Elec4621:
Advanced Digital Signal Processing
Chapter 6: Random Processes and Power Spectrum Estimation

Dr. D. S. Taubman

April 8, 2011

1 Random Variables and Vectors

We will consistently use upper case (as in X) to refer to random variable (RV), and lower case (as in x) to refer to a realization (i.e., the outcome, or actual value) of the random variable. You may think about X as a place-holder for the unknown outcome, x, which carries around our prior expectations about the likelihood of each possible outcome. This likelihood information is captured by the Probability Density Function (PDF) of X, which is written $f_X(x)$.

Definition 1 *The PDF, $f_X(x)$, of a RV, X, is defined by*

$$f_X(x) = \lim_{\|\delta(x)\| \to 0} \frac{P\left(X \in \delta\left(x\right)\right)}{\|\delta\left(x\right)\|}$$

where $\delta(x)$ is a small interval which contains x, and $\|\delta(x)\|$ is its length.

It is easy to see that

$$\int f_X\left(x\right) dx = 1$$

Definition 2 *The mean (or expectation) of X, μ_X is defined by*

$$\mu_X = E[X] = \int x f_X(x) dx$$

Definition 3 *The variance of X is defined by*

$$\sigma_X^2 = var(X) = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2 = \left(\int x^2 f_X(x) dx\right) - \mu_X^2$$

More generally, if $g()$ is any function which maps x to y according to

$$y = g(x)$$

we may view the y values as outcomes of another random variable, Y. We write $Y = g(X)$ and we can easily show that the expectation of Y satisfies

$$\mu_Y = E[Y] = \int y f_Y(y)\, dy = \int g(x)\, f_X(x)\, dx$$

Thus, the variance of X is nothing other than the mean of another random variable, Y, whose outcome is $y = (x - \mu_X)^2$, whenever x is the outcome of X.

It follows that expectation is a linear operator, with

$$E[\alpha X] = \alpha E[X], \quad \alpha \text{ a deterministic constant}$$

and

$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(Y)]$$

More generally, given any two random variables, X and Y,

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$$

1.1 Random Vectors

It is convenient to introduce the notion of a random vector, $\mathbf{X}$. An m-dimensional random vector is nothing other than a collection of m random variables, X_1 through X_m, which we may write as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}$$

We write $f_{\mathbf{X}}(\mathbf{x})$ for the PDF of $\mathbf{X}$, which is the joint probability density function of the m random variables. We sometimes write this as $f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m)$ to emphasize the individual random variables whose joint probability distribution is being considered. The joint PDF is defined in the natural way as

$$f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) \triangleq \lim_{\|\delta(x_1)\|,\ldots,\|\delta(x_m)\| \to 0} \frac{P(X_1 \in \delta x_1 \text{ and } \ldots \text{ and } X_m \in \delta(x_m))}{\|\delta(x_1)\| \cdot \ldots \cdot \|\delta(x_m)\|}$$

We refer to the individual distributions of the random variables, $f_{X_k}(x_k)$, as their "marginal" distributions or marginal PDF's. The k^{th} marginal distribution is obtained by integrating the joint PDF over all variables except x_k.

It is easy to show that the mean of $\mathbf{X}$ is just the vector formed from the means of the individual random variables. That is,

$$\mu_{\mathbf{X}} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_m] \end{pmatrix} = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \vdots \\ \mu_{X_m} \end{pmatrix}$$

1.2 Independence and Correlation

Definition 4 *Two random variables, X and Y, are said to be statistically independent if their distribution is separable, i.e.*

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

This definition agrees with our intuition concerning independent events. Specifically, if two events, X and Y, have independent outcomes and we want to know the probability that $X \in \delta(x)$ and $Y \in \delta(y)$ simultaneously, we expect that this should be the product of the probability that $X \in \delta(x)$ and the probability that $Y \in \delta(y)$. That is, we expect independent events to satisfy

$$\begin{aligned} f_{X,Y}(x,y) &\triangleq \lim_{\|\delta(x)\|,\|\delta(y)\| \to 0} \frac{P(X \in \delta(x) \quad \text{and} \quad Y \in \delta(y))}{\|\delta(x)\| \cdot \|\delta(y)\|} \\ &= \lim_{\|\delta(x)\|,\|\delta(y)\| \to 0} \frac{P(X \in \delta(x)) \cdot P(Y \in \delta(y))}{\|\delta(x)\| \cdot \|\delta(y)\|} \\ &= f_X(x) \cdot f_Y(y) \end{aligned}$$

The notion of independence is perhaps even more intuite if we consider conditional probability.

