \theta_2)$ over θ_1 and θ_2. In a similar fashion to the previous example we need to minimize

$$S'(\theta_1, \theta_2) = \sum_{n=0}^{N-1} [x(n) - \theta_1 n - \theta_2]^2 \tag{2.26}$$

The problem and solution are recast more conveniently in matrix notation. Let

$$\mathbf{x} = [x(0)\ x(1) \ldots x(N-1)]^T$$
$$\boldsymbol{\theta} = [\theta_1\ \theta_2]^T$$
$$\mathbf{e} = [e(0)\ e(1) \ldots e(N-1)]^T$$

$$\mathbf{H} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ N-1 & 1 \end{bmatrix}$$

Then Eq. (2.25) becomes

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e} \tag{2.27}$$

where $\mathbf{H}$ is a known matrix, $\boldsymbol{\theta}$ is a vector of unknown parameters, and $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$. $N(\boldsymbol{\mu}, \mathbf{K})$ denotes a multivariate Gaussian PDF with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$, and $\mathbf{I}$ is the identity matrix. Equation (2.26) now becomes

$$S'(\theta) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \tag{2.28}$$

To find the MLE, we take the gradient of S' by making use of the following identities:

$$\frac{\partial \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta} \qquad \text{if } \mathbf{A}^T = \mathbf{A}$$

$$\frac{\partial \boldsymbol{\theta}^T\mathbf{b}}{\partial \boldsymbol{\theta}} = \mathbf{b}$$

Hence,

$$S'(\boldsymbol{\theta}) = \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}$$

$$\frac{\partial S'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = \mathbf{0}$$

which results in

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{2.29}$$

$\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ under the conditions stated above.

It frequently occurs in practice that an estimation problem can be expressed in the form of Eq. (2.27) but that $\mathbf{e}$ is not composed of uncorrelated zero-mean Gaussian random variables. The least-squares modified Yule-Walker equation estimator of Section 2.5.8 is one such example. In such a case $\hat{\boldsymbol{\theta}}$ can still be used as an estimator of $\boldsymbol{\theta}$ although there are no optimality properties associated with the estimator; $\hat{\boldsymbol{\theta}}$ is called the *least-squares estimator* since it minimizes a sum of squares as given by Eq. (2.28). That is, it picks the $\hat{\boldsymbol{\theta}}$ such that the difference of the sum of the squares of the output produced by $\hat{\boldsymbol{\theta}}$ (i.e., $\mathbf{H}\hat{\boldsymbol{\theta}}$) and the output actually observed (i.e., $\mathbf{x}$) is minimized over all possible $\hat{\boldsymbol{\theta}}$.

2.3.3 A Useful PDF

The PDF for a sum of squares of independent $N(0, 1)$ random variables is of interest in spectral estimation. If $y = \Sigma_{n=0}^{N-1} x^2(n)$ where $x(n) \sim N(0, 1)$ for $n = 0, 1, \ldots, N-1$, and all the $x(n)$'s are independent, then the PDF of y is

$$p(y) = \begin{cases} \dfrac{1}{2^{N/2}\Gamma(N/2)} y^{N/2-1} \exp(-y/2) & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases} \tag{2.30}$$

where $\Gamma(u)$ is the gamma integral. The random variable y is said to be distributed according to a chi-squared distribution with N degrees of freedom or

$$y \sim \chi_N^2.$$

It is easily shown that the mean and variance are

$$E[y] = N \tag{2.31}$$

$$\text{VAR}[y] = 2N \tag{2.32}$$

2.4 NONPARAMETRIC SPECTRAL ESTIMATION

In this section we will discuss several popular methods of nonparametric spectral estimation. The Fourier or classical methods are the periodogram, which makes use of Eq. (2.4), and the Blackman-Tukey approach, which relies on Eq. (2.3). A recently proposed method termed the minimum variance spectral estimator will also be described. In all cases it will become apparent that for a fixed data record length we can reduce either the bias or the variance of the estimator but not both simultaneously.

2.4.1 Periodogram Spectral Estimator

The periodogram spectral estimator relies on Eq. (2.4):

$$P_x(\omega) = \lim_{M\to\infty} E\left[\frac{1}{2M+1}\left|\sum_{n=-M}^{M} x(n) \exp(-j\omega n)\right|^2\right] \tag{2.4}$$

Recall that the observed data set is $\{x(0), x(1), \ldots, x(N-1)\}$. By neglecting the expectation operator and by using the available data, we define the periodogram spectral estimator as

$$\hat{P}_{\text{PER}}(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n) \exp(-j\omega n)\right|^2 \tag{2.33}$$

An interesting interpretation of the periodogram estimator becomes apparent if we replace ω by ω_0 to emphasize that we are estimating the PSD at a particular frequency and rewrite Eq. (2.33) as

$$\hat{P}_{\text{PER}}(\omega_0) = \left[N\left|\sum_{k=0}^{N-1} h(n-k)x(k)\right|^2\right]\Bigg|_{n=0}$$

where

$$h(n) = \begin{cases} \dfrac{1}{N}\exp(j\omega_0 n) & \text{for } n = -(N-1), \ldots, -1, 0 \\ 0 & \text{otherwise} \end{cases}$$

and $h(n)$ is the impulse response of a linear shift-invariant filter with frequency response

$$\begin{aligned} H(\omega) &= \sum_{n=-(N-1)}^{0} h(n)\exp(-j\omega n) \\ &= \frac{\sin[N(\omega-\omega_0)/2]}{N\sin[(\omega-\omega_0)/2]}\exp\left[j\left(\frac{N-1}{2}\right)(\omega-\omega_0)\right] \end{aligned} \tag{2.34}$$

This is a bandpass filter with center frequency at $\omega = \omega_0$. Hence, the periodogram estimates the power at frequency ω_0 by filtering the data with a bandpass filter, sampling the output at $n = 0$, and computing the magnitude squared. The N factor is necessary to account for the bandwidth of the filter, which can be shown to be approximately $1/N$ [5], assuming a 3-dB bandwidth. The power when divided by $1/N$ yields the spectral estimate.

It might be supposed that if enough data were available, say $N \to \infty$, then

$$\hat{P}_{\text{PER}}(\omega) \to P_x(\omega)$$

or that the periodogram is a consistent estimator of the PSD. This was the case for estimation of the mean. To test this hypothesis we will consider the periodogram of zero-mean white Gaussian noise. In Fig. 2.4 we have computed the periodogram for several data record lengths. Each N-point data record was obtained by taking the first N points of a 1024-point data record. It appears that the random fluctuation or variance of the periodogram does not decrease with increasing N and hence that the periodogram is *not* a consistent estimator.

To verify this observation we will derive the PDF of the periodogram for white Gaussian noise. We will assume for simplicity that $\omega = \omega_k = k2\pi/N$ for $k = 0, 1, \ldots, N/2$ (N even). Note that $\hat{P}_{\text{PER}}(-\omega_k) = \hat{P}_{\text{PER}}(\omega_k)$, so we need not consider $k = -N/2 + 1, \ldots, -1$. The general results for an arbitrary frequency may be found in [6]. We show in the Appendix at the end of this chapter that

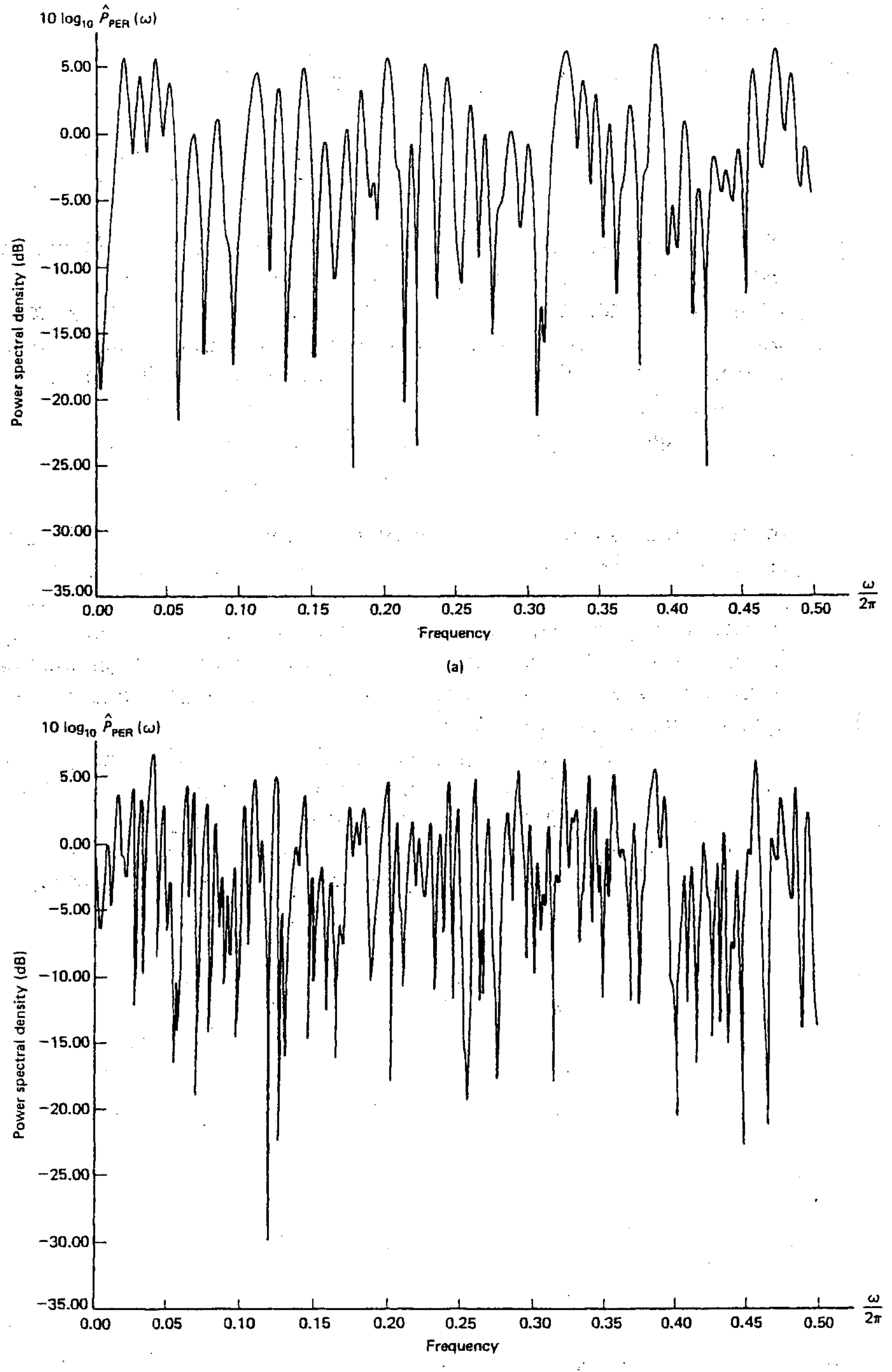
$$\begin{aligned} \frac{2\hat{P}_{\text{PER}}(\omega_k)}{P_x(\omega)} &\sim \chi_2^2, \qquad k = 1, 2, \ldots, N/2-1 \\ \frac{\hat{P}_{\text{PER}}(\omega_k)}{P_x(\omega)} &\sim \chi_1^2, \qquad k = 0, N/2 \end{aligned} \tag{2.35}$$

It immediately follows from Eqs. (2.31) and (2.32) that

$$E[\hat{P}_{\text{PER}}(\omega_k)] = P_x(\omega_k), \qquad k = 0, 1, \ldots, N/2 \tag{2.36}$$

$$\text{VAR}[\hat{P}_{\text{PER}}(\omega_k)] = \begin{cases} P_x^2(\omega_k), & k = 1, 2, \ldots, N/2-1 \\ 2P_x^2(\omega_k), & k = 0, N/2 \end{cases} \tag{2.37}$$

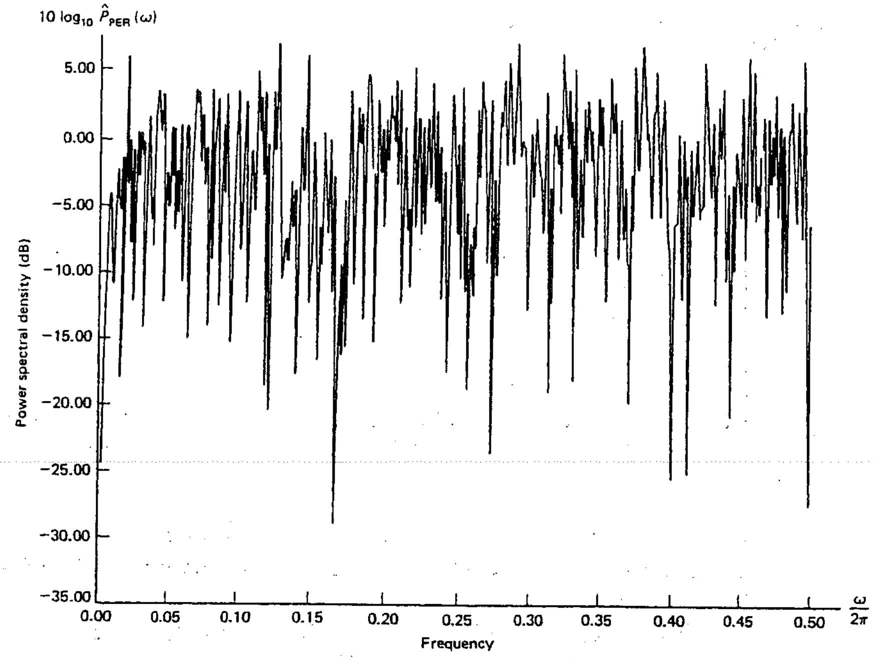
We see that the periodogram is an unbiased estimator of the PSD but that it is *not* consistent in that the variance does not decrease with increasing data record length. This accounts for the appearance of Fig. 2.4. The periodogram estimator is unreliable because the standard deviation, which is the square root of the variance, is as large



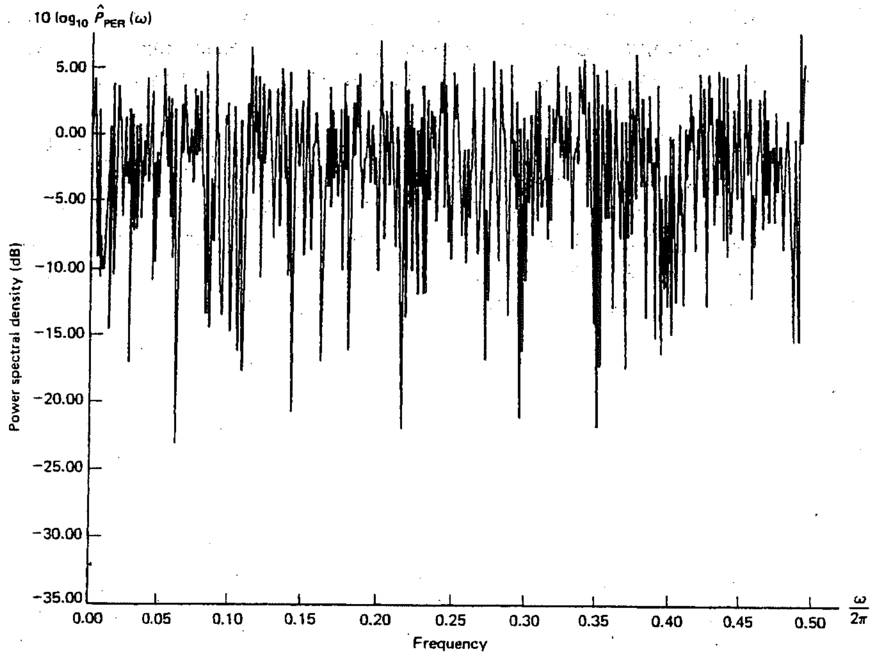
(a)

(b)

Figure 2.4 Illustration of the inconsistency of the periodogram for white Gaussian noise ($\sigma_x^2 = 1$). (a) $N = 128$, (b) $N = 256$, (c) $N = 512$, (d) $N = 1024$.



(c)



(d)

Figure 2.4 *(cont.)*

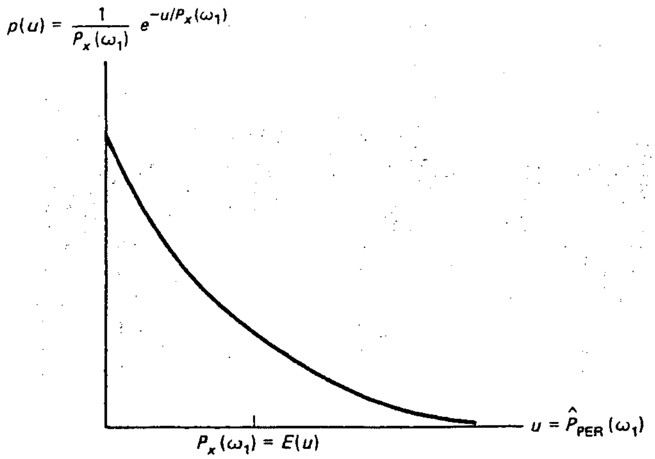


Figure 2.5 Probability density function of periodogram estimator.

as the mean, the quantity to be estimated. The PDF of $\hat{P}_{\text{PER}}(\omega_1)$ is given in Fig. 2.5 (see Section 2.3.3), which shows that the probability of obtaining an estimate near $P_x(\omega_1)$ is small.

The reason the periodogram is a poor estimator is that, given N data points, we are attempting to estimate about $N/2$ unknown parameters, i.e., $\{P_x(\omega_0), P_x(\omega_1), \ldots, P_x(\omega_{N/2})\}$. As N increases, our estimator does not improve because we are estimating proportionally more parameters. In the case of mean estimation the number of unknown parameters was fixed at one. Similar conclusions about the performance of the periodogram can be drawn for arbitrary frequencies and processes with arbitrary PSDs.

The way out of this dilemma is to use an averaged periodogram estimator in an attempt to approximate the expectation operator of Eq. (2.4). Assume we are given K data records uncorrelated with each other and all for the interval $0 \le n \le L - 1$. Also assume that they are drawn from the same random process. The data are $\{x_0(n), 0 \le n \le L-1; x_1(n), 0 \le n \le L-1; \ldots; x_{K-1}(n), 0 \le n \le L-1\}$. Then the averaged periodogram estimator is

$$\hat{P}_{\text{AVPER}}(\omega) = \frac{1}{K}\sum_{m=0}^{K-1} \hat{P}_{\text{PER}}^{(m)}(\omega) \tag{2.38}$$

where

$$\hat{P}_{\text{PER}}^{(m)}(\omega) = \frac{1}{L}\left|\sum_{n=0}^{L-1} x_m(n)\exp(-j\omega n)\right|^2 .$$

The expected value of the averaged periodogram for white Gaussian noise will be the true PSD as before, but the variance will be decreased by a factor of K, the number of periodograms averaged. Since the data records are uncorrelated and hence independent, the individual periodograms are independent and hence uncorrelated. It follows from Eq. (2.37) that for $k \neq 0, N/2$,

$$\begin{aligned}\text{VAR}[\hat{P}_{\text{AVPER}}(\omega_k)] &= \frac{1}{K^2}\sum_{m=0}^{K-1} \text{VAR}[\hat{P}^{(m)}_{\text{PER}}(\omega_k)] \\ &= \frac{1}{K^2} K P_x^2(\omega_k) \\ &= \frac{1}{K} P_x^2(\omega_k)\end{aligned} \tag{2.39}$$

As an example, consider the averaged periodogram estimate for white noise with $K = 8$ and $L = 128$ as shown in Fig. 2.6. A comparison with Fig. 2.4(a) illustrates the reduction in variance. In practice we seldom have uncorrelated data sets, but only one data record of length N on which to base the spectral estimator. A frequent approach is to segment the data into K nonoverlapping blocks of length L, where $N = KL$. In this manner we can use Eq. (2.38) with

$$x_m(n) = x(n + mL), \qquad n = 0, 1, \ldots, L-1; \quad m = 0, 1, \ldots, K-1$$

Since the blocks are contiguous, they cannot be uncorrelated for any other process but white noise. The variance reduction factor will in general be less than K. For processes not exhibiting sharp resonances, the autocorrelation function will damp out rapidly so that the use of Eq. (2.39) will be a good approximation.