Definition 5 *We write $f_{X|Y}(x|y)$ for the conditional probability distribution of X given that $Y = y$, where $f_{X|Y}(x|y)$ is defined by*

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,y)dx}$$

It is easy to see that $f_{X|Y}(x|y) = f_X(x)$ if and only if X and Y are statistically independent. That is, when X and Y are independent, the outcome of Y is irrelevant to the distribution of X.

Definition 6 *The correlation of two random variables, X and Y, is defined by*

$$E[XY] = \int\int xy f_{X,Y}(x,y)dxdy$$

where $f_{X,Y}(\cdot,\cdot)$ is the joint distribution (PDF) of X and Y.

Definition 7 *The covariance of X and Y is defined by*

$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

Two random variables whose covariance is 0 *are said to be "uncorrelated."*

Definition 8 *The correlation matrix, $R_{\mathbf{X}}$, of a random vector, $\mathbf{X}$, is defined by*

$$R_{\mathbf{X}} = E\left[\mathbf{X}\cdot\mathbf{X}^t\right] = \begin{pmatrix} E[X_1X_1] & E[X_1X_2] & \cdots & E[X_1X_m] \\ E[X_2X_1] & E[X_2X_2] & \cdots & E[X_2X_m] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_mX_1] & E[X_mX_2] & \cdots & E[X_mX_m] \end{pmatrix}$$

Definition 9 *The covariance matrix, $C_{\mathbf{X}}$, of a random vector, $\mathbf{X}$, is defined by*

$$C_{\mathbf{X}} = E\left[(\mathbf{X}-\mu_{\mathbf{X}})\cdot(\mathbf{X}-\mu_{\mathbf{X}})^t\right] = R_{\mathbf{X}} - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^t$$

The entry at the i'th row and j'th column of this matrix is $cov(X_iY_j) = E[X_iX_j] - \mu_{X_i}\mu_{X_j}$.

Evidently, the correlation and covariance matrices are both symmetric:

$$R_{\mathbf{X}} = R_{\mathbf{X}}^t, \quad C_{\mathbf{X}} = C_{\mathbf{X}}^t$$

If X and Y are statistically independent, then they must be uncorrelated. To see this, observe that

$$\begin{aligned} E[XY] &= \int\int xyf_{X,Y}(x,y)dxdy \\ &= \int xf_X(x)dx \cdot \int yf_Y(y)dy \\ &= \mu_X\mu_Y \end{aligned}$$

More generally, if the m random variables, X_1 through X_m are mutually independent, then they must be mutually uncorrelated, meaning that $C_{\mathbf{X}}$ is a diagonal matrix,

$$C_{\mathbf{X}} = \begin{pmatrix} \mathrm{cov}(X_1,X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_m) \\ \mathrm{cov}(X_2,X_1) & \mathrm{cov}(X_2,X_2) & \cdots & \mathrm{cov}(X_2,X_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_m,X_1) & \mathrm{cov}(X_m,X_2) & \cdots & \mathrm{cov}(X_m,X_m) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m^2 \end{pmatrix}$$

Unfortunately, the converse is not generally true. That is, uncorrelated random variables are not generally statistically independent, despite the fact that we often use the word "uncorrelated" informally to suggest statistical independence. One important exception occurs when $\mathbf{X}$ has a Gaussian distribution. In this case, mutually uncorrelated random variables are always independent, and vice-versa.

Definition 10 *We say that* $\mathbf{X}$ *has a Gaussian distribution, or equivalently that* X_1 *through* X_m *are mutually Gaussian, if the joint PDF,* $f_{\mathbf{X}}(\mathbf{x})$ *has the form:*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m \|\mathbf{C}_{\mathbf{X}}\|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{X}})^t \mathbf{C}_{\mathbf{X}}^{-1}(\mathbf{x}-\mu_{\mathbf{X}})}$$

Here, $\|\mathbf{C}\|$ *identifies the determinant of the matrix,* $\mathbf{C}$*. It can be shown that the determinant of a covariance matrix must be non-negative.*

The Gaussian distribution evidently depends only upon means and covariances, i.e. the expectations of polynomials of degree 1 or 2 in the m random variables in $\mathbf{X}$. We call these the *first and second order statistics*. The distribution of a Gaussian random vector is determined entirely by its first and second order statistics. This is one of the reasons why so much of statistical signal processing is restricted to consideration of the first and second order statistics.

If $\mathbf{X}$ is Gaussian and its random variables are all mutually uncorrelated, $C_{\mathbf{X}}$ is diagonal and the Gaussian PDF expression separates into

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_{X_i}^2}} e^{-\frac{1}{2}\frac{x_i^2}{\sigma_{X_i}^2}}$$

Gaussian distributions are the only distributions for which statistical independence is equivalent to zero covariance.

Questions:

- Under what conditions are uncorrelated random variables statistically independent?
- Can you construct two non-trivial (i.e. not equal to a constant with probability 1) random variables, X and Y, whose covariance is zero, but which are not statistically independent? (hint: you can make the variables deterministically related to one another so that they are definitely not independent, e.g. $Y = f(X)$ for some function, $f(\cdot)$).

2 Random Processes

Random processes are a generalization of random vectors, to infinitely long sequences of random variables. The time sequence, $x[n]$, is understood as a realiztion of some underlying random process, $X[n]$.

Example 1 *Consider the addition of a deterministic signal,* $x_0[n]$*, to a random noise process,* $\nu[n]$*, yielding*

$$x[n] = x_0[n] + \nu[n]$$

Since the noise is random, so is $x[n]$*. The corresponding random processes satisfy*

$$X[n] = x_0[n] + V[n]$$

Note that any random component, such as noise, in the time sequence renders it a random process. Consequently, almost all signals which we deal with in signal processing are realizations of random processes.

In general, we cannot gain access to the statistics of a random process by monitoring it over time, because the entire time signal itelf is only a single realization of the process. We will need to restrict the random process to a narrower class of "ergodic" random processes for this to be possible.

Definition 11 *We say that a random process is stationary if the statistics are invariant to shifts. Formally, let $\mathcal{I} = \{i_1, i_2, \ldots, i_k\}$ be any finite set of k time indices and let $\mathbf{X}_\mathcal{I}$ be the k-dimensional vector formed from the random variables, $X[i_1], X_[i_2], \ldots, X_[i_k]$. We say that $X[n]$ is stationary if for all finite index sets, $\mathcal{I}$, and all delays, m, the joint distribution of $\mathbf{X}_\mathcal{I}$ is identical to the joint distribution of the random vector, $\mathbf{X}_{\mathcal{I}+m}$, where $\mathcal{I} + m = \{i_1 + m, \ldots, i_k + m\}$.*

It is obvious that a stationary random process has a constant mean. That is,

$$\mu_{X[n]} = E[X[n]] = \int x f_{X[n]}(x) dx = \int x f_{X[0]}(x) dx = \mu_{X[0]}$$

Definition 12 *We define the autocorrelation sequence for $X[n]$ by*

$$R_{XX}[n, n-m] = E[X[n] \cdot X[n-m]]$$

If $X[n]$ is stationary, we may simplify this to

$$R_{XX}[n, n-m] = E[X[0]X[-m]] = R_{XX}[m]$$

since that the autocorrelation sequence depends only upon the lag, m.

For stationary sequences, we may define the autocovariance sequence, $C_{XX}[m] = \mathrm{cov}(X[n], X[n-m])$, as a function of only the lag, m.