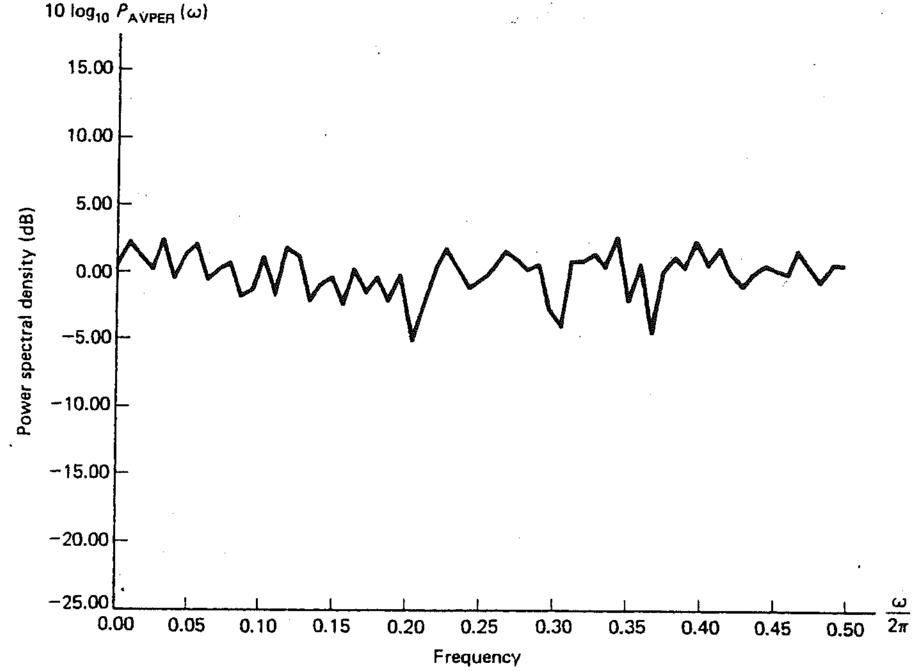


Figure 2.6 Averaged periodogram for white noise with $\sigma_x^2 = 1$, $L = 128$, and $K = 8$.

The astute reader may now ask why we could not just segment the data record into more and more subrecords in an effort to reduce the variance. The problem with this approach is that as the number of subrecords increases, the bias of the averaged periodogram estimator will increase for *any other process but white noise*. To see this we now derive the expected value of the averaged periodogram:

$$E[\hat{P}_{\text{AVPER}}(\omega)] = E\left[\frac{1}{K}\sum_{m=0}^{K-1} \hat{P}^{(m)}_{\text{PER}}(\omega)\right] = E[\hat{P}^{(0)}_{\text{PER}}(\omega)] \tag{2.40}$$

where we have used the fact that all the individual periodograms have the same PDF or are identically distributed. It can be shown, however, that the periodogram defined in Eq. (2.33) can also be written as

$$\hat{P}_{\text{PER}}(\omega) = \sum_{k=-(N-1)}^{N-1} \hat{r}_x(k)\exp(-j\omega k) \tag{2.41}$$

where

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=0}^{N-1-|k|} x(n)x(n+|k|) \tag{2.42}$$

and $\hat{r}_x(k)$ is observed to be an estimator of the autocorrelation function. Using Eq. (2.41) with N replaced by L in Eq. (2.40), we have

$$E[\hat{P}_{\text{AVPER}}(\omega)] = E\left[\sum_{k=-(L-1)}^{L-1} \hat{r}^{(0)}_x(k)\exp(-j\omega k)\right]$$

where

$$\hat{r}^{(0)}_x(k) = \frac{1}{L}\sum_{n=0}^{L-1-|k|} x(n)x(n+|k|) \tag{2.43}$$

so that

$$E[\hat{P}_{\text{AVPER}}(\omega)] = \sum_{k=-(L-1)}^{L-1} E[\hat{r}^{(0)}_x(k)]\exp(-j\omega k)$$

But from Eq. (2.43),

$$E[\hat{r}^{(0)}_x(k)] = \left(1 - \frac{|k|}{L}\right)r_x(k) \qquad \text{for } |k| \le L-1$$

so

$$E[\hat{P}_{\text{AVPER}}(\omega)] = \sum_{k=-(L-1)}^{L-1}\left(1 - \frac{|k|}{L}\right)r_x(k)\exp(-j\omega k)$$

If we define

$$w_B(k) = \begin{cases} 1 - |k|/L, & |k| \le L-1 \\ 0, & |k| > L \end{cases}$$

then we see that

$$E[\hat{P}_{\text{AVPER}}(\omega)] = \mathcal{F}[w_B(k)r_x(k)] = \int_{-\pi}^{\pi} W_B(\omega - \xi)P_x(\xi)\frac{d\xi}{2\pi} \tag{2.44}$$

where $\mathcal{F}$ denotes the Fourier transform operator and $W_B(\omega)$ is the Fourier transform of $w_B(k)$. The sequence $w_B(k)$ is sometimes referred to as a *triangular* or *Bartlett window*. Its Fourier transform is easily shown to be

$$W_B(\omega) = \frac{1}{L}\left(\frac{\sin \omega L/2}{\sin \omega/2}\right)^2 \tag{2.45}$$

and is plotted in Fig. 2.7. The result of the convolution operation is to produce an average spectral estimate that is smeared. An example is shown in Fig. 2.8. To avoid the smearing, the Bartlett window length L must be chosen so that the width of the main lobe of $W_B(\omega)$ is much less than the width of the narrowest peak in $P_x(\omega)$. Since the 3-dB bandwidth of the main lobe of $W_B(\omega)$ is about $\Delta\omega = 2\pi/L$, we cannot resolve details in the PSD finer than $2\pi/L$. The spectral estimator is then said to have a resolution of $1/L$ cycles/sample. Clearly, for maximum resolution we should choose L as large as possible or $L = N - 1$, which results in the standard periodogram. However, we know that for good variance reduction we should choose $K = N/L$ large according to Eq. (2.39) or L small. Since both goals cannot be met simultaneously, we are forced to trade off bias (or, equivalently, resolution) for variance by adjusting L. In practice, a good strategy is to compute several averaged periodogram spectral estimates, each successive one having a larger L. If the spectral

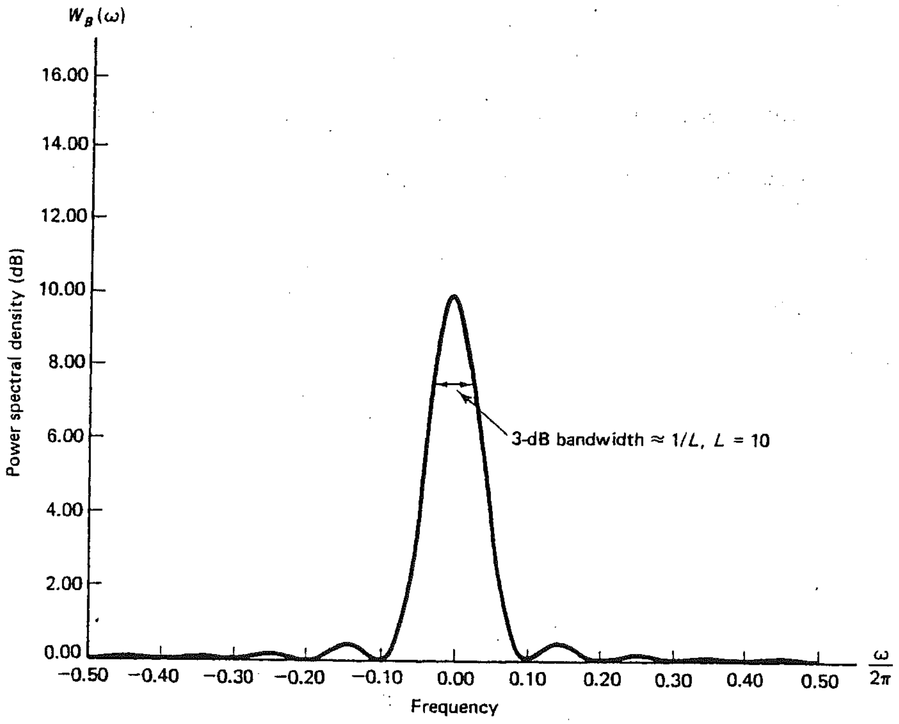


Figure 2.7 Fourier transform of Bartlett lag window.

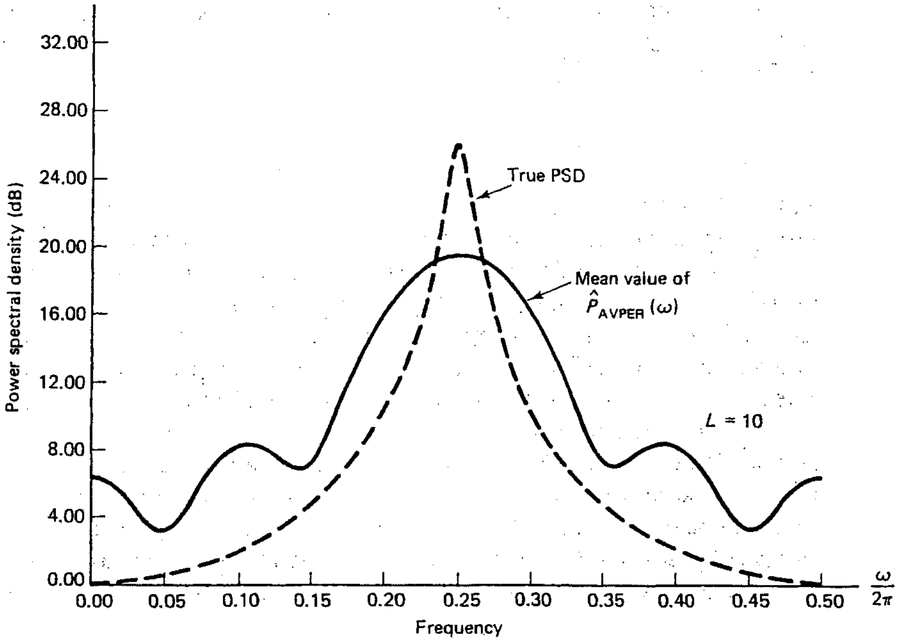


Figure 2.8 Example of mean value of averaged periodogram.

estimate does not change significantly as L is increased, then all the spectral detail has been found. This technique, known as *window closing* [6], may suffer from statistical instability problems. Another technique that is useful in practice is that of prewhitening the data prior to spectral estimation. Recall from Eq. (2.44) that the expected value of the averaged periodogram is a convolution of the Fourier transform of the Bartlett window with the PSD. If the process is white noise so that $P_x(\omega) = \sigma_x^2$, then

$$\begin{aligned} E[\hat{P}_{\text{AVPER}}(\omega)] &= \sigma_x^2 \int_{-\pi}^{\pi} W_B(\omega - \xi)\frac{d\xi}{2\pi} \\ &= \sigma_x^2 \int_{-\pi}^{\pi} W_B(\xi)\frac{d\xi}{2\pi} \\ &= \sigma_x^2 w_B(0) = \sigma_x^2 = P_x(\omega) \end{aligned} \tag{2.46}$$

Hence, regardless of the value of L, the estimator is unbiased, due of course to the lack of peaks and valleys in the PSD. If this is the case, then L can be made small to reduce the variance. In practice, if we have some idea of the general shape of the PSD but possibly not the details, a good technique is to filter the data with a linear shift-invariant filter to yield a PSD that is flatter at the output of the filter than at the input. To do so, we find the spectral estimate of the filter output and then divide it by $|H(\omega)|^2$, where $H(\omega)$ is the frequency response of the prewhitener. Such an approach is termed *prewhitening the data*.

It is extremely important in interpreting spectral estimates to be able to ascertain whether spectral detail is due to statistical fluctuation or is actually present. In other words, we need some measure of confidence in the spectral estimate. Assuming that the spectral estimator is approximately unbiased, we can derive a confidence interval for the estimator. (See Section 2.3.1 for a discussion of confidence intervals.) The unbiased assumption requires the bandwidth of the narrowest peak or valley of the PSD to be much larger than the bandwidth of the Bartlett spectral window, $W_B(\omega)$. If we recall the bandpass filtering interpretation of the periodogram, then at least within the vicinity of the frequency under consideration, we can replace the data by a white noise process. It is then possible to use the previous results to yield the approximation

$$\frac{2\hat{P}_{\text{PER}}(\omega)}{P_x(\omega)} \sim \chi_2^2 \tag{2.47}$$

(We omit $\omega = 0$ and $\omega = \pi$ from further consideration.) From Eq. (2.38),

$$\hat{P}_{\text{AVPER}}(\omega) = \frac{1}{K}\sum_{m=0}^{K-1} \hat{P}_{\text{PER}}^{(m)}(\omega)$$

so

$$\frac{2K\hat{P}_{\text{AVPER}}(\omega)}{P_x(\omega)} = \sum_{m=0}^{K-1} \frac{2\hat{P}_{\text{PER}}^{(m)}(\omega)}{P_x(\omega)}$$

But according to Eq. (2.47), each random variable in the summation is a χ_2^2 random variable or the sum of the squares of two independent $N(0,1)$ random variables. Furthermore, the blocks of data used to form each periodogram are approximately uncorrelated and hence independent due to the Gaussian nature of the data. Each random variable in the summation is then independent. The PDF for the sum of squares $2K\,N(0,1)$ random variables is χ_{2K}^2 or

$$\frac{2K\hat{P}_{\text{AVPER}}(\omega)}{P_x(\omega)} \sim \chi_{2K}^2 \tag{2.48}$$

We now define the α percentage point of a χ_{2K}^2 cumulative distribution function as

$$\text{Prob}\left[\chi_{2K}^2 \le \chi_{2K}^2(\alpha)\right] = \alpha$$

Then, from Eq. (2.48),

$$\text{Prob}\left[\chi_{2K}^2(\alpha/2) \le \frac{2K\hat{P}_{\text{AVPER}}(\omega)}{P_x(\omega)} \le \chi_{2K}^2(1-\alpha/2)\right] = 1-\alpha$$

$$\text{Prob}\left[\frac{2K\hat{P}_{\text{AVPER}}(\omega)}{\chi_{2K}^2(\alpha/2)} \ge P_x(\omega) \ge \frac{2K\hat{P}_{\text{AVPER}}(\omega)}{\chi_{2K}^2(1-\alpha/2)}\right] = 1-\alpha$$

so that a $(1-\alpha) \times 100\%$ confidence interval is

$$\left(\frac{2K\hat{P}_{\text{AVPER}}(\omega)}{\chi_{2K}^2(1-\alpha/2)}, \frac{2K\hat{P}_{\text{AVPER}}(\omega)}{\chi_{2K}^2(\alpha/2)}\right) \tag{2.49}$$

If we plot the PSD in decibels, the confidence interval becomes a constant-length interval for all ω, or

$$10 \log_{10} \hat{P}_{\text{AVPER}}(\omega) \begin{cases} +10 \log_{10} \dfrac{2K}{\chi^2_{2K}(\alpha/2)} \\ -10 \log_{10} \dfrac{\chi^2_{2K}(1-\alpha/2)}{2K} \end{cases} \text{dB} \tag{2.50}$$

As an example, for a 95% confidence interval with $K = 10$, $\alpha = 0.05$, $\chi^2_{20}(0.025) = 10.85$, $\chi_{20}(0.975) = 31.41$, and the confidence interval is

$$10 \log_{10} \hat{P}_{\text{AVPER}}(\omega) \begin{cases} +2.65 \\ -1.96 \end{cases} \text{dB}$$

This means that the interval

$$(10 \log_{10} \hat{P}_{\text{AVPER}}(\omega) - 1.96,\ 10 \log_{10} \hat{P}_{\text{AVPER}}(\omega) + 2.65)$$

will cover the true value of $10 \log_{10} P_x(\omega)$ with a probability of 0.95. Hence, spectral peaks and valleys of more than a few decibels should be considered as actually being present and not due to statistical fluctuation.

Before concluding our discussion of periodogram spectral estimators it is worthwhile to note the use of the fast Fourier transform (FFT) in computing them. Since we cannot expect to compute $\hat{P}_{\text{PER}}(\omega)$ for a continuum of frequencies, we are forced to sample it. Typically, one uses $\omega_k = 2\pi k/N$ for $k = 0, 1, \ldots, N-1$, so

$$\begin{aligned} \hat{P}_{\text{PER}}(\omega_k) &= \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n) \exp(-j\omega_k n)\right|^2 \\ &= \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n) \exp(-j2\pi kn/N)\right|^2 \end{aligned} \tag{2.51}$$

The periodogram for $0 \leq \omega \leq \pi$ is found from the samples $k = 0, 1, \ldots, N/2$, assuming N is even. Equation (2.51) is in the form of a discrete Fourier tranform (DFT) and hence the FFT may be used to efficiently perform the computation. To approximate $\hat{P}_{\text{PER}}(\omega)$ more closely, we may need to have a finer frequency spacing. This is accomplished by zero padding the data, i.e., by defining

$$x'(n) = \begin{cases} x(n), & n = 0, 1, \ldots, N-1 \\ 0, & n = N, N+1, \ldots, N'-1 \end{cases} \tag{2.52}$$

Then, letting $\omega_k' = 2\pi k/N'$ for $k = 0, 1, \ldots, N'-1$, we have

$$\begin{aligned} \hat{P}_{\text{PER}}(\omega_k') &= \frac{1}{N}\left|\sum_{n=0}^{N'-1} x'(n) \exp(-j2\pi kn/N')\right|^2 \\ &= \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n) \exp(-j2\pi kn/N')\right|^2 \end{aligned} \tag{2.53}$$

so the effective frequency spacing of the periodogram samples is $2\pi/N' < 2\pi/N$. No extra resolution is afforded by zero padding, but we achieve a better evaluation of the periodogram. See Fig. 2.9 for an example.

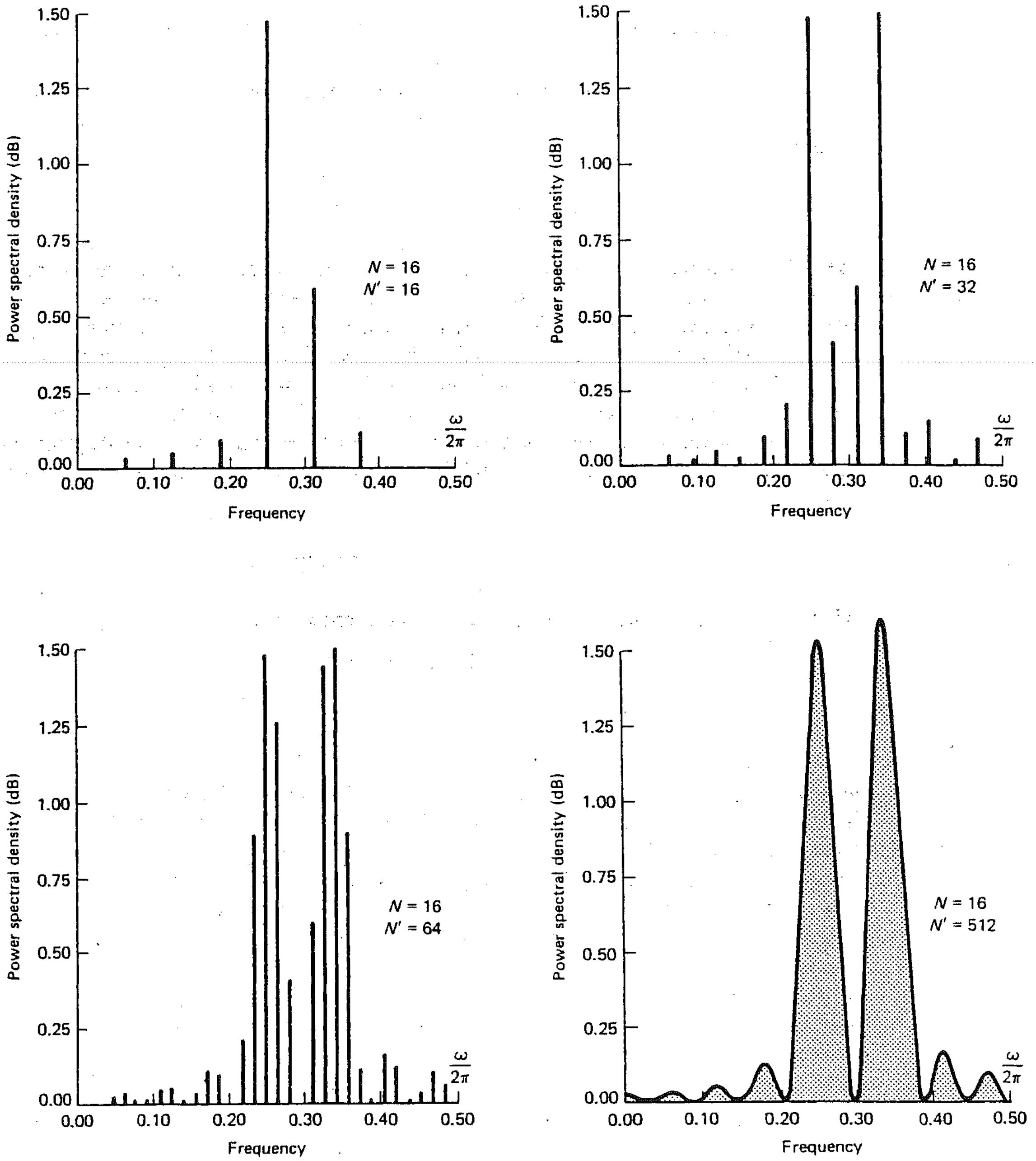


Figure 2.9 Effect of zero padding the data for periodogram computation via the FFT.