In general, stationarity means a great deal more than having mean and autocorrelation (and autocovariance) properties which are independent of time. Since this is often all we need, we define Wide-Sense Stationarity (WSS) as the property of having stationary mean and autocorrelation functions. A stationary process is certainly WSS, but a WSS process may not be truly stationary in the sense that all of its statistical properties might not be invariant to time shifts. Nevertheless, at least the mean and autocorrelation functions of a WSS process are invariant to time shifts. Wide sense stationarity is often all we need, since we largely restrict our attention to second order statistics. Note that Gaussian processes are described entirely by their second order statistics and so for Gaussian processes wide sense stationarity is equivalent to stationarity. A WSS Gaussian process is described completely by its mean, μ_X, and autocorrelation, $R_{XX}[m]$.

2.1 Time Averages

In this section, we restrict our attention to WSS random processes. Recall that the mean is defined by $\mu_{X[n]} = E[X[n]]$, which for WSS processes is independent of the index, n. We can also define the time average of a specific realization,

$$m_x = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n]$$

The random process is said to be *mean-ergodic* if, for realizations, $x[n]$, we find that the time average, m_x, is equal the *ensemble* (or statistical) average, μ_X, with probability 1. That is, regardless of which realization we look at, with probability 1, we will find that the time average turns out to be identical to μ_X.

It is clear that $E[m_X] = \mu_X$; that is, the ensemble average of the time averages must equal the ensemble average. This is because

$$\begin{aligned} E[m_X] &= E\left[\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} X[n]\right] \\ &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} E[X[n]] \\ &= \mu_X \end{aligned}$$

However, it is not clear that the variance of the time averages, m_X, should necessarily be zero so that only one value has any likelihood of occurring. In fact, it is possible to construct examples of WSS and even stationary random processes for which the time average has a non-zero variance.

Definition 13 *In the specific case where the time average always comes out to the ensemble average, with probability* 1*, we say that the random process is "mean ergodic".*

We can define a time-autocorrelation for any specific realization, $x[n]$, of the random process, $X[n]$, according to

$$r_{xx}[m] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n]x[n-m]$$

Again, in the case of a *correlation ergodic* random process, this time average always yields the same autocorrelation sequence, $r_x[m]$, for all realizations of the underlying random process, with probability 1 and so we can equate $r_x[m] = R_X[m]$. Note carefully that $r_x[m]$ is calculated from a single realization, $x[n]$, while $R_X[m]$ is a statistical property of the underlying random process.

We generally restrict our attention to suitably ergodic random processes, because this allows us to access the statistical properties of the random process by taking time averages. In a little while we shall investigate specific methods for performing the relevant time averages.

2.2 Effect of LTI Systems on Random Processes

In this section, we restrict our attention to WSS ergodic random processes.

Definition 14 *We may define the PDS (Power Density Spectrum) for the random process, $X[n]$, as the DTFT of its autocorrelation sequence, $R_{XX}[m]$, i.e.*

$$S_{XX}(\omega) = \hat{R}_{XX}(\omega) = \sum_m R_{XX}[m] e^{-j\omega m}$$

We shall see shortly why this is called the Power Density Spectrum.

Note that $S_{XX}(\omega)$ is entirely real-valued, since the autocorrelation sequence is symmetric, i.e., $R_{XX}[m] = R_{XX}[-m]$.

Definition 15 *We define the cross-correlation between two WSS random processes, $X[n]$ and $Y[n]$, by*

$$R_{XY}[m] = E[X[n], Y[n-m]]$$

The WSS property ensures that this definition is independent of n, which is merely a place-holder to remind us of the interpretation.

When the process is appropriately ergodic, the cross-correlation is equal (with probability 1) to the time average,

$$r_{xy}[m] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n]y[n-m]$$

for all realizations, $x[n]$ and $y[n]$, of $X[n]$ and $Y[n]$. Note that $R_{XY}[m] = R_{YX}[-m]$.