2.4.2 Blackman-Tukey Spectral Estimator

We saw in the previous section that the periodogram estimator could be expressed as

$$\hat{P}_{\text{PER}}(\omega) = \sum_{k=-(N-1)}^{N-1} \hat{r}_x(k) \exp(-j\omega k) \tag{2.41}$$

where

$$\hat{r}_x(k) = \frac{1}{N} \sum_{n=0}^{N-1-|k|} x(n)x(n + |k|) \tag{2.42}$$

Here $\hat{r}_x(k)$ is a biased estimator of the autocorrelation function, and in this form the periodogram is seen to be an estimator based on the Wiener-Khinchin theorem. The poor performance of the periodogram may be attributed to the poor performance of the autocorrelation function estimator. In fact, from Eq. (2.41), $r_x(N - 1)$ is estimated by $(1/N)x(0)x(N - 1)$ no matter how large N is. This estimator will be highly variable because of the lack of averaging of lag products, and it will be biased as well. The higher lags of the autocorrelation function will be poorer estimates since they involve fewer lag products. One way to avoid this problem is to weight the higher lags less, or

$$\hat{P}_{BT}(\omega) = \sum_{k=-(N-1)}^{N-1} w(k)\hat{r}_x(k) \exp(-j\omega k) \tag{2.54}$$

where $w(k)$ is a *lag window* with the following properties:

$$\begin{aligned} &1. \quad 0 \le w(k) \le w(0) = 1 \\ &2. \quad w(-k) = w(k) \\ &3. \quad w(k) = 0 \qquad \text{for } |k| > M \end{aligned} \tag{2.55}$$

where $M \le N - 1$. Equation (2.54) is called the Blackman-Tukey spectral estimator. It is equivalent to the periodogram if $w(k) = 1$ for $|k| \le M = N - 1$. The Blackman-Tukey spectral estimator is sometimes called the *weighted covariance* estimator. The approach of reducing the variance of a random variable by weighting it by a factor less than 1 was discussed in Section 2.3.1. Here, the weighting is applied to the autocorrelation function estimator and, as we saw previously, we can expect an increase in the bias. Many windows are available. Table 2.1 lists a few of them. We must be careful to ensure that the chosen window will always lead to a nonnegative spectral estimate. To see how a negative spectral estimate might occur as a result of windowing, note from Eq. (2.54) that

$$\begin{aligned} \hat{P}_{\text{BT}}(\omega) &= \mathcal{F}\{w(k)\hat{r}_x(k)\} \\ &= \int_{-\pi}^{\pi} W(\omega - \xi)\hat{P}_{\text{PER}}(\xi)\frac{d\xi}{2\pi} \end{aligned} \tag{2.56}$$

since $\mathcal{F}\{\hat{r}_x(k)\} = \hat{P}_{\text{PER}}(\omega)$. Although $\hat{P}_{\text{PER}}(\omega) \ge 0$, $W(\omega)$ may be negative enough to cause $\hat{P}_{\text{BT}}(\omega)$ to be negative. To ensure that this will not be the case, $w(k)$ should have

TABLE 2.1 COMMON LAG WINDOWS

Name	Definition	Fourier Transform
Rectangular	$w(k) = \begin{cases} 1, & \lvert k \rvert \le M \\ 0, & \lvert k \rvert > M \end{cases}$	$W(\omega) = W_R(\omega) = \dfrac{\sin \frac{\omega}{2}(2M + 1)}{\sin \omega/2}$
Bartlett	$w(k) = \begin{cases} 1 - \dfrac{\lvert k \rvert}{M}, & \lvert k \rvert \le M \\ 0, & \lvert k \rvert > M \end{cases}$	$W(\omega) = W_B(\omega) = \dfrac{1}{M}\left(\dfrac{\sin M\omega/2}{\sin \omega/2}\right)^2$
Hanning	$w(k) = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\cos\dfrac{\pi k}{M}, & \lvert k \rvert \le M \\ 0, & \lvert k \rvert > M \end{cases}$	$W(\omega) = \dfrac{1}{4}W_R(\omega - \pi/M) + \dfrac{1}{2}W_R(\omega) + \dfrac{1}{4}W_R(\omega + \pi/M)$
Hamming	$w(k) = \begin{cases} 0.54 + 0.46\cos\dfrac{\pi k}{M}, & \lvert k \rvert \le M \\ 0, & \lvert k \rvert > M \end{cases}$	$W(\omega) = 0.23\, W_R(\omega - \pi/M) + 0.54\, W_R(\omega) + 0.23\, W_R(\omega + \pi/M)$
Parzen	$w(k) = \begin{cases} 2\left(1 - \dfrac{\lvert k \rvert}{M}\right)^3 - \left(1 - 2\dfrac{\lvert k \rvert}{M}\right)^3, & \lvert k \rvert \le M/2 \\ 2\left(1 - \dfrac{\lvert k \rvert}{M}\right)^3, & \dfrac{M}{2} < k \le M \\ 0, & \lvert k \rvert > M \end{cases}$	$W(\omega) = \dfrac{8}{M^3}\left(\dfrac{3}{2}\dfrac{\sin^4 M\omega/4}{\sin^4 \omega/2} - \dfrac{\sin^4 M\omega/4}{\sin^2 \omega/2}\right)$

a Fourier transform that is nonnegative. A window that satisfies these requirements is the Bartlett window. Of the other windows given in Table 2.1 only the Parzen window has this property.

We now examine the bias and variance of the Blackman-Tukey spectral estimator. From Eq. (2.56) the mean is

$$E[\hat{P}_{\mathrm{BT}}(\omega)] = \int_{-\pi}^{\pi} W(\omega - \xi)E[\hat{P}_{\mathrm{PER}}(\xi)]\frac{d\xi}{2\pi}$$

If we assume that the periodogram is approximately an unbiased estimator, then

$$E[\hat{P}_{\mathrm{BT}}(\omega)] \approx \int_{-\pi}^{\pi} W(\omega - \xi)P_x(\xi)\frac{d\xi}{2\pi} \tag{2.57}$$

The unbiased assumption will be valid if the data record is long enough so that $P_x(\omega)$ is smooth over any $2\pi/N$ interval. This follows from Eq. (2.44) if we consider $L = N$.

Note from Eq. (2.57) that the mean of the Blackman-Tukey spectral estimator is a smeared version of the true PSD. It is said that $W(\omega)$ acts as a *spectral window*. The PSDs that will be heavily biased are those with nonflat spectra.

The variance of the Blackman-Tukey spectral estimator may be shown [6] to be

$$\mathrm{VAR}[\hat{P}_{\mathrm{BT}}(\omega)] \approx \frac{P_x^2(\omega)}{N} \int_{-\pi}^{\pi} W^2(\xi)\frac{d\xi}{2\pi} \tag{2.58}$$

Equation (2.58) is derived by making the assumptions that $P_x(\omega)$ is smooth over the main lobe of the spectral window ($\approx 4\pi/M$) and that $N \gg M$. Using Parseval's theorem, we can rewrite Eq. (2.58) as

$$\mathrm{VAR}[\hat{P}_{\mathrm{BT}}(\omega)] \approx \frac{P_x^2(\omega)}{N} \sum_{k=-M}^{M} w^2(k) \tag{2.59}$$

As an example, for the Bartlett window,

$$w_B(k) = \begin{cases} 1 - |k|/M, & |k| \le M - 1 \\ 0, & |k| > M \end{cases}$$

$$\mathrm{VAR}[\hat{P}_{\mathrm{BT}}(\omega)] \approx \frac{2M}{3N} P_x^2(\omega) \tag{2.60}$$

Again a bias-variance tradeoff is evident if we examine Eqs. (2.57) and (2.60). For a small bias we would like M large to make the spectral window in Eq. (2.57) behave as an impulse function. On the other hand, for a small variance, M should be small according to Eq. (2.60). Much of the art in nonparametric spectral estimation is in choosing an appropriate window, both in type and in length.

A confidence interval for the Blackman-Tukey spectral estimator may be derived in a similar fashion to that for the averaged periodogram. The results are

$$10 \log_{10} \hat{P}_{\mathrm{BT}}(\omega) \begin{cases} +10 \log_{10} \dfrac{\nu}{\chi_\nu^2(\alpha/2)} \\ -10 \log_{10} \dfrac{\chi_\nu^2(1 - \alpha/2)}{\nu} \end{cases} \mathrm{dB} \tag{2.61}$$

where $\nu = 2N/\sum_{k=-M}^{M} w^2(k)$ = degrees of freedom [6]. As an example, for the Bartlett window, $\nu \approx 3N/M$.

2.4.3 Minimum Variance Spectral Estimation

The minimum variance spectral estimator (MVSE) estimates the PSD by effectively measuring the power out of a set of narrowband filters [7, 8]. The popularly used name maximum likelihood method (MLM) is actually a misnomer in that the spectral estimator is not a maximum likelihood spectral estimator *nor does it possess any of the properties of a maximum likelihood estimator*. Even the name MVSE that has been chosen to describe this estimator is not meant to imply that the spectral estimator is one that possesses minimum variance but is used only to describe the origins of the estimator. The MVSE is also referred to as the Capon spectral estimator [9]. In the MVSE the shapes of the narrowband filters are, in general, dependent on the frequency under consideration, in contrast to the periodogram, for which the shapes of the narrowband filters are the same for all frequencies. In the MVSE the filters adapt to the process for which the PSD is sought with the advantage that the filter sidelobes can be adjusted to reduce the response to spectral components outside the band of interest.

Recall from the discussion of Section 2.4.1 that the periodogram estimates the PSD by forming a bank of narrowband filters with frequency responses given by Eq. (2.34). The frequency response of each filter is unity at $\omega = \omega_0$, the frequency of interest. For a good spectral estimate the filter output power should be due only to the PSD near $\omega = \omega_0$. However, because of the high sidelobes of Eq. (2.34), it is possible to observe a large output power owing to the PSD outside the band of interest. In an attempt to combat this *leakage* it is desirable to design a bank of filters that will adaptively adjust its sidelobes to minimize the power at the filter outputs due to "out of band" spectral components. To do so we can design the filters to minimize the power at their output or we choose the filter to minimize

$$\rho = \int_{-\pi}^{\pi} |H(\omega)|^2 P_x(\omega) \frac{d\omega}{2\pi} \tag{2.62}$$

subject to the constraint $H(\omega_0) = 1$. The filter frequency response is of the form

$$H(\omega) = \sum_{n=-(N-1)}^{0} h(n) \exp(-j\omega n)$$

in accordance with Eq. (2.34). Note that the filter coefficients will in general be complex as was the case for the periodogram. To minimize ρ, note that it may be expressed as

$$\begin{aligned}
\rho &= \int_{-\pi}^{\pi} \sum_{k=-(N-1)}^{0} h(k) \exp(-j\omega k) \sum_{\ell=-(N-1)}^{0} h^*(\ell) \exp(j\omega\ell) P_x(\omega) \frac{d\omega}{2\pi} \\
&= \sum_{k=-(N-1)}^{0} \sum_{\ell=-(N-1)}^{0} h(k)h^*(\ell) \int_{-\pi}^{\pi} P_x(\omega) \exp[j\omega(\ell - k)] \frac{d\omega}{2\pi} \\
&= \sum_{k=-(N-1)}^{0} \sum_{\ell=-(N-1)}^{0} h(k)h^*(\ell) r_x(\ell - k) \\
&= \mathbf{h}^H \mathbf{R}_x \mathbf{h}
\end{aligned} \tag{2.63}$$

where $\mathbf{h}^* = [h(0)\, h(-1) \ldots h(-(N-1))]^T$, $\mathbf{R}_x$ is the $N \times N$ autocorrelation matrix with (i,j) element $r_x(i - j)$, and H denotes the complex conjugate transpose. The constraint of unity frequency response at $\omega = \omega_0$ can also be rewritten as

$$\mathbf{h}^H \mathbf{e} = 1$$

where $\mathbf{e} = [1 \exp(j\omega_0) \ldots \exp[j(N - 1)\omega_0]^T$. This constrained minimization may be accomplished by making use of the identity

$$\mathbf{h}^H \mathbf{R}_x \mathbf{h} = (\mathbf{h} - \bar{\mathbf{h}})^H \mathbf{R}_x (\mathbf{h} - \bar{\mathbf{h}}) + \bar{\mathbf{h}}^H \mathbf{R}_x \bar{\mathbf{h}} \tag{2.64}$$

where

$$\bar{\mathbf{h}} = \frac{\mathbf{R}_x^{-1}\mathbf{e}}{\mathbf{e}^H\mathbf{R}_x^{-1}\mathbf{e}}$$

and which holds if $\mathbf{h}^H\mathbf{e} = 1$. To verify this identity, note that

$$\begin{aligned}(\mathbf{h} - \bar{\mathbf{h}})^H\mathbf{R}_x(\mathbf{h} - \bar{\mathbf{h}}) + \bar{\mathbf{h}}^H\mathbf{R}_x\bar{\mathbf{h}} &= \mathbf{h}^H\mathbf{R}_x\mathbf{h} + (2\bar{\mathbf{h}}^H\mathbf{R}_x\bar{\mathbf{h}} - \bar{\mathbf{h}}^H\mathbf{R}_x\mathbf{h} - \mathbf{h}^H\mathbf{R}_x\bar{\mathbf{h}})\\ &= \mathbf{h}^H\mathbf{R}_x\mathbf{h} + \frac{1}{\mathbf{e}^H\mathbf{R}_x^{-1}\mathbf{e}}(2 - \mathbf{e}^H\mathbf{h} - \mathbf{h}^H\mathbf{e})\end{aligned}$$

Since $\mathbf{h}^H\mathbf{e} = 1$, it follows that $\mathbf{e}^H\mathbf{h} = 1$, and hence the identity is proved. To minimize the variance, observe from Eq. (2.64) that $\bar{\mathbf{h}}^H\mathbf{R}_x\bar{\mathbf{h}}$ does not depend on $\mathbf{h}$ and $(\mathbf{h} - \bar{\mathbf{h}})^H\mathbf{R}_x(\mathbf{h} - \bar{\mathbf{h}}) \geq 0$ because $\mathbf{R}_x$ is a positive-definite matrix. Therefore, the minimum value is obtained by letting $\mathbf{h} = \bar{\mathbf{h}}$, or

$$\mathbf{h} = \frac{\mathbf{R}_x^{-1}\mathbf{e}}{\mathbf{e}^H\mathbf{R}_x^{-1}\mathbf{e}} \tag{2.65}$$

Finally, substituting Eq. (2.65) into Eq. (2.63), we obtain the minimum power

$$\rho = \frac{1}{\mathbf{e}^H\mathbf{R}_x^{-1}\mathbf{e}} \tag{2.66}$$

As an example, assume that $x(n)$ is a first-order autoregressive process with parameter $a_1 = -r$. The PSD is that of a lowpass process with its power concentrated at $\omega = 0$ (see Fig. 2.11b in the next section for an example). The autocorrelation function is shown in Section 2.5.2 to be

$$r_x(k) = \frac{\sigma^2}{1 - a_1^2}(-a_1)^{|k|}$$

Using this, we can easily verify that

$$\mathbf{R}_x^{-1} = \frac{1}{\sigma^2}\begin{bmatrix} 1 & a_1 & 0 & 0 & \cdots & 0 \\ a_1 & 1 + a_1^2 & a_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_1 & 1 + a_1^2 & a_1 \\ 0 & \cdots & 0 & 0 & a_1 & 1 \end{bmatrix}$$

so that

$$\mathbf{R}_x^{-1}\mathbf{e} = \frac{1}{\sigma^2}\begin{bmatrix} A^*(\omega_0) \\ \exp(j\omega_0)|A(\omega_0)|^2 \\ \vdots \\ \exp[j\omega_0(N-2)]|A(\omega_0)|^2 \\ \exp[j\omega_0(N-1)]A(\omega_0) \end{bmatrix}$$

where $A(\omega_0) = 1 - r\exp(-j\omega_0)$. It follows that

$$\mathbf{e}^H\mathbf{R}_x^{-1}\mathbf{e} = \frac{1}{\sigma^2}[N - 2(N-1)r\cos\omega_0 + (N-2)r^2]$$

The filter coefficients and hence the frequency response can be found by using these two expressions in Eq. (2.65). The magnitude of the frequency response is plotted in Fig. 2.10 for $N = 10$, $\omega_0/2\pi = 0.25$ for various values of r. For $r = 0$, i.e., $a_1 = 0$, the noise is white and the filter has the usual sinc-type response as given in Eq. (2.34) (shown as a dashed curve in the figure). For other values of r, the optimal filter attempts to reject the noise by adjusting the response so as to attenuate that region of the frequency band where the noise PSD is the largest. In this case the PSD of the noise is (see Section 2.5.2)

$$P_{\text{AR}}(\omega) = \frac{\sigma^2}{|1 + a_1\exp(-j\omega)|^2} = \frac{\sigma^2}{|1 - r\exp(-j\omega)|^2} \tag{2.67}$$

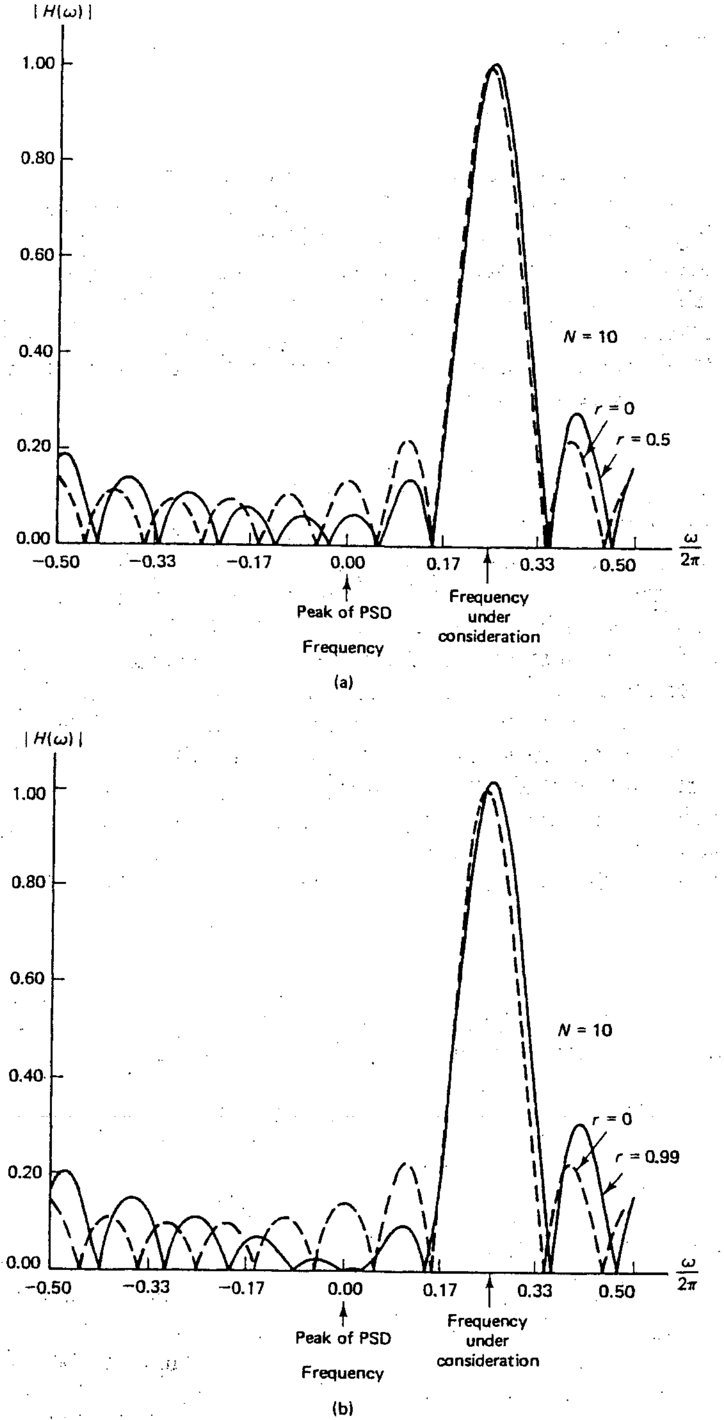
The PSD has a peak at $\omega/2\pi = 0$ and the sharpness of the peak increases with r. This is reflected in the frequency response, in which the attenuation in the frequency band centered about $\omega/2\pi = 0$ increases with r. In fact, as $r \rightarrow 1$, so that the PSD approaches a Dirac delta function located at $\omega/2\pi = 0$, it can be shown that $H(0) \rightarrow 0$. In effect, a null is placed at the frequency location where the noise power is greatest. As expected, the gain of the filter at $\omega/2\pi = \omega_0/2\pi = 0.25$ is unity.

It is now apparent that the frequency response of the optimal filter will depend on the noise background near the center frequency $\omega = \omega_0$. For a given PSD **h** will form a different filter for each assumed value of ω_0. The filter will adjust itself to reject noise components with frequencies not near $\omega = \omega_0$ and to pass noise components at and near $\omega = \omega_0$. The power out of the filter is therefore a good indication of the power in the process in the vicinity of ω_0. A spectral estimator can thus be defined as

$$\hat{P}_{\text{MV}}(\omega) = \frac{p}{\mathbf{e}^H\hat{\mathbf{R}}_x^{-1}\mathbf{e}} \tag{2.68}$$

where $\mathbf{e} = [1\ \exp(j\omega)\ \exp(j2\omega)\ \ldots\ \exp(j\omega(p-1))]^T$. The factor p is included to account for the bandwidth of the filter, i.e., to yield a power spectral *density* estimate by dividing the estimate of the power by $1/p$, which is approximately the 3-dB bandwidth of the optimal filter. To compute the spectral estimate we need only an estimate of the $p \times p$ autocorrelation matrix $\mathbf{R}_x$. Note that the dimensions of the autocorrelation matrix should be much less than $N \times N$ to allow the higher order autocorrelation samples to be reliably estimated. The spectral estimator $\hat{P}_{MV}(\omega)$ has been referred to as the MLM or Capon method.

In practice, the MVSE exhibits more resolution than the periodogram and Blackman-Tukey spectral estimators but less than an autoregressive (AR) spectral estimator [8]. Also, the variance of the MVSE appears to be less than that of the AR spectral estimator when both are based on the same number of autocorrelation lags. The critical choice in the MVSE is in the value of p, the dimension of the estimated autocorrelation matrix. For large p, the filter bandwidth will be small, which will yield high-resolution spectral estimates. However, the variance may also be large due to the



(a)

(b)

Figure 2.10 Frequency response magnitude of optimal filter: (a) $r = 0.5$, (b) $r = 0.99$.

large number of autocorrelation lags estimated. As usual, then, a tradeoff must be effected between bias and variance.

In summary, nonparametric spectral estimation based on a limited data set involves trading off bias for variance. A basic problem is that the PSD depends on an infinite number of autocorrelation function lags, all of which need to be estimated to obtain a good spectral estimate. Noting the impossibility of the task, one may reasonably ask whether it might be better to assume a model for the PSD or autocorrelation function that depends on only a finite set of parameters. If the number of data points is large relative to the number of PSD parameters, then good estimates of the parameters and hence the PSD would be expected. This approach is termed *parametric spectral estimation* and is based on classical models of time series analysis. We will study this alternative spectral estimation approach in the next section.

2.5 PARAMETRIC SPECTRAL ESTIMATION

2.5.1 Time Series Models

The most general time series model is one in which the random process is assumed to have been generated by exciting a linear shift-invariant causal pole-zero filter with white noise, or

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + \sum_{k=0}^{q} b_k \epsilon(n-k) \tag{2.69}$$

where $b_0 = 1$ and $\epsilon(n)$ is a white noise process with zero mean and variance σ^2. Such a model is termed an autoregressive moving average (ARMA) model and is given the abbreviation ARMA(p, q). We see that $x(n)$ is the output of the linear shift-invariant filter with transfer function

$$\begin{aligned} H(z) &= \frac{B(z)}{A(z)} \\ &= \frac{1 + \sum_{k=1}^{q} b_k z^{-k}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \end{aligned} \tag{2.70}$$

There are no restrictions on the number of poles p or zeros q. It is assumed that both $A(z)$ and $B(z)$ have their zeros inside the unit circle of the z-plane so that $H(z)$ as well as $1/H(z)$ are stable and causal filters. The latter requirements are reasonable since the time series models are usually justified by physical considerations.

The PSD of $x(n)$ follows from Eq. (2.70) as

$$\begin{aligned} P_{\text{ARMA}}(\omega) &= |H(\exp(j\omega))|^2 P_\epsilon(\omega) \\ &= \frac{\sigma^2 |1 + \sum_{k=1}^{q} b_k \exp(-j\omega k)|^2}{|1 + \sum_{k=1}^{p} a_k \exp(-j\omega k)|^2} \end{aligned} \tag{2.71}$$

Equation (2.71) represents a model for the PSD based on the time series model of Eq. (2.69) for the data. Note that the PSD depends *only on the filter coefficients and white noise variance*. To estimate the PSD we need only estimate $\{b_1, b_2, \ldots, b_q, a_1, a_2, \ldots, a_p, \sigma^2\}$ and substitute the estimated values into Eq. (2.71). We hope that $N \gg p + q + 1$ so that we will obtain good estimates of the unknown parameters. Equivalently, the autocorrelation function must also be a function of the same model parameters and, hence, we have achieved our objective of replacing estimation of an infinite set of autocorrelation function lags by estimation of a finite set of parameters.

It should be observed that the time series model of Eq. (2.69) is fundamentally different from the deterministic signal model discussed in Chapter 1. For the deterministic signal it was appropriate to model it as an impulse response, with the error being the difference between the actual signal and the modeled signal. For the spectral estimation problem the "signal" itself is inherently random, and the error is the difference between the true PSD and the estimated one. Alternately, for our problem there does not exist a single "signal" that is to be modeled in terms of its *sequence values* but rather an ensemble of sequences for which the *average distribution of power with frequency* is sought. Curiously, though, some of the estimation algorithms are nearly identical, but this equivalence is coincidental, attributable to the need for easily implementable algorithms; in general the *optimal algorithms* for the two cases will be quite different. Finally, assessing the performance of an estimation algorithm must be done in the context of the fundamental data assumptions made. Failure to do so results in a comparison of "apples and oranges."

Less general, although more useful, models in practice are found by setting either $b_k = 0$ for $k = 1, 2, \ldots, q$ or $a_k = 0$ for $k = 1, 2, \ldots, p$ in Eq. (2.69). Adopting the latter leads to the moving average (MA) times series model

$$x(n) = \sum_{k=0}^{q} b_k \epsilon(n-k) \tag{2.72}$$

and a corresponding PSD model

$$P_{\text{MA}}(\omega) = \sigma^2 \left| 1 + \sum_{k=1}^{q} b_k \exp(-j\omega k) \right|^2 \tag{2.73}$$

The abbreviation MA(q) is used and q is termed the MA model order. The b_k's and σ^2 are called the MA parameters. If we set the b_k's equal to zero in the ARMA model, we then have the autoregressive (AR) time series model

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + \epsilon(n) \tag{2.74}$$

and a corresponding PSD model

$$P_{\text{AR}}(\omega) = \frac{\sigma^2}{\left| 1 + \sum_{k=1}^{p} a_k \exp(-j\omega k) \right|^2} \tag{2.75}$$

The abbreviation AR(p) is used, and p is termed the AR model order. The a_k's and σ^2 are called the AR parameters.

Usually, the appropriate model is chosen based on physical modeling, as in vocal tract modeling for speech [10]. Frequently, in practice one does not know which of the given models yields an accurate representation of the PSD. An important result from the Wold decomposition and Kolmogorov theorems is that any ARMA or MA process can be represented by an AR process of possibly infinite order; likewise, any ARMA or AR process can be represented by an MA process of possibly infinite order [11,12]. Hence, if we choose the wrong model among the three, we may still obtain a reasonable approximation by using a high enough model order.

An important relationship for parametric spectral estimation is that between the parameters of the model and the autocorrelation function. Considering the general ARMA model, we now derive the celebrated Yule-Walker equations. First, multiply Eq. (2.69) by $x(n-k)$ and take the expectation:

$$E[x(n)x(n-k)] = -\sum_{\ell=1}^{p} a_\ell E[x(n-\ell)x(n-k)] + \sum_{\ell=0}^{q} b_\ell E[\epsilon(n-\ell)x(n-k)]$$

$$r_x(k) = -\sum_{\ell=1}^{p} a_\ell r_x(k-\ell) + \sum_{\ell=0}^{q} b_\ell r_{x\epsilon}(k-\ell) \tag{2.76}$$

where $r_{x\epsilon}(k) = E[x(n)\epsilon(n+k)]$. To evaluate $r_{x\epsilon}(k)$ note that $x(n)$ is the output of a filter with impulse response $h(n)$ excited at the input by $\epsilon(n)$ so that

$$\begin{aligned} r_{x\epsilon}(k) &= E\left[\sum_{\ell=-\infty}^{\infty} h(n-\ell)\epsilon(\ell)\epsilon(n+k)\right] \\ &= \sum_{\ell=-\infty}^{\infty} h(n-\ell)\sigma^2\delta(n+k-\ell) \qquad (2.77) \\ &= h(-k)\sigma^2 \end{aligned}$$

Since the filter is causal, $r_{x\epsilon}(k) = 0$ for $k > 0$. Equation (2.76) becomes

$$r_x(k) = \begin{cases} -\sum_{\ell=1}^{p} a_\ell r_x(k-\ell) + \sigma^2 \sum_{\ell=0}^{q-k} h(\ell) b_{\ell+k} & k = 0, 1, \ldots, q \\ -\sum_{\ell=1}^{p} a_\ell r_x(k-\ell) & k \geq q+1 \end{cases} \tag{2.78}$$

Equations (2.78) are termed the Yule-Walker equations. They relate the autocorrelation function to the ARMA parameters and are useful for estimating the ARMA parameters given estimates of the autocorrelation function lags.

We now give examples of the autocorrelation function and PSD for low-order AR, MA, and ARMA models. By doing so we will observe the types of spectra that are representable by these models.

2.5.2 Examples of AR Processes

Consider first an AR(p) model. From Eq. (2.78), on letting $q = 0$, we have

$$r_x(k) = \begin{cases} -\sum_{\ell=1}^{p} a_\ell r_x(k-\ell) + \sigma^2 h(k), & k = 0 \\ -\sum_{\ell=1}^{p} a_\ell r_x(k-\ell), & k \geq 1 \end{cases}$$

But using Eqs. (2.70), we have

$$h(0) = \lim_{z\to\infty} H(z) = 1$$

so

$$r_x(k) = \begin{cases} -\sum_{\ell=1}^{p} a_\ell r_x(\ell) + \sigma^2, & k = 0 \\ -\sum_{\ell=1}^{p} a_\ell r_x(k-\ell), & k \geq 1 \end{cases} \tag{2.79}$$

Equations (2.79) allow us to compute the autocorrelation function recursively given the AR parameters, an example of which will be given shortly. The process is reversible since given $\{r_x(0), r_x(1), \ldots, r_x(p)\}$ we can determine the AR parameters using Eq. (2.79) for $k = 1, 2, \ldots, p$ as follows

$$\begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(p-1) \\ r_x(1) & r_x(0) & \cdots & r_x(p-2) \\ \vdots & \vdots & \ddots & \ddots \\ r_x(p-1) & r_x(p-2) & \cdots & r_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = -\begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(p) \end{bmatrix} \tag{2.80}$$

$$\sigma^2 = r_x(0) + \sum_{\ell=1}^{p} a_\ell r_x(\ell)$$

The AR parameters can be found by solving this set of linear equations. As explained in Chapter 1, due to the special nature of the matrix and right-hand vector, the Levinson algorithm may be used to solve Eq. (2.80). The algorithm is summarized below.

$$a_{11} = -\frac{r_x(1)}{r_x(0)}$$

$$\sigma_1^2 = (1 - a_{11}^2) r_x(0)$$

For $k = 2, 3, \ldots, p$,

$$\begin{aligned} a_{kk} &= -\frac{r_x(k) + \sum_{\ell=1}^{k-1} a_{k-1,\ell} r_x(k-\ell)}{\sigma_{k-1}^2} \\ a_{ki} &= a_{k-1,i} + a_{kk} a_{k-1,k-i}, \qquad i = 1, 2, \ldots, k-1 \\ \sigma_k^2 &= (1 - a_{kk}^2)\sigma_{k-1}^2 \end{aligned} \tag{2.81}$$

The solution of Eq. (2.80) is given by $a_i = a_{pi}$, $i = 1, 2, \ldots, p$, and $\sigma^2 = \sigma_p^2$. For further details see Chapter 1.

As an example of the utility of Eq. (2.79) consider an AR(1) process. Then, for $p = 1$, Eq. (2.79) yields

$$r_x(k) = -a_1 r_x(k-1), \qquad k \geq 1$$

which leads to the solution

$$r_x(k) = r_x(0)(-a_1)^{|k|}$$

To find $r_x(0)$ we use Eq. (2.79) for $k = 0$ to obtain

$$\begin{aligned} \sigma^2 &= r_x(0) + a_1 r_x(1) \\ &= r_x(0) - a_1^2 r_x(0) \end{aligned}$$

which results in

$$r_x(0) = \frac{\sigma^2}{1 - a_1^2}$$

or finally

$$r_x(k) = \frac{\sigma^2}{1 - a_1^2}(-a_1)^{|k|} \tag{2.82}$$

Note that $|a_1| < 1$ to guarantee stability of $1/A(z)$. The autocorrelation function $r_x(k)$ is plotted in Fig. 2.11 for $a_1 < 0$ and $a_1 > 0$. The corresponding PSDs are given by

$$P_{\text{AR}}(\omega) = \frac{\sigma^2}{|1 + a_1 \exp(-j\omega)|^2} \tag{2.83}$$

and are also plotted in decibels in Fig. 2.11. Note that $a_1 < 0$ yields a lowpass process while $a_1 > 0$ yields a highpass process. Thus, an AR(1) process cannot model a bandpass process. To do so we must use an AR(2) process with complex conjugate poles, i.e., poles at $z = r\exp(\pm j\omega_0)$. It can be shown [13] that for this case

$$r_x(k) = \frac{\sigma^2 \dfrac{1+r^2}{1-r^2}\sqrt{1 + \left(\dfrac{1-r^2}{1+r^2}\right)^2 \cot^2(\omega_0)}}{1 - 2r^2\cos 2\omega_0 + r^4} r^{|k|}\cos(|k|\omega_0 - \phi) \tag{2.84}$$

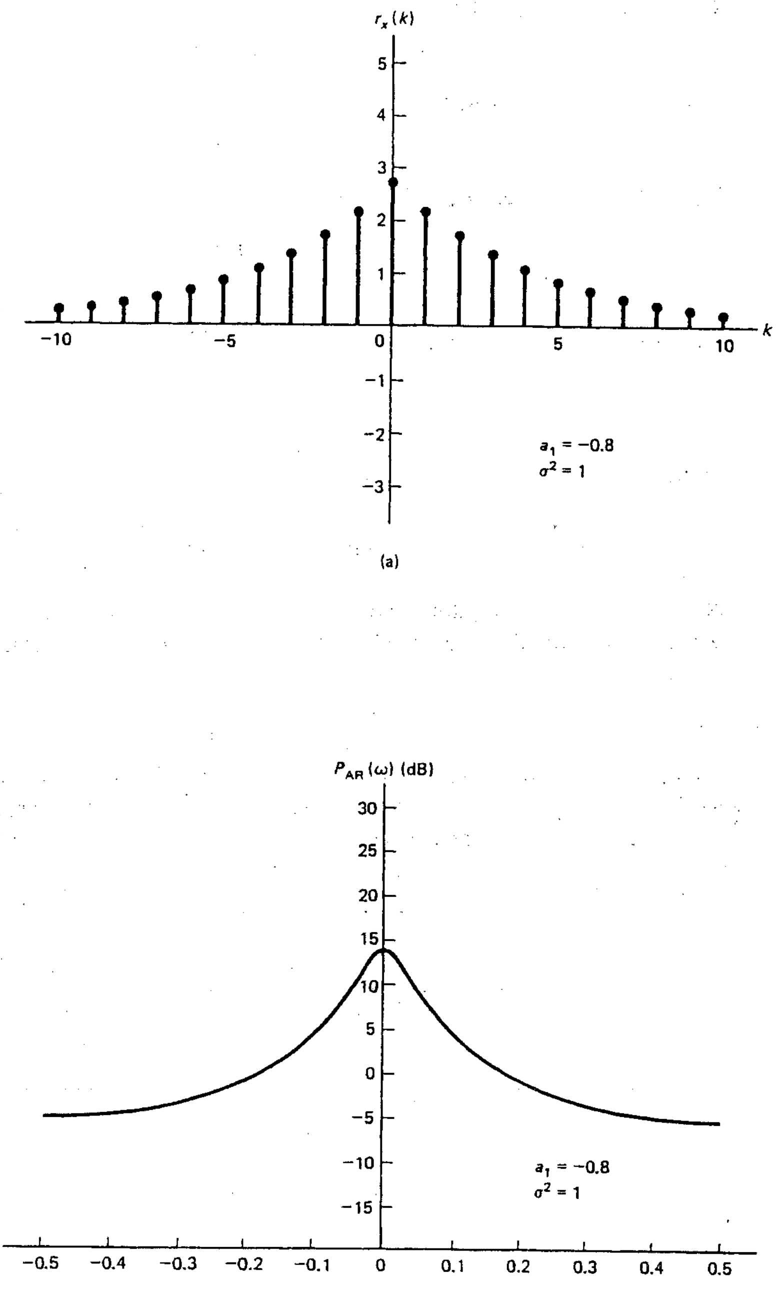
where

$$\phi = \arctan\left[\frac{1-r^2}{1+r^2}\cot\omega_0\right]$$

$$a_1 = -2r\cos\omega_0, \qquad a_2 = r^2$$

The PSD is

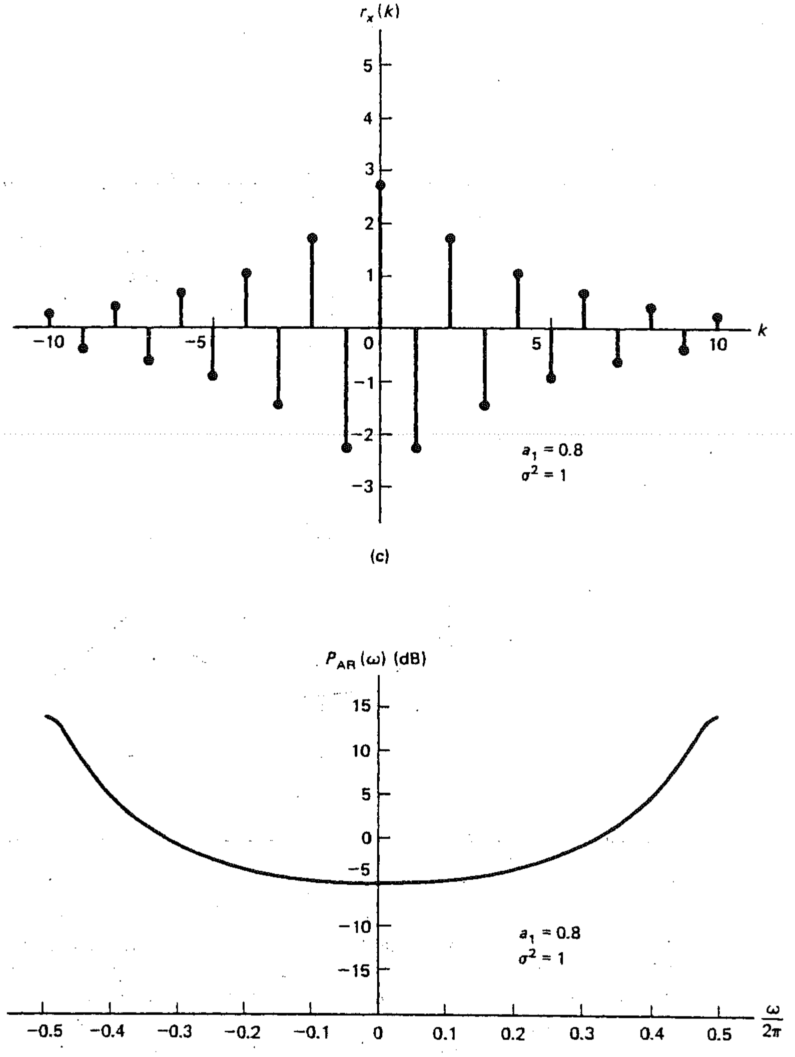
$$\begin{aligned} P_{\text{AR}}(\omega) &= \frac{\sigma^2}{|1 + a_1\exp(-j\omega) + a_2\exp(-j2\omega)|^2} \\ &= \frac{\sigma^2}{|1 - r\exp[-j(\omega - \omega_0)]|^2 |1 - r\exp[-j(\omega + \omega_0]|^2} \end{aligned} \tag{2.85}$$



(a)

(b)

Figure 2.11 (a) Autocorrelation of AR(1) process; (b) power spectral density of AR(1) process; (c) autocorrelation of AR(1) process; (d) power spectral density of AR(1) process.



(c)

(d)

Figure 2.11 (*cont.*)

Examples of the autocorrelation function and PSD for complex poles at $r \exp(\pm j\pi/2)$ are given in Fig. 2.12. As $r \to 1$ the PSD becomes more peaked about $\omega = \omega_0$ and the autocorrelation function becomes more sinusoidal. In general, to represent L spectral peaks none of which are at $\omega = 0$ or π requires a model order of $p = 2L$.