Definition 16 *We write $S_{XY}(\omega) = \hat{R}_{XY}(\omega)$ for the Fourier transform (DTFT) of $R_{XY}[m]$. We might call this the* cross-power spectral density*.*

Now suppose that $Y[n]$ is the output of an LTI system with impulse response, $h[n]$, which is excited by the random process, $X[n]$. Then, the following relationships can be shown to hold:

- The cross-correlation sequence, $R_{YX}[m]$, is given by the convolution of the autocorrelation sequence, $R_{XX}[m]$, with the impulse response, $h[m]$, i.e.

  $$R_{YX}[m] = \sum_k h[k] R_{XX}[m-k]$$

  In the Fourier domain this becomes

  $$S_{YX}(\omega) = S_{XX}(\omega)\hat{h}(\omega)$$

  where x^* denotes the complex conjugate of x.

- Since $R_{XY}[m] = R_{YX}[-m]$, we also have

$$\begin{aligned} R_{XY}[m] &= \sum_k h[-k] R_{XX}[m-k] \\ S_{XY}(\omega) &= S_{XX}(\omega)\hat{h}^*(\omega) \end{aligned}$$

- The autocorrelation of the output random process, $R_{YY}[m]$, is obtained by convolving $R_{XX}[m]$ by the sequence,

$$h_2[m] = \left(h \star \tilde{h}\right)[m] = \sum_k h[k]\, h[k-m]$$

In the Fourier (DTFT) domain, this is easily expressed as

$$\begin{aligned} S_{YY}(\omega) &= S_{XX}(\omega)\, \hat{h}(\omega)\, \hat{h}^*(\omega) \\ &= S_{XX}(\omega) \cdot \left|\hat{h}(\omega)\right|^2 \end{aligned}$$

The above relationships suggest a practical method for measuring the impulse response, $h[n]$, of a system. Suppose we excite the system by white noise, i.e. a random process for which $R_{XX}[m] = \delta[m]$ and hence $S_{XX}(\omega) = 1$. Suppose then that we use time averages to estimate the cross-correlation between the input and output sequences, $R_{YX}[m]$. We find that

$$S_{YX}(\omega) = S_{XX}(\omega)\hat{h}(\omega) = \hat{h}(\omega)$$

and so the cross-correlation sequence is the system impulse response which we seek.

2.3 Interpretation of Power Density Spectrum

The above relationships also suggest a good reason for using the term "Power Density Spectrum" to refer to the Fourier transform of the autocorrelation sequence. To see this, observe firstly that

$$R_{XX}[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{XX}(\omega) e^{j\omega m} d\omega$$

which is just the inverse DTFT formula. Evaluating this at $m = 0$, we find that

$$E[X[n]^2] = R_{XX}[0] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{XX}(\omega) d\omega$$

Thus, integrating $S_{XX}(\omega)$ over frequency, gives us the "power" of the random process, $X[n]$.

Now suppose we design a band-pass filter which suppresses all but a narrow range of frequencies in the neighbourhood of ω_0, i.e.

$$\hat{h}(\omega) = \begin{cases} 1 & \text{if } |\omega - \omega_0| < \delta, \\ 0 & \text{otherwise.} \end{cases}$$

and we measure the power of the output random process from this filter. Then we will obtain

$$\begin{aligned} E[Y[n]^2] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{YY}(\omega)d\omega \\ &= \frac{1}{2\pi}\int_{-\omega_0-\delta}^{\omega_0+\delta} S_{XX}(\omega)d\omega \end{aligned}$$

So we see that the power around frequency, ω_0, is directly related to the value of the PDS, $S_{XX}(\omega)$, in the neighbourhood of ω_0. This justifies our terminology.

3 Estimating Autocorrelation and Power Spectra

We have seen that for a WSS ergodic random process, $X[n]$, the time average, $r_{xx}[m]$ is equal to the ensemble average, $R_{XX}[m]$, for all realizations, $x[n]$. This suggests that we can measure $R_{XX}[m]$ by taking time averages, which is true, except that in practice we do not have an infinite length sequence over which to take this average. In this section, we consider various techniques for estimating $R_{XX}[m]$ from a finite length sequence, $x[n]$.