2.5.3 Examples of MA Processes

To find the autocorrelation function for an MA(q) process we use Eq. (2.78) with $p = 0$. Then

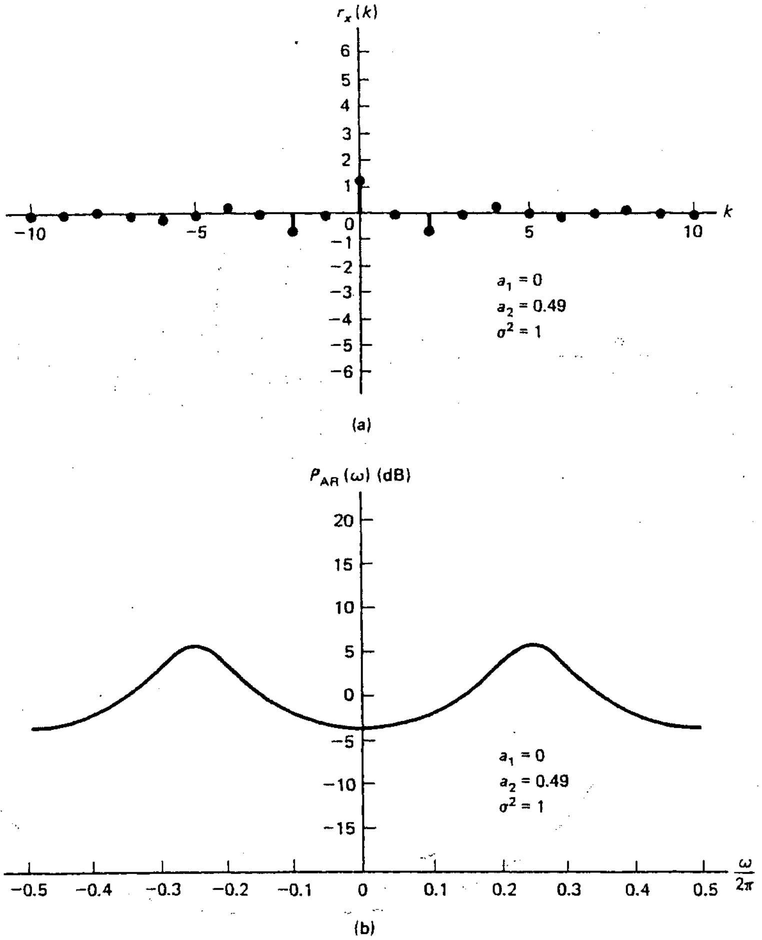


Figure 2.12 (a) Autocorrelation of AR(2) process with complex conjugate poles at $0.7\exp[\pm j2\pi(0.25)]$; (b) power spectral density of AR(2) process with complex conjugate poles at $0.7\exp[\pm j2\pi(0.25)]$; (c) autocorrelation of AR(2) process with complex conjugate poles at $0.95\exp[\pm j2\pi(0.25)]$; (d) power spectral density of AR(2) process with complex conjugate poles at $0.95\exp[\pm j2\pi(0.25)]$.

$$r_x(k) = \begin{cases} \sigma^2 \sum_{\ell=0}^{q-k} h(\ell) b_{\ell+k}, & k = 0, 1, \ldots, q \\ 0, & k \geq q + 1 \end{cases}$$

But

$$h(k) = \begin{cases} b_k, & k = 0, 1, \ldots, q \\ 0, & \text{otherwise} \end{cases}$$

which results in

$$r_x(k) = \begin{cases} \sigma^2 \sum_{\ell=0}^{q-|k|} b_\ell b_{\ell+k}, & |k| \leq q \\ 0, & |k| > q \end{cases} \tag{2.86}$$

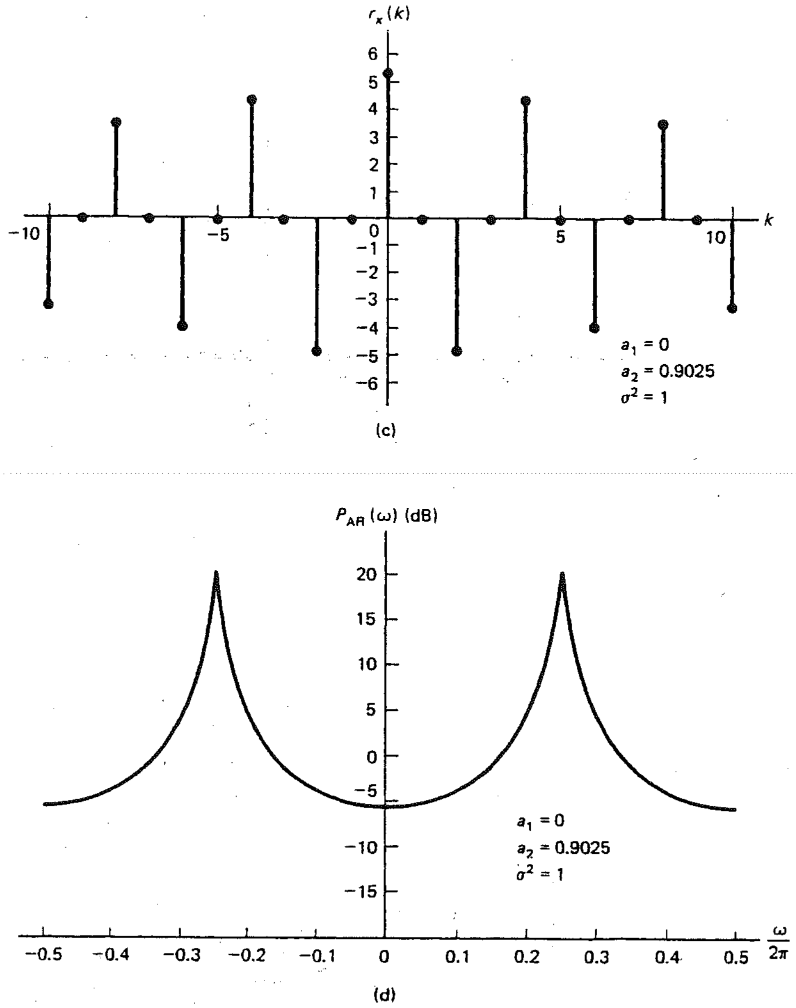


Figure 2.12 (*cont.*)

For an MA(1) process we have, on using Eq. (2.86),

$$r_x(k) = \begin{cases} \sigma^2(1 + b_1^2), & k = 0 \\ \sigma^2 b_1, & k = 1 \\ 0, & k \geq 2 \end{cases} \tag{2.87}$$

with a corresponding PSD

$$P_{MA}(\omega) = \sigma^2 |1 + b_1 \exp(-j\omega)|^2 \tag{2.88}$$

An example is given in Fig. 2.13 in which we see that the PSD exhibits a dip at $\omega = 0$, i.e., at the zero location of $B(z)$. For spectral valleys at other frequency locations we

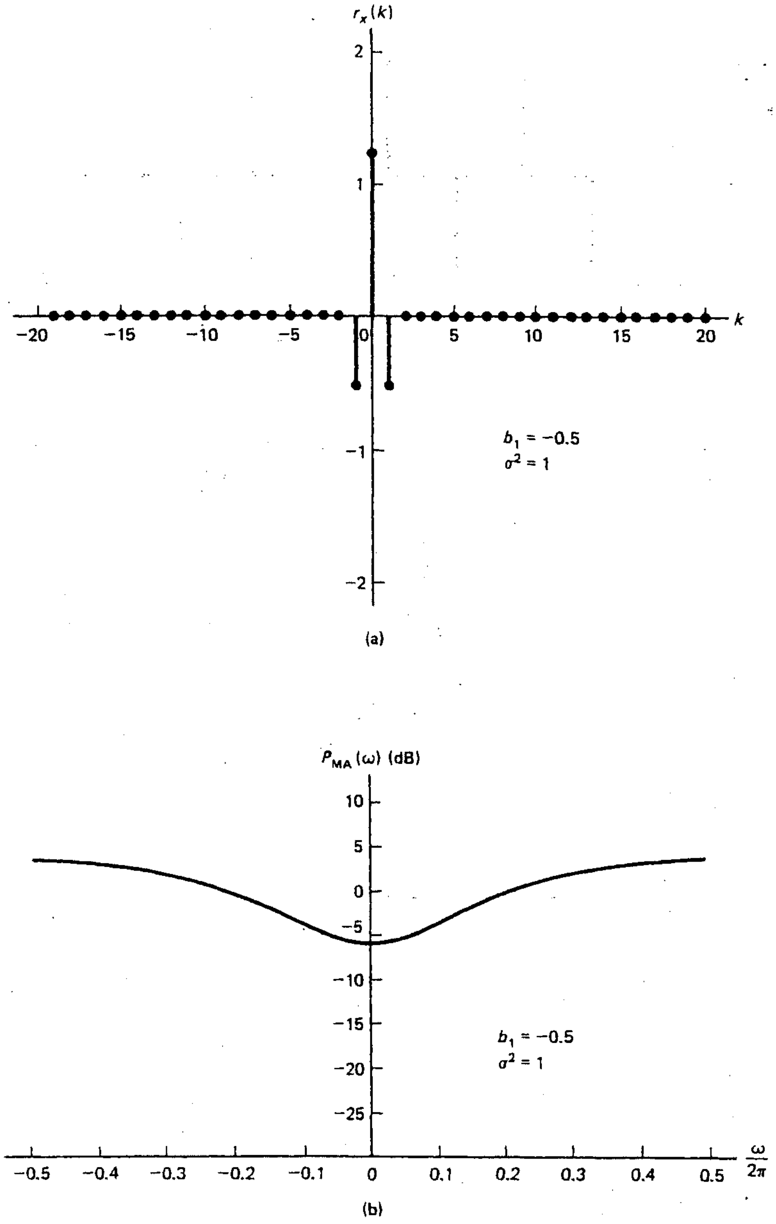
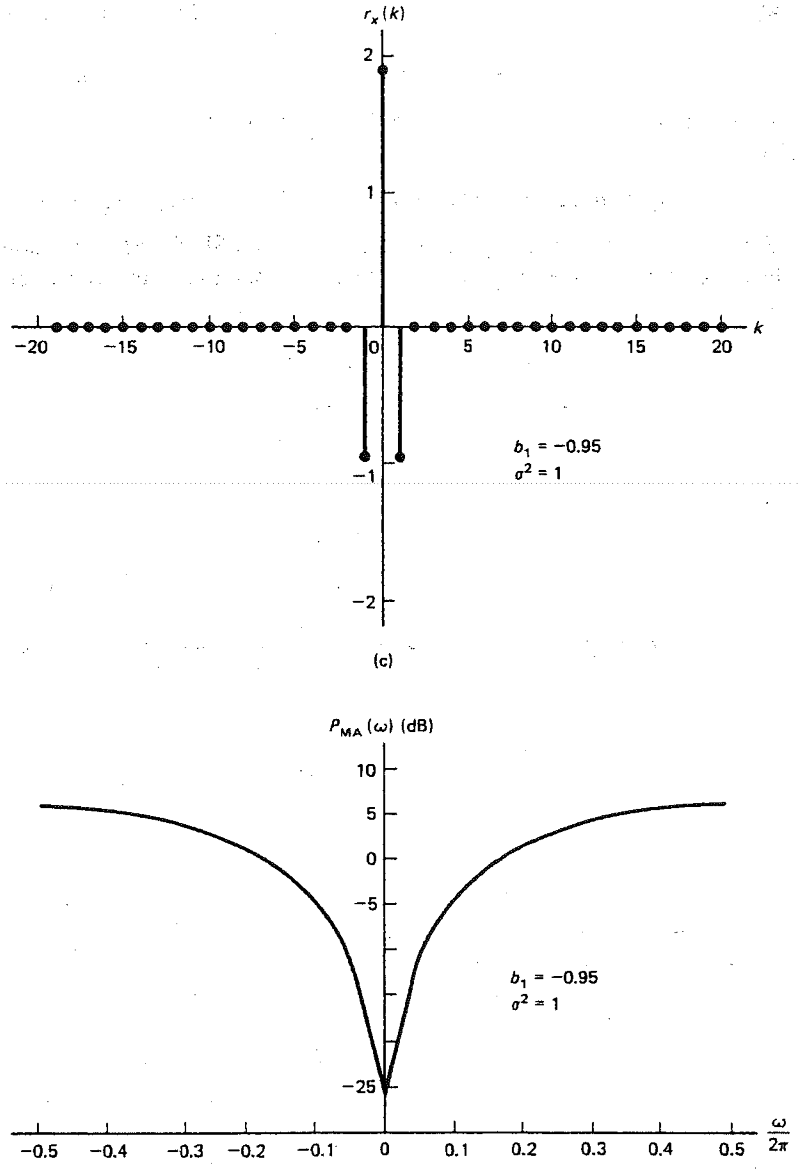


Figure 2.13 (a) Autocorrelation of MA(1) process; (b) power spectral density of MA(1) process; (c) autocorrelation of MA(1) process; (d) power spectral density of MA(1) process.



(c)

(d)

Figure 2.13 (*cont.*)

must use a higher-order MA model. For an MA(2) process,

$$r_x(k) = \begin{cases} \sigma^2(1 + b_1^2 + b_2^2), & k = 0 \\ \sigma^2(b_1 + b_1 b_2), & k = 1 \\ \sigma^2 b_2, & k = 2 \\ 0, & k \geq 3 \end{cases} \tag{2.89}$$

and

$$P_{\mathrm{MA}}(\omega) = \sigma^2 |\, 1 + b_1 \exp(-j\omega) + b_2 \exp(-j2\omega)\,|^2 \tag{2.90}$$

Assuming complex zeros at $r \exp(\pm j\omega_0)$ so that $b_1 = -2r \cos \omega_0$, $b_2 = r^2$, the PSD becomes

$$P_{\mathrm{MA}}(\omega) = \sigma^2 |\, 1 - r \exp[-j(\omega - \omega_0)]\,|^2 |\, 1 - r \exp[-j(\omega + \omega_0)]\,|^2 \tag{2.91}$$

Examples are shown in Fig. 2.14. Note that the PSD for MA processes tends to be broadband but may exhibit nulls if the zeros are close to the unit circle.

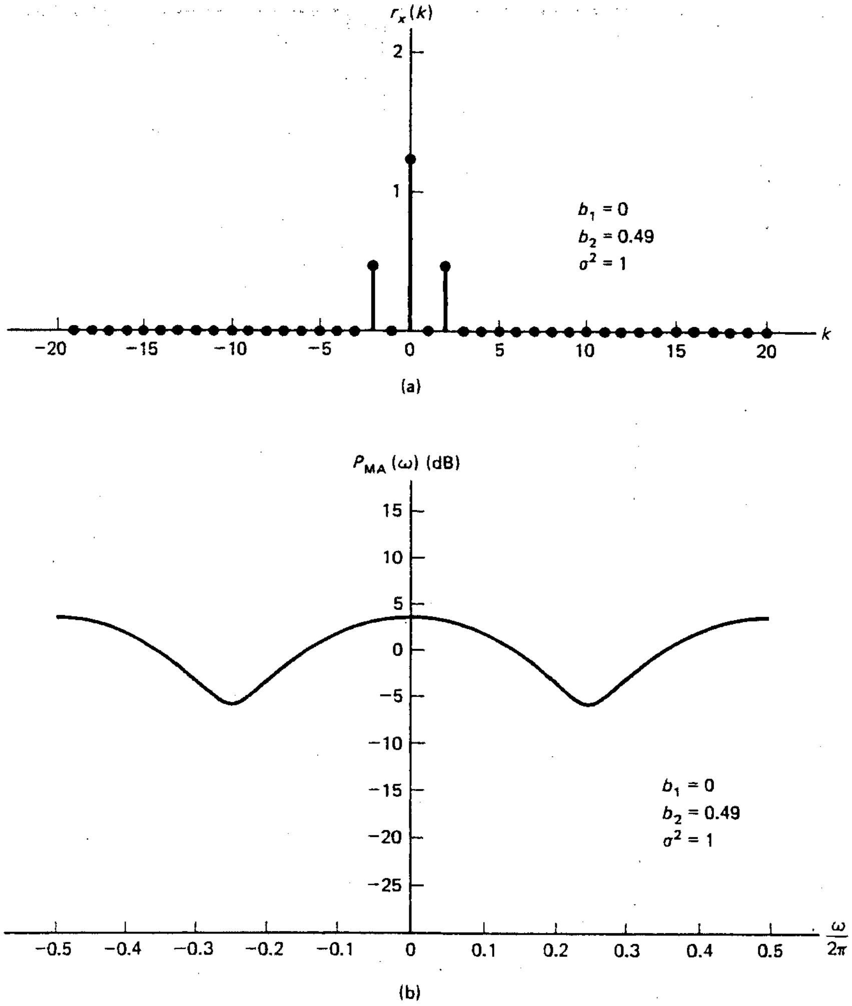
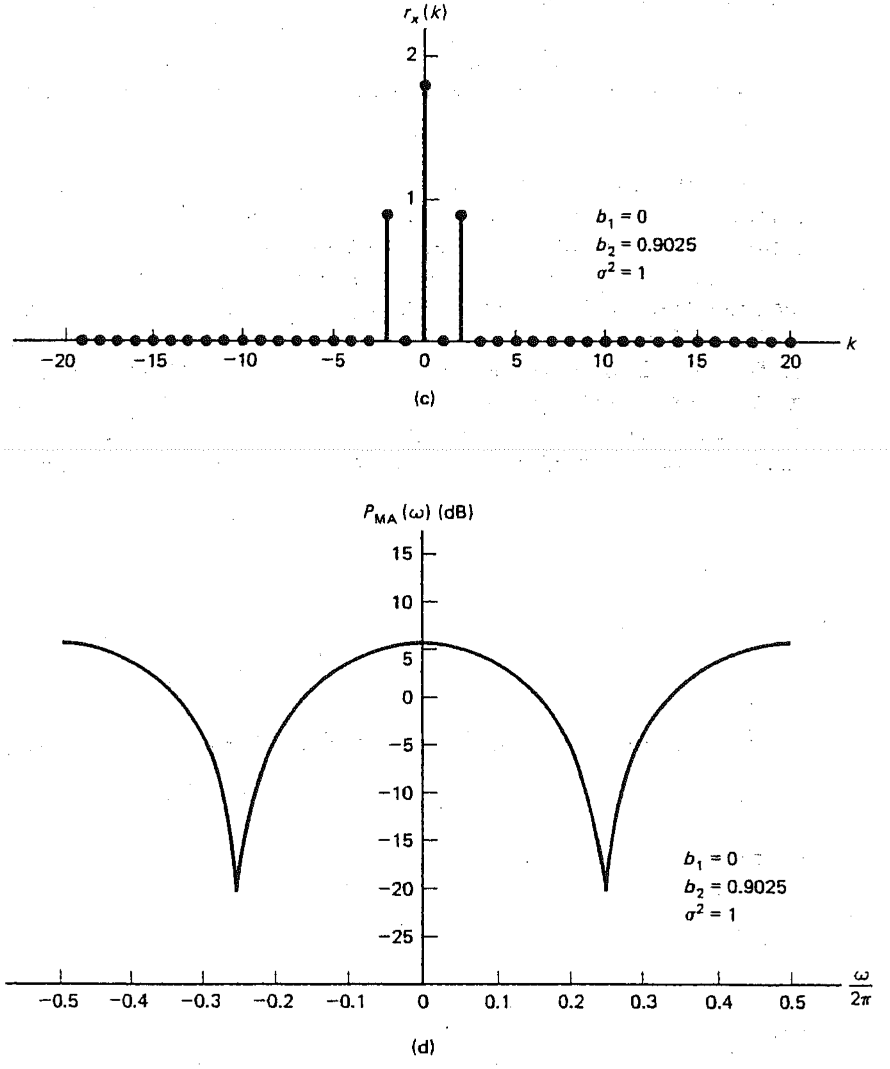


Figure 2.14 (a) Autocorrelation of MA(2) process with complex conjugate zeros at $0.7 \exp[\pm j2\pi(0.25)]$; (b) power spectral density of MA(2) process with complex conjugate zeros at $0.7 \exp[\pm j2\pi(0.25)]$; (c) autocorrelation of MA(2) process with complex conjugate zeros at $0.95 \exp[\pm j2\pi(0.25)]$; (d) power spectral density of MA(2) process with complex conjugate zeros at $0.95 \exp[\pm j2\pi(0.25)]$.



(c)

(d)

Figure 2.14 (*cont.*)

2.5.4 ARMA Processes

ARMA processes have PSDs that in general exhibit both spectral peaks and spectral valleys. If the zeros of the process are not near the unit circle, then an AR model may be appropriate. Likewise, if the poles of the process are not near the unit circle, then an MA process may suffice. If neither of these conditions is satisfied, then the general ARMA model will be necessary. The autocorrelation function and PSD for an ARMA process are similar in nature to those for the MA and AR processes and so will be omitted here. Examples may be found in [14].