3.1 Biased and Unbiased Estimates of $R_{XX}[m]$

Given a finite number of time samples, $x[n]$, in the range $0 \le n < N$, perhaps the most natural estimate for $R_{XX}[m]$ is the finite time average

$$r_{xx}[m] = \frac{1}{N-|m|}\sum_{n=\max\{0,m\}}^{N-1-m+\max\{0,m\}} x[n]\,x[n-m], \quad |m| < N \qquad (1)$$

The summation limits here are chosen to ensure that both $x[n]$ and $x[n-m]$ correspond to available samples, $x[n]$, in the range $0 \le n < N$. Evidently, there are $N-|m|$ such samples, which is why we have divided by $N-|m|$. For lag values, $|m| \ge N$ we have no information whatsoever on which to base an estimate of $r_{XX}[m]$, so we use the value 0.

The estimate in equation (1) is said to be "unbiased", because the statistical average of $r_{xx}[m]$, taken over all length N outcomes, $x[n]$, of the underlying random process, $X[n]$, is identical to the actual autocorrelation, $R_{XX}[m]$. To

see this, observe that

$$\begin{aligned} E\left[r_{XX}\left[m\right]\right] &= E\left[\frac{1}{N-|m|}\sum_{n=\max\{0,m\}}^{N-1-m+\max\{0,m\}} x\left[n\right]x\left[n-m\right]\right] \\ &= \frac{1}{N-|m|}\sum_{n=\max\{0,m\}}^{N-1-m+\max\{0,m\}} E\left[X\left[n\right]X\left[n-m\right]\right] \\ &= \frac{1}{N-|m|}\sum_{n=\max\{0,m\}}^{N-1-m+\max\{0,m\}} R_{XX}\left[m\right] \\ &= R_{XX}\left[m\right] \end{aligned}$$

One obvious problem with the estimate in equation (1), however, is that the estimates cannot be expected to be very reliable at lags, m, approaching $\pm N$. In the extreme case, where $|m| = N-1$, the "average" involves only one time sample: $r_{xx}\left[N-1\right] = r_{xx}\left[1-N\right] = x\left[0\right]x\left[N-1\right]$. Not surprisingly, then, the variance of the estimate, $r_{XX}\left[m\right]$, taken over all realizations of the random process, $X\left[n\right]$, tends to increase markedly as $|m|$ approaches N. As we shall see, this property can have devastating consequences on the usefulness of the Power Density estimate obtained by taking the DTFT of the estimated autocorrelation sequence.

In view of the above problem, a more conservative estimate might be formed from

$$\begin{aligned} r_{xx}^{\text{biased}}\left[m\right] &= \frac{1}{N}\sum_{n=\max\{0,m\}}^{N-1-m+\max\{0,m\}} x\left[n\right]x\left[n-m\right], \quad |m| < N \qquad (2) \\ &= \frac{N-|m|}{N} r_{xx}^{\text{unbiased}}\left[m\right] \end{aligned}$$

Obviously this is a biased estimate of the true autocorrelation sequence, since

$$\begin{aligned} E\left[r_{xx}^{\text{biased}}\left[m\right]\right] &= \frac{N-|m|}{N} E\left[r_{xx}^{\text{unbiased}}\left[m\right]\right], \quad |m| < N \\ &= \frac{N-|m|}{N} R_{XX}\left[m\right], \quad |m| < N \\ &= w_B\left[m\right] \cdot R_{XX}\left[m\right] \end{aligned}$$

where $w_B\left[m\right]$ is the Bartlett window, obtained by sampling the function,

$$w_B\left(t\right) = \begin{cases} 1 - \left|\frac{t}{\tau}\right| & |t| < \tau \\ 0 & |t| \geq \tau \end{cases}$$

with $\tau = N$. Bias in the mean value of the autocorrelation estimate is the price paid for a reduction in the variance of the estimate. Obviously, the variance in $r_{xx}^{\text{biased}}\left[m\right]$ is reduced by $\left(w_B\left[m\right]\right)^2$, with respect to the variance in the unbiased estimate.