2.5.5 Mixed Processes

In some cases the actual process may be modeled well by the sum of two uncorrelated time series. As an example, consider an AR(2) process plus white noise, or

$$y(n) = x(n) + w(n) \tag{2.92}$$

where $w(n)$ is zero-mean white noise with variance σ_w^2, and $x(n)$ is an AR(2) process. A process given by Eq. (2.92) is termed a mixed process. If we were to use an AR(2) model for $y(n)$ and solve for the AR parameters via the Yule-Walker equations using $r_y(k) = r_x(k) + \sigma_w^2\delta(k)$, we would obtain unsatisfactory results. For a signal-to-noise ratio (SNR) of SNR $= 10 \log_{10} r_x(0)/\sigma_w^2 = 5$ dB the PSD that is obtained is shown in Fig. 2.15(a). If $p = 5$ is used, a better spectral fit is obtained, as shown in Fig. 2.15(b). This is to be expected since according to the Wold decomposition an AR(∞) model can represent any time series. The true model for an AR(p) process in white noise can be found in the following manner. Let $P_x(z)$ denote the z-transform of the autocorrelation function of $x(n)$. Then it follows from Eq. (2.92) that

$$\begin{aligned} P_y(z) &= P_x(z) + P_w(z) \\ &= \frac{\sigma^2}{A(z)A(z^{-1})} + \sigma_w^2 \\ &= \frac{\sigma^2 + \sigma_w^2 A(z)A(z^{-1})}{A(z)A(z^{-1})} \end{aligned} \tag{2.93}$$

It can be shown [15] that the numerator of Eq. (2.93) may be factored as

$$\sigma^2 + \sigma_w^2 A(z)A(z^{-1}) = \sigma_\eta^2 B(z)B(z^{-1}) \tag{2.94}$$

where $B(z) = 1 + \sum_{k=1}^{p} b_k z^{-k}$ has its zeros inside the unit circle and $\sigma_\eta^2 > 0$. Hence,

$$P_y(z) = \frac{\sigma_\eta^2 B(z)B(z^{-1})}{A(z)A(z^{-1})}$$

We see that $y(n)$ is actually an ARMA(p,p) process with its AR and MA parameters linked according to Eq. (2.94). Estimation of the PSD of $y(n)$ should be based on the appropriate ARMA model. This is an important consideration in that many time series of interest are actually composed of signals in noise.

2.5.6 Estimation of AR Power Spectral Densities

For good estimates of the AR parameters an MLE is usually employed (see Section 2.3.2). Assuming $x(n)$ is a Gaussian random process, we now show that an approximate MLE of the AR parameters is found by solving a set of simultaneous linear equations. This estimator is identical in form to the covariance method for signal modeling, which was discussed in Chapter 1. Assuming that $\mathbf{x} = [x(0)x(1) \cdots x(N-1)]^T$ is observed, we wish to estimate $\mathbf{a} = [a_1 a_2 \cdots a_p]^T$ and σ^2. The MLE

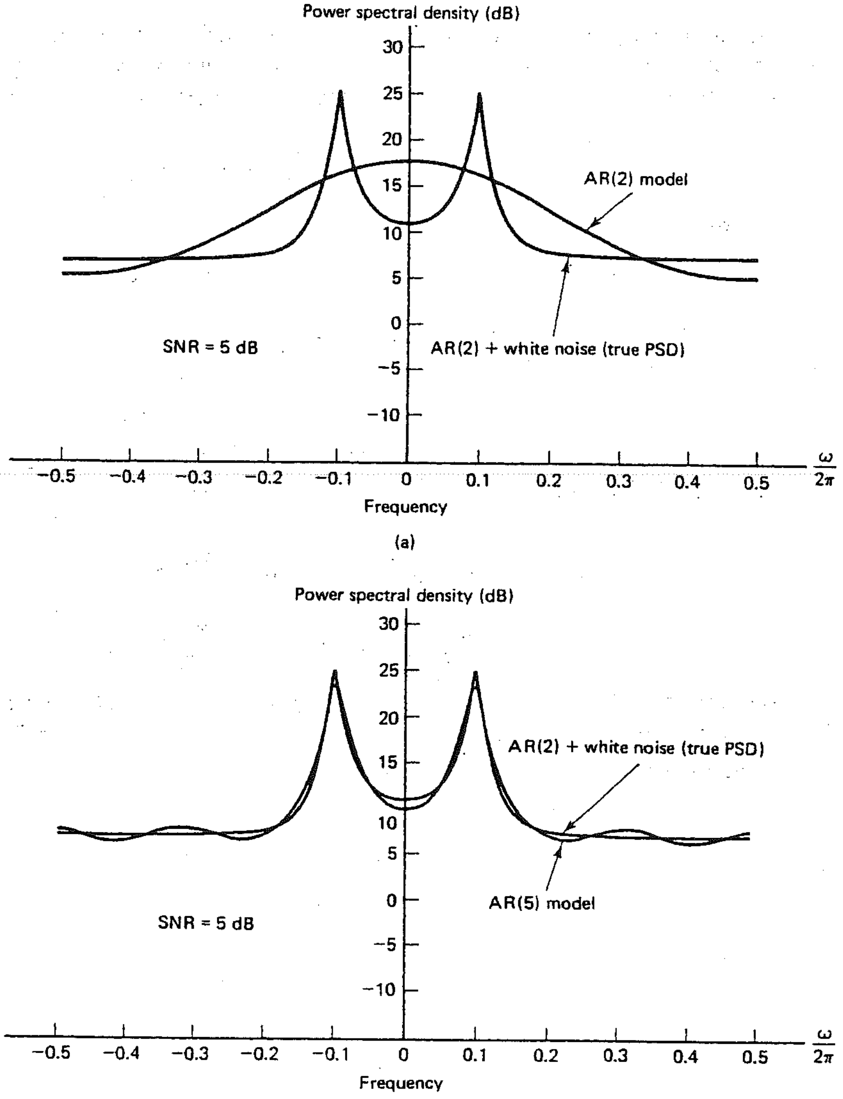


Figure 2.15 (a) AR(2) modeling of noise-corrupted AR(2) power spectral density; (b) AR(5) modeling of noise-corrupted AR(2) power spectral density.

is found by maximizing the joint PDF $p(\mathbf{x}; \mathbf{a}, \sigma^2)$ over $\mathbf{a}$ and σ^2 when $\mathbf{x}$ is replaced by the observed data samples. The PDF can also be written conditioned on the first p data samples as

$$p(x(p), x(p+1), \ldots, x(N-1) \mid x(0), x(1), \ldots, x(p-1); \mathbf{a}, \sigma^2) \\ p(x(0), x(1), \ldots, x(p-1); \mathbf{a}, \sigma^2) \qquad (2.95)$$

For large data records the maximization of the PDF can be effected by maximizing only the conditional PDF in Eq. (2.95) as long as the poles are not too close to the unit circle [16]. We thus seek to maximize the conditional PDF. From Eq. (2.74),

$$\epsilon(n) = x(n) + \sum_{k=1}^{p} a_k x(n-k)$$

or, for $n = p, p+1, \ldots, N-1$,

$$\begin{aligned} \epsilon(p) &= x(p) + a_1 x(p-1) + \cdots + a_p x(0) \\ \epsilon(p+1) &= x(p+1) + a_1 x(p) + \cdots + a_p x(1) \\ &\vdots \\ \epsilon(N-1) &= x(N-1) + a_1 x(N-2) + \cdots + a_p x(N-p-1) \end{aligned} \tag{2.96}$$

Noting that the $\epsilon(n)$'s are uncorrelated Gaussian random variables and hence independent, the PDF of $\boldsymbol{\epsilon} = [\epsilon(p)\epsilon(p+1)\cdots\epsilon(N-1)]^T$ is

$$\begin{aligned} p(\boldsymbol{\epsilon}) &= \prod_{n=p}^{N-1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\epsilon^2(n)\right] \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=p}^{N-1}\epsilon^2(n)\right] \end{aligned} \tag{2.97}$$

We now transform the PDF of $\{\epsilon(p), \epsilon(p+1), \ldots, \epsilon(N-1)\}$ to the PDF of $\{x(p), x(p+1), \ldots, x(N-1)\}$ by using the transformation of Eq. (2.96). The transformation may be rewritten as

$$\boldsymbol{\epsilon} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ a_1 & 1 & 0 & 0 & \cdots & 0 \\ a_2 & a_1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & a_p & a_{p-1} & \cdots & 1 \end{bmatrix}}_{\mathbf{J}} \mathbf{x}' + \begin{bmatrix} \sum_{i=1}^{p} a_i x(p-i) \\ \sum_{i=2}^{p} a_i x(p+1-i) \\ \vdots \\ a_p x(p-1) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $\mathbf{x}' = [x(p)x(p+1)\cdots x(N-1)^T$. The Jacobian of the transformation $\partial\boldsymbol{\epsilon}/\partial\mathbf{x}'$ is just $\mathbf{J}$ and hence $|\det(\mathbf{J})| = 1$. Then, using Eq. (2.97), we have

$$\begin{aligned} p(\mathbf{x}'|x(0), x(1), \ldots, x(p-1); \mathbf{a}, \sigma^2) &= p(\boldsymbol{\epsilon}(\mathbf{x}'))\left|\det\left(\frac{\partial\boldsymbol{\epsilon}}{\partial\mathbf{x}'}\right)\right| = p(\boldsymbol{\epsilon}(\mathbf{x}')) \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=p}^{N-1}\left[x(n) + \sum_{j=1}^{p} a_j x(n-j)\right]^2\right] \end{aligned} \tag{2.98}$$

To maximize Eq. (2.98) over **a** we need only minimize

$$S_1(\mathbf{a}) = \sum_{n=p}^{N-1} \left[x(n) + \sum_{j=1}^{p} a_j x(n-j) \right]^2 \tag{2.99}$$

Since S_1 is a quadratic function of **a**, differentiating Eq. (2.99) will produce a global minimum. Performing the differentiation, we have

$$\sum_{j=1}^{p} \hat{a}_j \sum_{n=p}^{N-1} x(n-j)x(n-k) = -\sum_{n=p}^{N-1} x(n)x(n-k), \quad k = 1, 2, \ldots, p \tag{2.100}$$

or in matrix form the MLE of **a** is found by solving

$$\underbrace{\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix}}_{\hat{\mathbf{a}}} = -\underbrace{\begin{bmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{p0} \end{bmatrix}}_{\mathbf{c}} \tag{2.101}$$

where

$$c_{jk} = \frac{1}{N-p} \sum_{n=p}^{N-1} x(n-j)x(n-k)$$

The factor $1/(N-p)$ has been introduced by dividing both sides of Eq. (2.100) by $N-p$. In this form c_{jk} is seen to be an estimate of $r_x(j-k)$ and hence *the MLE is obtained by replacing the true ACF in the Yule-Walker equations of Eq. (2.80) by suitable estimates*. It should be noted that **C** is symmetric and positive semidefinite. To solve Eq. (2.101) we may use a Cholesky decomposition or the fast solution for the covariance equations as discussed in Chapter 1.

To find $\hat{\sigma}^2$ we first substitute $\hat{\mathbf{a}}$ into Eq. (2.98) and then differentiate with respect to σ^2. Equivalently, we may take the logarithm of Eq. (2.98) since it is a monotonic function and differentiate to yield

$$\frac{\partial p}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left[-\frac{N-p}{2}\ln 2\pi - \frac{N-p}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}S_1(\hat{\mathbf{a}}) \right] = 0$$

which yields

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{N-p} S_1(\hat{\mathbf{a}}) \\ &= \frac{1}{N-p} \sum_{n=p}^{N-1} \left[x(n) + \sum_{j=1}^{p} \hat{a}_j x(n-j) \right]^2 \end{aligned} \tag{2.102}$$

An alternate expression for $\hat{\sigma}^2$ is found by observing from Eq. (2.100) that

$$\sum_{n=p}^{N-1} \left[x(n) + \sum_{j=1}^{p} \hat{a}_j x(n-j) \right] \sum_{k=1}^{p} \hat{a}_k x(n-k) = 0$$

Hence, Eq. (2.102) becomes

$$\hat{\sigma}^2 = \frac{1}{N-p}\sum_{n=p}^{N-1} x^2(n) + \sum_{j=1}^{p} \hat{a}_j \frac{1}{N-p}\sum_{n=p}^{N-1} x(n)x(n-j)$$

or, finally,

$$\hat{\sigma}^2 = c_{00} + \sum_{j=1}^{p} \hat{a}_j c_{0j} \tag{2.103}$$

The estimation of the AR parameters via Eqs. (2.101) and (2.103) is called the *covariance method*. Once the estimates have been computed, the PSD estimate is given by

$$\hat{P}_{\text{AR}}(\omega) = \frac{\hat{\sigma}^2}{|1 + \hat{a}_1 \exp(-j\omega) + \cdots + \hat{a}_p \exp(-j\omega p)|^2} \tag{2.104}$$

Note that the approximate MLE was obtained by minimizing a sum of squared errors given by Eq. (2.99). If we consider $-\Sigma_{j=1}^{p} a_j x(n-j)$ as a linear prediction of $x(n)$ based on $\{x(n-1), x(n-2), \ldots, x(n-p)\}$ and $x(n) - [-\Sigma_{j=1}^{p} a_j x(n-j)]$ as a prediction error, then it may be said that the MLE of the AR parameters is found by finding the best pth-order linear predictor of $x(n)$ based on the previous p samples. In this manner, the name *linear prediction* is often associated with AR spectral estimation [3].

In addition to the covariance method, there are numerous other techniques that also are approximate MLEs. As an example it is easily shown that if $N \gg p$, then

$$c_{jk} \approx \hat{r}_x(j-k)$$

where $\hat{r}_x(k)$ is given by Eq. (2.42). When this autocorrelation function estimator is substituted into Eq. (2.101), a set of Yule-Walker equations results, with the theoretical autocorrelation function replaced by the biased autocorrelation function estimator. This approach is termed the *autocorrelation method* (see Chapter 1). Some other methods that have been found to work well are the forward-backward or modified covariance method [17,18] and the Burg method [2]. We will summarize these estimators. For further details see Chapter 1.

In the forward-backward method we minimize

$$S_2(\mathbf{a}) = \sum_{n=p}^{N-1}\left[x(n) + \sum_{j=1}^{p} a_j x(n-j)\right]^2 + \sum_{n=0}^{N-1-p}\left[x(n) + \sum_{j=1}^{p} a_j x(n+j)\right]^2 \tag{2.105}$$

which is the sum of forward and backward prediction error energies. The minimization proceeds analogously to that of the covariance method with the results given by Eq. (2.101) and Eq. (2.103) but with c_{jk} defined as

$$c_{jk} = \frac{1}{2(N-p)}\left[\sum_{n=p}^{N-1} x(n-j)x(n-k) + \sum_{n=0}^{N-1-p} x(n+j)x(n+k)\right] \tag{2.106}$$

One can show that C for the forward-backward method is symmetric and positive semidefinite. The inversion of C can be avoided by using a Cholesky decomposition or the recursive algorithm in [19].

The second method, the Burg algorithm, makes use of the Levinson algorithm and an estimate for the *reflection coefficients,* which are defined as $k_i = a_{ii}$ in Eq. (2.81). The Burg algorithm is initialized by

$$\begin{aligned} e_0(n) &= x(n) \\ b_0(n) &= x(n) \\ \sigma_0^2 &= \frac{1}{N}\sum_{n=0}^{N-1} x^2(n) \end{aligned}$$

For $i = 1, 2, \ldots, p,$

$$k_i = \frac{-2\sum_{n=i}^{N-1} e_{i-1}(n)b_{i-1}(n-1)}{\sum_{n=i}^{N-1} [e_{i-1}^2(n) + b_{i-1}^2(n-1)]}$$

$$a_{ij} = \begin{cases} k_1 & \text{for } i = j = 1 \\ a_{i-1,j} + k_i a_{i-1,i-j} & \text{for } j = 1, 2, \ldots, i-1; \\ & \quad i = 2, 3, \ldots, p \end{cases} \tag{2.107}$$

$$\sigma_i^2 = (1 - k_i^2)\sigma_{i-1}^2$$

$$e_i(n) = e_{i-1}(n) + k_i b_{i-1}(n-1)$$

$$b_i(n) = b_{i-1}(n-1) + k_i e_{i-1}(n)$$

The estimates of the AR parameters are $\hat{a}_i = a_{pi}$, $i = 1, 2, \ldots, p$, and $\hat{\sigma}^2 = \sigma_p^2$.

2.5.7 Estimation of MA Power Spectral Densities

The MA spectral estimator was defined to be

$$P_{\text{MA}}(\omega) = \sigma^2 \left| 1 + \sum_{k=1}^{q} b_k \exp(-j\omega k) \right|^2 \tag{2.108}$$

This can also be written as

$$P_{\text{MA}}(\omega) = \sum_{k=-q}^{q} r_x(k) \exp(-j\omega k) \tag{2.109}$$

by expanding the factors in Eq. (2.108) with the autocorrelation function $r_x(k)$ given by Eq. (2.86). A natural estimator of $P_{\text{MA}}(\omega)$ would then seem to be

$$\hat{P}_{\text{MA}}(\omega) = \sum_{k=-q}^{q} \hat{r}_x(k) \exp(-j\omega k) \tag{2.110}$$

where $\hat{r}_x(k)$ is some suitable estimator of the autocorrelation function. The MA spectral estimator bears a strong resemblance to the Blackman-Tukey spectral estimator. A subtle difference between the two estimators is that the MA spectral estimator is based on the MA(q) model and hence by assumption $\hat{r}_x(k) = 0$ for $|k| > q$. The Blackman-Tukey spectral estimator, on the other hand, can be applied to any process. Furthermore, in the Blackman-Tukey spectral estimator the autocorrelation function estimator is truncated at $k = M$ (as well as weighted) due to a finite-length data record.

To estimate the MA PSD we need to find the MLE of the MA parameters. Durbin has derived them for large data records. The exact derivation [20,14] relies heavily on advanced statistical theory and so will not be presented here. Instead we offer an intuitive justification for Durbin's method. The PSD of an MA process generalized to the z-plane is

$$P_{\mathrm{MA}}(z) = \sigma^2 B(z) B(z^{-1})$$

We can approximate $B(z)$ by $1/A(z)$, where $A(z) = 1 + \Sigma_{k=1}^{L} a_k z^{-k}$ to yield

$$P_{\mathrm{MA}}(z) = \frac{\sigma^2}{A(z)A(z^{-1})} \tag{2.111}$$

As an example, consider an MA(1) process. Then

$$\begin{aligned} B(z) &\approx \frac{1}{A(z)} \\ 1 + b_1 z^{-1} &\approx \frac{1}{A(z)} \end{aligned} \tag{2.112}$$

or

$$A(z) \approx \frac{1}{1 + b_1 z^{-1}} = \sum_{k=0}^{\infty} (-b_1)^k z^{-k}$$

If we choose $a_k = (-b_1)^k$ for $k = 1, 2, \ldots, L$, and if $(-b_1)^k \approx 0$ for $k > L$, then the approximation will be a good one. Clearly, the approximation will be better when the zero of $B(z)$ is not near the unit circle or when $|b_1|$ is small. If L can be chosen such that the approximation is a good one, then the MA(q) process is equivalent to an AR(L) process. Consequently, the MLE of the AR parameters can be found using any of the methods of Section 2.5.6. Call these estimates $\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_L, \hat{\sigma}^2\}$. To find the estimates of the b_k's, note from Eq. (2.112) that

$$B(z)B(z^{-1}) \approx \frac{1}{A(z)A(z^{-1})}$$

or

$$A(z)A(z^{-1}) \approx \frac{1}{B(z)B(z^{-1})} \tag{2.113}$$

The right-hand side of Eq. (2.113) represents the PSD of an AR(q) process. To estimate the b_k's we need only estimate the autocorrelation lags for $0 \leq k \leq q$, assuming we use the autocorrelation method of AR parameter estimation. But these lag estimates can be found from Eq. (2.113) by taking the inverse z-transform of $\hat{A}(z)\hat{A}(z^{-1})$, or

$$\hat{r}_a(k) = \frac{1}{L+1} \sum_{n=0}^{L-|k|} \hat{a}_n \hat{a}_{n+|k|} \tag{2.114}$$

The $1/(L+1)$ scale factor has been added to allow $\hat{r}_a(k)$ to be interpreted as an autocorrelation function estimator.

Durbin's method can be summarized as follows.

1. Fit a large-order AR model to the data. Specifically, using the Levinson algorithm solve

$$\begin{bmatrix} \hat{r}_x(0) & \hat{r}_x(1) & \cdots & \hat{r}_x(L-1) \\ \hat{r}_x(1) & \hat{r}_x(0) & \cdots & \hat{r}_x(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_x(L-1) & \hat{r}_x(L-2) & \cdots & \hat{r}_x(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_L \end{bmatrix} = - \begin{bmatrix} \hat{r}_x(1) \\ \hat{r}_x(2) \\ \vdots \\ \hat{r}_x(L) \end{bmatrix} \tag{2.115}$$

where

$$\hat{r}_x(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n)x(n+k)$$

The estimate of σ^2 is given by

$$\hat{\sigma}^2 = \hat{r}_x(0) + \sum_{k=1}^{L} \hat{a}_k \hat{r}_x(k) \tag{2.116}$$

2. Using the data sequence $\{1, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_L\}$, fit an AR($q$) model. As before, solve

$$\begin{bmatrix} \hat{r}_a(0) & \hat{r}_a(1) & \cdots & \hat{r}_a(q-1) \\ \hat{r}_a(1) & \hat{r}_a(0) & \cdots & \hat{r}_a(q-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_a(q-1) & \hat{r}_a(q-2) & \cdots & \hat{r}_a(0) \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_q \end{bmatrix} = - \begin{bmatrix} \hat{r}_a(1) \\ \hat{r}_a(2) \\ \vdots \\ \hat{r}_a(q) \end{bmatrix} \tag{2.117}$$

where

$$\hat{r}_a(k) = \frac{1}{L+1} \sum_{n=0}^{L-k} \hat{a}_n \hat{a}_{n+k}$$

The estimation of the AR parameters in steps 1 and 2 may be effected by any of the AR estimation methods. The choice of the autocorrelation method allows a simple solution to the linear equations via the Levinson algorithm.

The only difficulty in using Durbin's method is in choosing L. From the intuitive justification it is clear that L should be equal to the effective impulse response length of $1/B(z)$. Unfortunately, this is unknown *a priori* since it is the MA parameters we wish to estimate. Also, we require $L \ll N$ for the AR parameter estimates to be statistically accurate.

2.5.8 Estimation of ARMA Power Spectral Densities

The ARMA(p, q) spectral estimator was defined in Section 2.5.1 to be

$$\hat{P}_{\text{ARMA}}(\omega) = \frac{\hat{\sigma}^2 \left| 1 + \sum_{k=1}^{q} \hat{b}_k \exp(-j\omega k) \right|^2}{\left| 1 + \sum_{k=1}^{p} \hat{a}_k \exp(-j\omega k) \right|^2} \tag{2.118}$$

For reliable estimates of the ARMA parameters we would once again like to use an MLE. Unfortunately, in this case the MLE is exceedingly difficult to obtain. Even with several simplifying approximations to the PDF *the equations obtained by differentiating are extremely nonlinear*. To illustrate the problem we will examine the function that needs to be minimized. First we return to the problem of AR estimation. There we found that the function we needed to minimize was

$$S_1(\mathbf{a}) = \sum_{n=p}^{N-1} \left[x(n) + \sum_{j=1}^{p} a_j x(n-j) \right]^2 \tag{2.119}$$

If we consider $x(n) = 0$ outside of the $0 \leq n \leq N - 1$ interval and replace the limits of the summation with $n = 0$ to $n = N + p - 1$, which for N large will not significantly alter the sum, then

$$S_1(\mathbf{a}) = \sum_{n=-\infty}^{\infty} \left[x(n) + \sum_{j=1}^{p} a_j x(n-j) \right]^2 \tag{2.120}$$

Note that $S_1(\mathbf{a})$ represents the energy out of the inverse or whitening filter $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ for the input $x(n)$. For the true AR parameters, $x(n) + \sum_{j=1}^{p} a_j x(n-j)$ becomes $\epsilon(n)$, or we have whitened the $x(n)$ time series. It is not surprising then that for an ARMA process the approximate MLE is found by finding the inverse filter $A(z)/B(z)$, which whitens the ARMA time series. Whitening can also be shown to be equivalent to minimizing the output energy out of the filter. An explicit expression for the energy for the ARMA case can be obtained by transforming to the frequency domain. First for an AR process Eq. (2.120) can be rewritten using Parseval's theorem as

$$S_1(\mathbf{a}) = N \int_{-\pi}^{\pi} I(\omega) |A(\omega)|^2 \frac{d\omega}{2\pi}$$

where

$$I(\omega) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x(n) \exp(-j\omega n) \right|^2$$

which is just the periodogram. The different notation is meant to emphasize that $I(\omega)$ is to be regarded as a function of the data and not necessarily as a spectral estimator. Likewise, the approximate MLE of the ARMA parameters is found by minimizing

$$S_2(\mathbf{a}, \mathbf{b}) = N \int_{-\pi}^{\pi} I(\omega) \frac{|A(\omega)|^2}{|B(\omega)|^2} \frac{d\omega}{2\pi} \tag{2.121}$$

As an example, consider the ARMA(1, 1) case. Then, differentiating Eq. (2.121) to obtain a set of necessary conditions for the approximate MLE, we have

$$\begin{aligned} \frac{\partial S_2}{\partial a_1} &= N \int_{-\pi}^{\pi} \frac{I(\omega)}{|B(\omega)|^2} \frac{\partial}{\partial a_1} \{[1 + a_1 \exp(-j\omega)][1 + a_1 \exp(-j\omega)]\} \frac{d\omega}{2\pi} \\ &= N \int_{-\pi}^{\pi} \frac{I(\omega)}{|B(\omega)|^2} \{[1 + a_1 \exp(-j\omega)] \exp(j\omega) \\ &\qquad + \exp(-j\omega)[1 + a_1 \exp(j\omega)]\} \frac{d\omega}{2\pi} \qquad (2.122) \\ &= N \int_{-\pi}^{\pi} \frac{I(\omega)}{|B(\omega)|^2} (2 \cos \omega + 2a_1) \frac{d\omega}{2\pi} \\ &= 0 \end{aligned}$$

so

$$a_1 = - \frac{\displaystyle\int_{-\pi}^{\pi} \frac{I(\omega) \cos \omega}{|B(\omega)|^2} \frac{d\omega}{2\pi}}{\displaystyle\int_{-\pi}^{\pi} \frac{I(\omega)}{|B(\omega)|^2} \frac{d\omega}{2\pi}} \tag{2.123}$$

In a similar fashion we obtain

$$\frac{\partial S_2}{\partial b_1} = -N \int_{-\pi}^{\pi} \frac{I(\omega)|A(\omega)|^2}{|B(\omega)|^4} (2 \cos \omega + 2b_1) \frac{d\omega}{2\pi} = 0 \tag{2.124}$$

Substituting Eq. (2.123) for a_1 into Eq. (2.124) results in a very nonlinear equation in b_1. In general, we would obtain a set of nonlinear equations that need to be solved using iterative techniques. The standard Newton-Raphson technique has been proposed for this problem but suffers from convergence problems; furthermore, even if convergence is obtained, the solution may correspond to only a local and not a global minimum [21,14].

Due to the difficulty of computing the approximate MLE, several suboptimal approaches have been proposed. They rely on the theoretical relationship between the autocorrelation function and the ARMA parameters as embodied in the Yule-Walker equations. Recall from Eq. (2.78) that

$$r_x(k) = -\sum_{\ell=1}^{p} a_\ell r_x(k - \ell), \qquad k \geq q + 1 \tag{2.125}$$

If we rewrite this in matrix notation for $k = q + 1, q + 2, \ldots, q + p$, then

$$\underbrace{\begin{bmatrix} r_x(q) & r_x(q-1) & \cdots & r_x(q-p+1) \\ r_x(q+1) & r_x(q) & \cdots & r_x(q-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(q+p-1) & r_x(q+p-2) & \cdots & r_x(q) \end{bmatrix}}_{\mathbf{R}_x'} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = -\begin{bmatrix} r_x(q+1) \\ r_x(q+2) \\ \vdots \\ r_x(q+p) \end{bmatrix} \tag{2.126}$$

These equations are called the *modified Yule-Walker equations* and can be used to estimate the AR parameters of an ARMA(p, q) model if the theoretical autocorrelation lags are replaced by estimates. It is termed the modified Yule-Walker equation estimator. Once the AR parameter estimates have been obtained, the remaining parameters $\{b_1, b_2, \ldots, b_q, \sigma^2\}$ can be found as follows. Filter $x(n)$ with $\hat{A}(z)$, where $\hat{A}(z) = 1 + \sum_{k=1}^{p} \hat{a}_k z^{-k}$ and the $\hat{a}_k$'s have been found from Eq. (2.126). If $\hat{A}(z) \approx A(z)$, then the filter output will be very nearly an MA(q) process with parameters $\{b_1, b_2, \ldots, b_q, \sigma^2\}$. Now use Durbin's method on the output sequence to estimate the remaining parameters. Due to the memory of the $\hat{A}(z)$ filter the output sequence should be used only for $n = p, p+1, \ldots, N-1$.

The use of the modified Yule-Walker equations followed by Durbin's method produces good results at times. The technique can, however, yield highly variable spectral estimates, especially when the matrix $\mathbf{R}_x'$ is nearly singular [22]. The problem may be likened to attempting to fit a straight line through two data points that are subject to error. To overcome this deficiency it is better to use more than p equations. Since there will now be more equations than unknowns, we solve for the AR parameters in a least-squares sense, as explained in Section 2.3.2. Choosing the equations in Eq. (2.125) corresponding to $k = q+1, q+2, \ldots, M$ and rewriting in matrix notation, we have

$$\begin{bmatrix} r_x(q+1) \\ r_x(q+2) \\ \vdots \\ r_x(M) \end{bmatrix} = -\begin{bmatrix} r_x(q) & r_x(q-1) & \cdots & r_x(q-p+1) \\ r_x(q+1) & r_x(q) & \cdots & r_x(q-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(M-1) & r_x(M-2) & \cdots & r_x(M-p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{2.127}$$

By replacing $r_x(k)$ by an estimate, Eq. (2.127) will no longer hold exactly. An error term will need to be introduced so that Eq. (2.127) becomes

$$\begin{bmatrix} \hat{r}_x(q+1) \\ \hat{r}_x(q+2) \\ \vdots \\ \hat{r}_x(q+M) \end{bmatrix} = \begin{bmatrix} -\hat{r}_x(q) & -\hat{r}_x(q-1) & \cdots & -\hat{r}_x(q-p+1) \\ -\hat{r}_x(q+1) & -\hat{r}_x(q) & \cdots & -\hat{r}_x(q-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{r}_x(M-1) & -\hat{r}_x(M-2) & \cdots & -\hat{r}_x(M-p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} e(q+1) \\ e(q+2) \\ \vdots \\ e(M) \end{bmatrix} \tag{2.128}$$

But this is exactly in the form of Eq. (2.27), or

$$\mathbf{r} = \mathbf{R}\mathbf{a} + \mathbf{e} \tag{2.129}$$

so that a least-squares estimator of **a** would minimize

$$S_3(\mathbf{a}) = (\mathbf{r} - \mathbf{R}\mathbf{a})^T(\mathbf{r} - \mathbf{R}\mathbf{a})$$

The solution that is found from Eq. (2.29) is

$$\hat{\mathbf{a}} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{r} \tag{2.130}$$

It has been found empirically that the estimator given by Eq. (2.130), which is termed the *least-squares modified Yule-Walker equation* (LSMYWE) estimator, yields better estimates than the modified Yule-Walker equation estimator. Neither technique, however, works well when the poles are sufficiently displaced from the unit circle. This is because the autocorrelation function damps out rapidly, leading to autocorrelation lag estimates dominated by the estimation error. It should be noted that the LSMYWE is not optimal in any reasonable sense. The relationship of the least-squares estimator with the MLE is not valid in this case since **R** is not a constant matrix but is random, and **e** is not distributed according to $N(\mathbf{0}, \sigma^2\mathbf{I})$. The LSMYWE minimizes the error in a theoretical equation, i.e., Eq. (2.127), when estimates are used. For this reason the technique is also referred to as an *equation error modeling* approach. Finally, once the AR parameter estimates have been obtained, Durbin's method can be used to estimate the remaining parameters based on the data at the output of $\hat{A}(z)$.

A difficulty of the LSMYWE is in choosing M. M should be chosen large to take advantage of the information in the higher-order lags but not so large that the autocorrelation function estimates are unreliable. Unless one has some *a priori* knowledge about the pole positions, this choice is difficult and is a problem that has not been resolved.

2.5.9 Model Order Determination

So far we have not addressed the question of how one actually chooses the appropriate model order for the time series model. The best way to make this decision is to base it on the physics of the process generating the data. For instance, in speech processing it is known that the vocal tract can be modeled as an all-pole filter having about 4 resonances in a 4-Khz band [3]. Hence, at least 8 complex conjugate poles are necessary and typically $p = 12$ is chosen in an AR model. When no such information is available, statistical tests can be used to estimate order. One such test is the Akaike information criterion (AIC) [23]. The AIC computes a measure over all possible model orders and chooses the model order that minimizes the measure. For an ARMA(p, q) process the AIC is defined as

$$\text{AIC}(i,j) = N \ln \hat{\sigma}^2_{ij} + 2(i + j) \tag{2.131}$$

where $\hat{\sigma}^2_{ij}$ is the MLE of the white noise variance for an assumed ARMA(i, j) process. The AIC attempts to balance the modeling error (or bias) that is manifested by $\hat{\sigma}^2_{ij}$ and that generally decreases with increasing model order and the need to maintain a small number of model parameters to be estimated (or variance), which is embodied in the

$2(i + j)$ term and which increases with increasing model order. In practice $\hat{\sigma}_{ij}^2$ is usually any good estimate of the white noise variance. For an AR or MA process the AIC is defined as

$$\mathrm{AIC}(i) = N \ln \hat{\sigma}_i^2 + 2i \tag{2.132}$$

where i is the assumed AR or MA model order. As an example consider the AR(2) process given by

$$x(n) = 1.34x(n-1) - 0.9025x(n-2) + \epsilon(n)$$

To estimate the AR parameters we use the Burg algorithm, so

$$\hat{\sigma}_i^2 = (1 - k_i^2)\hat{\sigma}_{i-1}^2$$

according to Eq. (2.107). Thus, the AIC becomes

$$\begin{aligned}\mathrm{AIC}(i) &= N \ln[(1 - k_i^2)\hat{\sigma}_{i-1}^2] + 2i \\ &= \mathrm{AIC}(i-1) + N \ln(1 - k_i^2) + 2\end{aligned}$$

Note that the term $N \ln(1 - k_i^2)$ causes the AIC to decrease with increasing i since $k_i^2 < 1$ while the constant term 2 causes it to increase with increasing i. For $N = 100$ the AIC yields the curve in Fig. 2.16. We can see that the minimum is attained at $i = 4$ even though the true model order is $p = 4$. This tendency to overestimate the true model order is characteristic of the AIC.

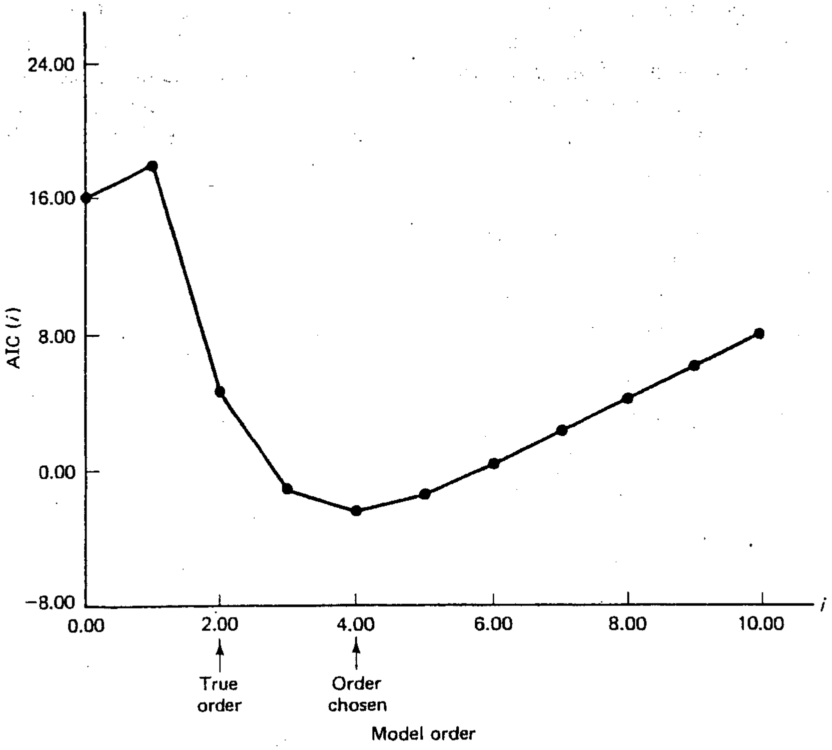


Figure 2.16 Example of Akaike information criterion in model order selection.

2.5.10 Maximum Entropy Spectral Estimation

Maximum entropy spectral estimation (MESE) is based on an explicit extrapolation of a segment of a known autocorrelation function for the samples that are not known [2]. In this way the characteristic smearing of the estimated PSD due to truncation of the autocorrelation function is alleviated. If $\{r_x(0), r_x(1), \ldots, r_x(p)\}$ is known, then the question arises as to how $\{r_x(p+1), r_x(p+2), \ldots\}$ should be specified in order to guarantee that the entire autocorrelation function is valid or that its Fourier transform is nonnegative. In general, there are an infinite number of possible extrapolations, all of which yield valid autocorrelation functions. In the MESE it is argued that the extrapolation should be chosen so that the time series characterized by the extrapolated autocorrelation function has maximum entropy. The time series will then be the most random one that has the known autocorrelation samples for its first $p + 1$ lag values. Alternatively, the PSD will be the one with the flattest (whitest) spectrum of all spectra for which the first $p + 1$ autocorrelation samples are equal to the known ones. The resultant spectral estimator is termed the MESE. The rationale for choosing the maximum entropy criterion is that it imposes the fewest constraints on the unknown time series by maximizing its randomness, thereby producing a minimum bias solution.

In particular, if one assumes a Gaussian random process, then the entropy per sample is proportional to

$$\int_{-\pi}^{\pi} \ln P_x(\omega) d\omega \tag{2.133}$$

The MESE is found by maximizing Eq. (2.133) subject to the constraints that the autocorrelation function corresponding to $P_x(\omega)$ has as its first $p + 1$ lags the known samples of the autocorrelation function, or

$$\int_{-\pi}^{\pi} P_x(\omega) \exp(j\omega k) \frac{d\omega}{2\pi} = r_x(k) \qquad \text{for } k = 0, 1, \ldots, p \tag{2.134}$$

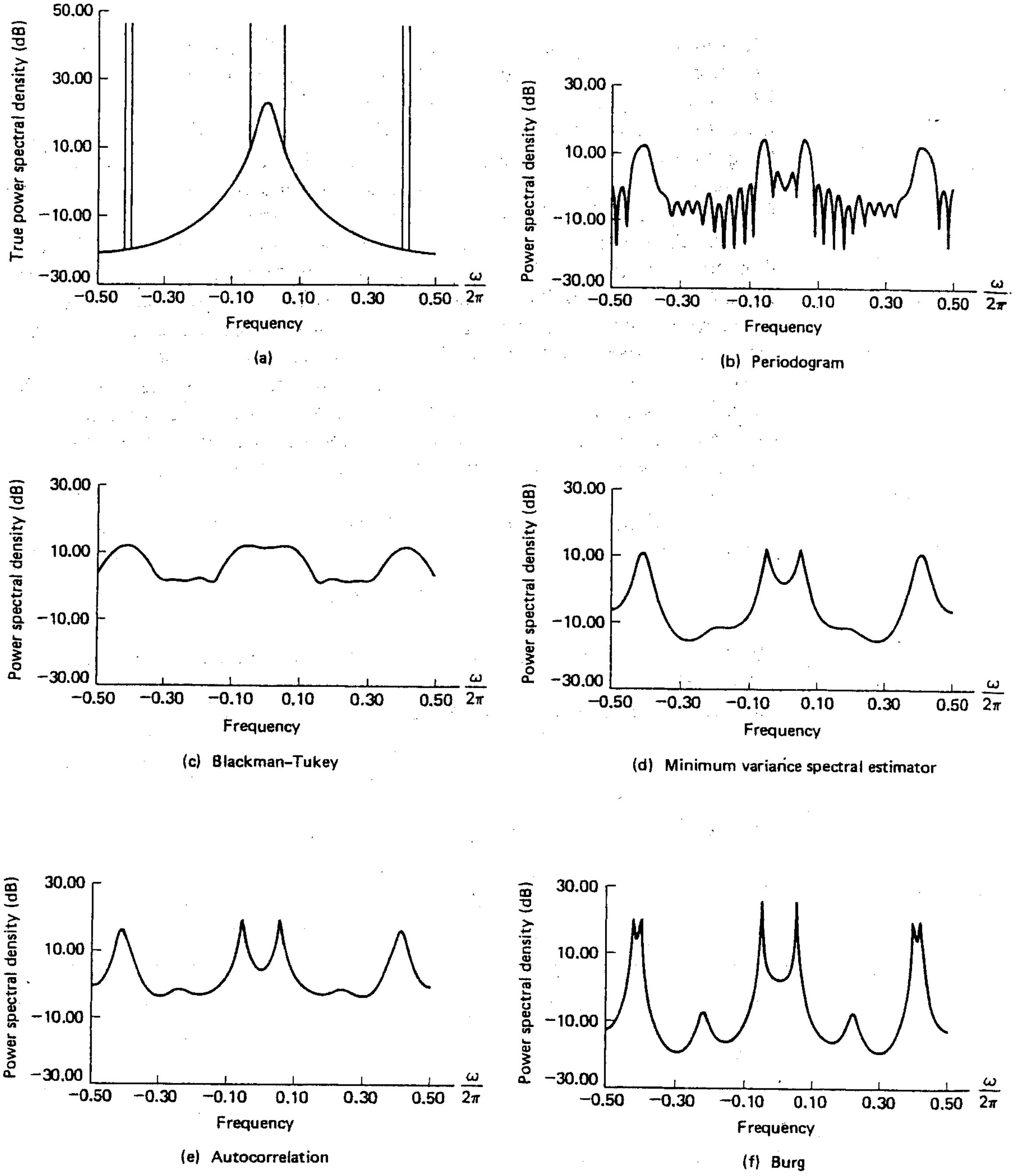
The solution that results from applying the technique of Lagrangian multipliers is

$$P_x(\omega) = \frac{\sigma^2}{\left| 1 + \sum_{k=1}^{p} a_k \exp(-j\omega k) \right|^2} \tag{2.135}$$

where $\{a_1, a_2, \ldots, a_p, \sigma^2\}$ are found by solving the Yule-Walker equations using the known samples of the autocorrelation function. Hence, with knowledge of $\{r_x(0), r_x(1), \ldots, r_x(p)\}$, the MESE is equivalent to the AR spectral estimator. This equivalence, however, is maintained only for Gaussian random processes and *known* autocorrelation function samples.

2.6 COMPUTER SIMULATION EXAMPLES

In an effort to illustrate the typical characteristics of the various spectral estimators that have been discussed and *not as a means of a quantitative comparison*, a test case was developed. The PSD is shown in Fig. 2.17(a). The process consists of narrowband



(a)

(b) Periodogram

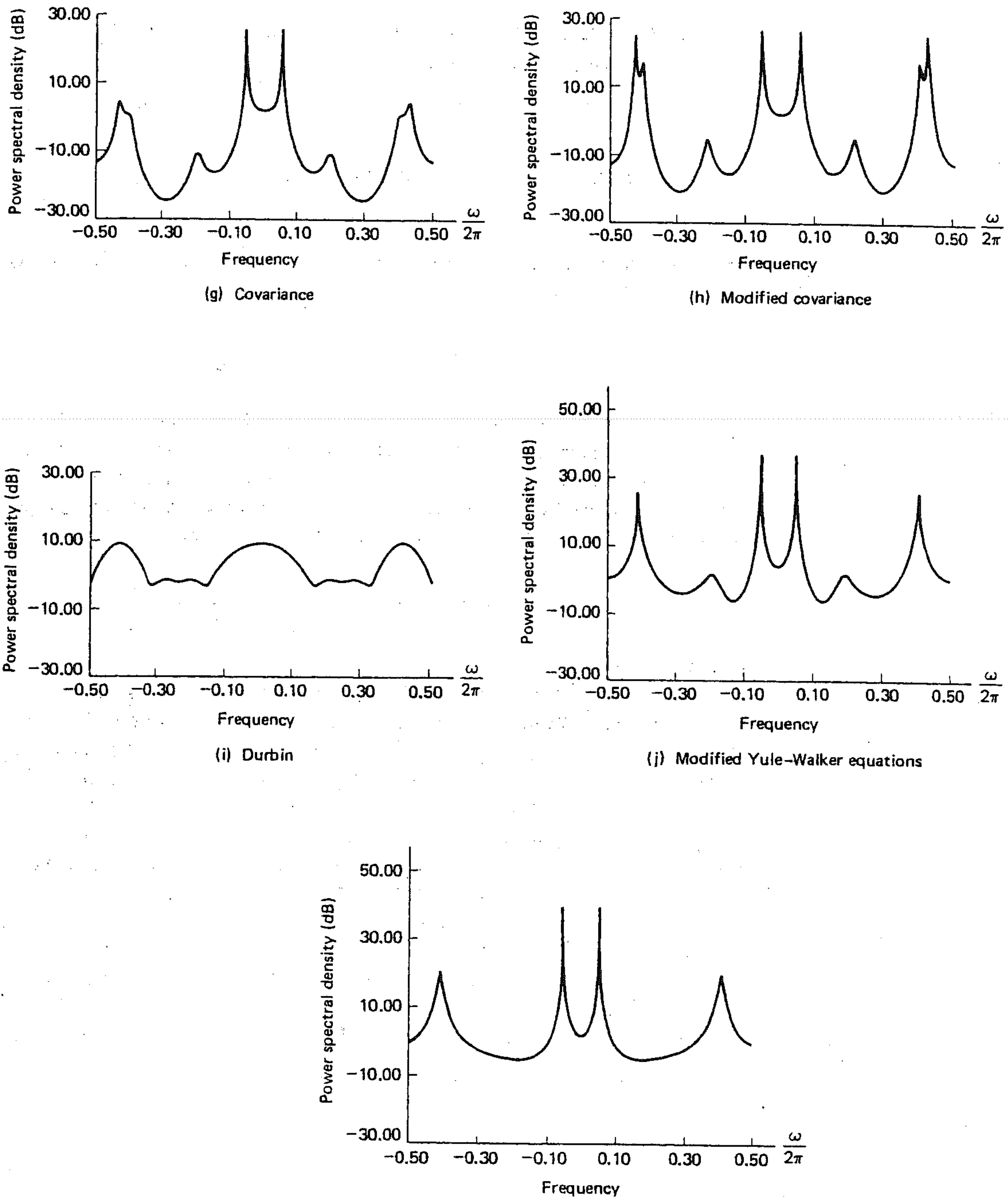
(c) Blackman-Tukey

(d) Minimum variance spectral estimator

(e) Autocorrelation

(f) Burg

Figure 2.17 Computer simulation results for test case data: (a) True power spectral density; (b) periodogram; (c) Blackman-Tukey; (d) MVSE; (e) autocorrelation; (f) Burg; (g) covariance; (h) modified covariance; (i) Durbin; (j) MYWE; (k) LSMYWE.



(g) Covariance

(h) Modified covariance

(i) Durbin

(j) Modified Yule-Walker equations

(k) Least-squares modified Yule-Walker equations

Figure 2.17 (*cont.*)

components or sinusoids as well as a broadband component. Specifically, 32 data points have been generated from the process:

$$x(n) = 2\cos(\omega_1 n) + 2\cos(\omega_2 n) + 2\cos(\omega_3 n) + z(n) \qquad (2.136)$$

for $n = 0, 1, \ldots, 31$. In Eq. (2.136) $\omega_1/2\pi = 0.05$, $\omega_2/2\pi = 0.40$, $\omega_3/2\pi = 0.42$, and $z(n)$ is an AR process of order 1, or

$$z(n) = -a_1 z(n-1) + \epsilon(n) \qquad (2.137)$$

The values of a_1 and σ^2 are -0.85 and 0.1, respectively. The PSD of $z(n)$ is given by

$$P_{zz}(\omega) = \frac{\sigma^2}{|1 + a_1 \exp(-j\omega)|^2} \qquad (2.138)$$

The sinusoidal components at $\omega_2/2\pi = 0.40$ and $\omega_3/2\pi = 0.42$ will not be resolved by a Fourier spectral estimator since their separation is less than the resolution limit of $1/N = 0.03$. The component at $\omega_1/2\pi = 0.05$ is well resolved since it is 0.1 cycle/sample apart from its nearest neighbor. The SNR is defined as the power of any sinusoidal component to the broadband noise power. Specifically, it is given by

$$\text{SNR} = 10\log_{10}\frac{2}{\sigma^2/(1 - a_1^2)}\ \text{dB}$$

and equals 7.4 dB.

Each spectral estimation method, other than the periodogram, is constrained to estimate the same number of parameters so that a qualitative comparison is fair. The exact choices for the model orders, lag windows, etc., were not made to yield the best results but only to illustrate the typical characteristics of each estimator. Other choices will yield similar, although not identical, results.

The Fourier spectral estimates are displayed in Fig. 2.17(b) and (c). As expected, the periodogram is unable to resolve the closely spaced sinusoids, which are less than $1/N = 0.03$ cycle/sample apart, and exhibits the usual sidelobe structure. The Blackman-Tukey estimate in Fig. 2.17(c) was based on a biased autocorrelation estimator for lags $k = 0, 1, \ldots, M = 10$ and employed a Bartlett window. Being a smoothed version of the periodogram (see Eq. 2.56), the spectral estimate displays less detail than the periodogram.

The minimum variance spectral estimate based on an autocorrelation matrix of dimension $p \times p = 11 \times 11$ is shown in Fig. 2.17(d). The closely spaced sinusoids are not discernible. A larger-dimension autocorrelation matrix results in a spectral estimate that resolves the closely spaced sinusoids but that also gives rise to many spurious peaks.

The AR spectral estimates based on a model order of $p = 10$ are shown in Fig. 2.17(e)–(h). At most 10 peaks may be present in the spectral estimates. It is observed that the sinusoidal components are resolved by all the estimators except the autocorrelation method, Fig 2.17(e), which has the lowest resolution. All the methods produce a spurious peak at about $\omega/2\pi = 0.2$ although it is less pronounced for the autocorrelation method (Fig. 2.17e). Note that a peak is visible near $\omega/2\pi = 0.2$ in the periodogram, which may be responsible for the spurious peak observed in the AR spectral estimates. In general, the AR methods are able to resolve closely spaced narrowband components but tend to exhibit peaky spectra even for broadband processes.

The MA spectral estimate based on a model order of $q = 10$ and using Durbin's algorithm with a large AR model order of $L = 15$ is shown in Fig. 2.17(i). It is unable to resolve any of the narrowband spectral components. The choice of the large AR model order L did not significantly affect the spectral estimate. It is observed to be nearly identical to the Blackman-Tukey estimate (Fig. 2.17c). This is also expected since the forms of the PSD are identical (see Section 2.5.7) with only the estimates of the autocorrelation function being different.

The ARMA spectral estimates based on model orders of $p = 7$ and $q = 3$ are shown in Fig. 2.17(j) and (k). Neither the modified Yule-Walker equation spectral estimator (part j) nor the LSMYWE spectral estimator (part k) is able to resolve the spectral components centered about $\omega/2\pi = 0.4$. If the AR model order p is increased, it is possible that the peaks may have been resolved. Both estimates are similar in appearance. The LSMYWE spectral estimator (part k) used $M = 15$ but did not appear to be oversensitive to the number of equations $M - q$ used.

2.7 FURTHER TOPICS

The spectral estimation methods discussed in this chapter are only some of the many approaches that have been proposed. Many other approximate maximum likelihood estimators for AR, MA, and ARMA processes are available. Also, the important problem of estimation of the frequencies of sinusoidal signals in white noise has not been addressed. The spectral estimation methods discussed in this chapter have been adapted to the frequency estimation problem although it can be more appropriately termed a parameter estimation problem. Such algorithms as the Pisarenko method, MUSIC method, the iterative filtering method, and the singular value decomposition or principal component AR method have been widely investigated. All these topics as well as numerous references may be found in [14].

2.8 SUMMARY

The principal methods of spectral estimation are the nonparametric approaches of the periodogram, the Blackman-Tukey, the minimum variance spectral estimators, and the parametric approaches based on the time series models. A general comparison of the various approaches is given in Table 2.2. For the nonparametric approaches a tradeoff must always be effected between the bias and the variance of the estimator because of the finite number of autocorrelation lags that can be estimated. The parametric models do not suffer from a truncated autocorrelation function since the model implicitly extrapolates it. Hence, when the model is accurate, spectral estimates with higher resolution and lower variability are obtained. However, when the model is incorrect, then no amount of data will yield a good spectral estimate.

TABLE 2.2 COMPARISON OF SPECTRAL ESTIMATION APPROACHES

Spectral Estimator Type	Assumed Model	PSD Estimator Form	Representable PSDs	Resolution Capability
Fourier	None	$\frac{1}{N}\left\|\sum_{n=0}^{N-1} x(n)e^{-j\omega n}\right\|^2$ $\sum_{k=-M}^{M} w(k)\hat{r}_x(k)e^{-j\omega k}$	Broadband only	Low
Minimum variance	None	$\frac{p}{\mathbf{e}^H\hat{\mathbf{R}}_x^{-1}\mathbf{e}}$	Broadband or narrowband but not both	Medium
Autoregressive	AR	$\frac{\hat{\sigma}^2}{\left\|1 + \sum_{k=1}^{p} \hat{a}_k e^{-j\omega k}\right\|^2}$	Broadband or narrowband but not both	High
Autoregressive moving average	ARMA	$\frac{\hat{\sigma}^2\left\|1 + \sum_{k=1}^{q} \hat{b}_k e^{-j\omega k}\right\|^2}{\left\|1 + \sum_{k=1}^{p} \hat{a}_k e^{-j\omega k}\right\|^2}$	Broadband or narrowband or both	High
Moving average	MA	$\hat{\sigma}^2\left\|1 + \sum_{k=1}^{q} \hat{b}_k e^{-j\omega k}\right\|^2$	Broadband only	Low

Confidence intervals are available for the Fourier spectral estimators. Unfortunately, the statistics of the minimum variance and parametric spectral estimators are unknown owing to the highly nonlinear relationship between the data and the spectral estimator. Consequently, no such confidence intervals for the parametric spectral estimators are available, making interpretation of the spectral estimate difficult.

In summary, nonparametric approaches are more robust due to the lack of restrictive assumptions, but where appropriate, parametric spectral estimators offer the promise of good spectral estimates even for short data records. The practitioner of spectral estimation will need to intelligently assess the characteristics of the data before applying an appropriate spectral estimation method (or methods).

REFERENCES

1. S. Kay and S. L. Marple, Jr., "Spectrum Analysis: A Modern Perspective," *Proc. IEEE*, Vol. 69, pp. 1380–1419, Nov. 1981.
2. J. P. Burg, "Maximum Entropy Spectral Analysis," Ph.D. Thesis, Dept. Geophysics, Stanford University, Stanford, CA, May 1975.

3. J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, Vol. 63, pp. 561–580, April 1975.
4. M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. II, Macmillan, New York, 1979.
5. F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE*, Vol. 66, pp. 51–83, Jan. 1978.
6. G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications*, Holden-Day, San Francisco, 1968.
7. J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," *Proc. IEEE*, Vol. 57, pp. 1408–1418, Aug. 1969.
8. R. T. Lacoss, "Data Adaptive Spectral Analysis Methods," *Geophysics*, Vol. 36, pp. 661–675, Aug. 1971.
9. R. N. McDonough, "Application of the Maximum-Likelihood and the Maximum-Entropy Method to Array Processing," in *Nonlinear Methods of Spectral Analysis*, S. Haykin, Ed., Springer-Verlag, New York, 1983.
10. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.
11. H. Wold, *A Study in the Analysis of Stationary Time Series*, Almqvist and Wiksell, Uppsala, 1954.
12. A. N. Kolmogorov, "Interpolation and Extrapolation von Stationaren Zufalligen Folgen," *Bull. Acad. Sci. U.S.S.R., Ser. Math.*, Vol. 5, pp. 3–14, 1941.
13. S. Kay, "Autoregressive Spectral Analysis of Narrowband Processes in White Noise with Application to Sonar Signals," Ph.D. Thesis, Georgia Institute of Technology, 1980.
14. S. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
15. M. Pagano, "Estimation of Models of Autoregressive Signal Plus White Noise," *Ann. Statistics*, Vol. 2, pp. 99–108, 1974.
16. S. Kay, "More Accurate Autoregressive Parameter and Spectral Estimates for Short Data Records," *Rec. 1st IEEE Workshop on Spectral Analysis*, Hamilton, Ont., Aug. 1981.
17. A. H. Nuttall, "Spectral Analysis of a Univariate Process with Bad Data Points, via Maximum Entropy and Linear Predictive Techniques," Naval Underwater Systems Center, Tech. Rep. 5303, New London, CT, March 26, 1976.
18. T. J. Ulrych and R. W. Clayton, "Time Series Modelling and Maximum Entropy," *Phys. Earth and Planetary Interiors*, Vol. 12, pp. 188–200, Aug. 1976.
19. S. L. Marple, Jr., "A New Autoregressive Spectrum Analysis Algorithm," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, pp. 441–454, Aug. 1980.
20 J. Durbin, "Efficient Estimation of Parameters in Moving-Average Models," *Biometrika*, Vol. 46, pp. 306–316, 1959.
21. H. Akaike, "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, Vol. 60, pp. 255–265, 1973.
22. S. Kay, "Noise Compensation for Autoregressive Spectral Estimates," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, pp. 292–303, June 1980.
23. H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automatic Control*, Vol. AC-19, pp. 716–723, Dec. 1974.

APPENDIX 2.1
DERIVATION OF PERIODOGRAM STATISTICS

The statistics of the periodogram are derived in this appendix for the case of zero-mean white Gaussian noise. It is assumed that the periodogram is evaluated at $\omega_k = k2\pi/N$ for $k = 0, 1, \ldots, N/2$, where N is even. Let

$$A(\omega_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \cos \omega_k n$$
$$B(\omega_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \sin \omega_k n \tag{2.1.1}$$

Then

$$\hat{P}_{\mathrm{PER}}(\omega_k) = A^2(\omega_k) + B^2(\omega_k) \tag{2.1.2}$$

Note that $B(\omega_0)$ and $B(\omega_{N/2})$ are identically zero. Since $x(n)$ is a Gaussian random process, $A(\omega_k)$ and $B(\omega_k)$ are jointly Gaussian. Furthermore, they are uncorrelated and hence independent, as we will now show.

$$E[A(\omega_k)] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} E[x(n)] \cos \omega_k n = 0$$

$$E[B(\omega_k)] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} E[x(n)] \sin \omega_k n = 0$$

$$\begin{aligned}
E[A(\omega_k)B(\omega_k)] &= \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} E[x(n)x(m)] \cos \omega_k n \sin \omega_k m \\
&= \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \sigma_x^2 \delta(n-m) \cos \omega_k n \sin \omega_k m \\
&= \frac{\sigma_x^2}{N} \sum_{n=0}^{N-1} \sin \omega_k n \cos \omega_k n \\
&= \frac{\sigma_x^2}{2N} \sum_{n=0}^{N-1} \sin 2\omega_k n \\
&= \frac{\sigma_x^2}{2N} \mathcal{I}m\left\{ \sum_{n=0}^{N-1} \exp(j2\omega_k n) \right\}
\end{aligned}$$

But

$$\sum_{n=0}^{N-1} \exp(j2\omega_k n) = \frac{\sin N\omega_k}{\sin \omega_k} \exp[j(N-1)\omega_k]$$

Hence,

$$\begin{aligned} E[A(\omega_k)B(\omega_k)] &= \frac{\sigma_x^2}{2N}\sin(N-1)\omega_k\frac{\sin N\omega_k}{\sin\omega_k} \\ &= \frac{\sigma_x^2}{2N}\sin[(N-1)2\pi k/N]\frac{\sin 2\pi k}{\sin 2\pi k/N} \\ &= 0 \qquad \text{for } k = 1, \ldots, N/2-1 \end{aligned}$$

which proves that $A(\omega_k)$ and $B(\omega_k)$ are independent. Next we compute the variances of $A(\omega_k)$ and $B(\omega_k)$.

$$\begin{aligned} \mathrm{VAR}[A(\omega_k)] &= E[A^2(\omega_k)] \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} E[x(n)x(m)]\cos\omega_k n\cos\omega_k m \\ &= \frac{\sigma_x^2}{N}\sum_{n=0}^{N-1}\cos^2\omega_k n \\ &= \begin{cases} \dfrac{\sigma_x^2}{2} & \text{for } k = 1, 2, \ldots, N/2-1 \\ \sigma_x^2 & \text{for } k = 0, N/2 \end{cases} \end{aligned}$$

since $\Sigma_{n=0}^{N-1}\cos^2\omega_k n = N/2$ for $k = 1, 2, \ldots, N/2 - 1$. Similarly,

$$\mathrm{VAR}[B(\omega_k)] = \sigma_x^2/2, \qquad k = 1, 2, \ldots, N/2-1$$

In summary,

$$\begin{aligned} A(\omega_k) &\sim N(0, \sigma_x^2/2), & k &= 1, 2, \ldots, N/2-1 \\ &\sim N(0, \sigma_x^2), & k &= 0, N/2 \\ B(\omega_k) &\sim N(0, \sigma_x^2/2), & k &= 1, 2, \ldots, N/2-1 \\ B(\omega_0) &= B(\omega_{N/2}) = 0 \end{aligned} \tag{2.1.3}$$

and $A(\omega_k)$ and $B(\omega_k)$ are independent. Hence,

$$\begin{aligned} \frac{\hat{P}_{\mathrm{PER}}(\omega_k)}{\sigma_x^2/2} &= \left[\frac{A(\omega_k)}{\sqrt{\sigma_x^2/2}}\right]^2 + \left[\frac{B(\omega_k)}{\sqrt{\sigma_x^2/2}}\right]^2, \qquad k = 1, 2, \ldots, N/2-1 \\ \frac{\hat{P}_{\mathrm{PER}}(\omega_k)}{\sigma_x^2} &= \left[\frac{A(\omega_k)}{\sigma_x}\right]^2, \qquad k = 0, N/2 \end{aligned} \tag{2.1.4}$$

Since each random variable in brackets is $N(0, 1)$, it follows from Section 2.3.3 that

$$\begin{aligned} \frac{2\hat{P}_{\mathrm{PER}}(\omega_k)}{\sigma_x^2} &\sim \chi_2^2, \qquad k = 1, 2, \ldots, N/2-1 \\ \frac{\hat{P}_{\mathrm{PER}}(\omega_k)}{\sigma_x^2} &\sim \chi_1^2, \qquad k = 0, N/2 \end{aligned}$$

Also, $P_x(\omega) = \sigma_x^2$, so

$$\frac{2\hat{P}_{\mathrm{PER}}(\omega_k)}{P_x(\omega)} \sim \chi_2^2, \qquad k = 1, 2, \ldots, N/2 - 1$$

$$\frac{\hat{P}_{\mathrm{PER}}(\omega_k)}{P_x(\omega)} \sim \chi_1^2 \qquad k = 0, N/2$$

ACKNOWLEDGMENTS

The author wishes to thank R. Gross, P. Lin, J. Mulhearn, and D. Sengupta of the University of Rhode Island for providing many of the computer-generated figures used throughout this chapter.