3.2 The Periodogram

Since the mean value of the biased autocorrelation estimate, $r_{xx}^{\text{biased}}[m]$, is a windowed version of the true autocorrelation, $R_{XX}[m]$, we should not be surprised to find that the biased power density spectrum (PDS) obtained by taking its Fourier transform, has a mean value equal to $S_{XX}(\omega)$ convolved by $\hat{w}_B(\omega)$. To see that this is actually the case, let

$$S_{xx}^{\text{biased}}(\omega) = \sum_{m=1-N}^{N-1} r_{xx}^{\text{biased}}[m]\, e^{-j\omega m}$$

and observe that

$$\begin{aligned} E\left[S_{XX}^{\text{biased}}(\omega)\right] &= \sum_m E\left[r_{XX}^{\text{biased}}[m]\right] e^{-j\omega m} \\ &= \sum_m w_B[m]\, R_{XX}[m]\, e^{-j\omega m} \\ &= \sum_k \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{w}_B(\theta)\, S_{XX}(\omega + 2\pi k - \theta)\, d\theta \\ &= \sum_k \left(\hat{w}_B \star S_{XX}\right)(\omega + 2\pi k) \end{aligned}$$

The summation over k in the above equation represents aliasing contributions, which must be taken into account since $S_{XX}(\omega)$ convolved with $\hat{w}_B(\omega)$ is not bandlimited. We have seen all of this before in connection with the window method for FIR filter design. In most cases, $S_{XX}(\omega)$ decays rapidly with ω so that $\hat{w}_B(\omega) \star S_{XX}(\omega)$ is approximately bandlimited and we can ignore the aliasing contributions.

The biased power spectrum, $S_{XX}^{\text{biased}}(\omega)$, may be understood in a very different way as follows. Define $y[n]$ to be the infinite length sequence whose samples agree with $x[n]$ for $0 \le n < M$ and are 0 everywhere else. That is,

$$y[n] = x[n]\, w_0[n]$$

where $w_0[n]$ is the all-or-nothing window,

$$w_0[n] = \begin{cases} 1 & 0 \le n < m \\ 0 & \text{otherwise} \end{cases}$$

The biased autocorrelation estimate may be expressed more simply as the autocorrelation of y. Specifically,

$$\begin{aligned} r_{xx}^{\text{biased}}[m] &= \frac{1}{N}\sum_{n=-\infty}^{\infty} y[n]\, y[n-m] \\ &= \frac{1}{N}\sum_{n=-\infty}^{\infty} y[n]\, \tilde{y}[m-n] \end{aligned}$$

where $\tilde{y}[m] = y[-m]$ is the mirror image of $y[m]$. Evidently, then, $r_{xx}^{\text{biased}}[m]$ is obtained by convolving $y[m]$ with its mirror image. This means that

$$\begin{aligned} S_{xx}^{\text{biased}}(\omega) &= \frac{1}{N}\hat{y}(\omega)\hat{y}^*(\omega) \\ &= \frac{1}{N}|\hat{y}(\omega)|^2 \end{aligned}$$

We conclude that $S_{xx}^{\text{biased}}(\omega)$ is the normalized, squared magnitude of the Fourier transform of the finite support sequence, $y[n]$. In our study of the DFT, we found that the N-point DFT of such a sequence is equivalent to its DFTF and that

$$Y[k] = \hat{y}(\omega)|_{\omega=\frac{2\pi}{N}k \bmod 2\pi}$$

With some abuse of notation, write $X_N[k]$ for the N point DFT of $x[n]$, i.e., the DFT obtained by taking the first N samples from the infinite sequence, $x[n]$. Then $X_N[k] = Y[k]$ and

$$\frac{1}{N}|X[k]|^2 = S_{xx}^{\text{biased}}(\omega)\big|_{\omega=\frac{2\pi}{N}k \bmod 2\pi}$$

This suggests a simple method for estimating the power spectrum of the random process, $X[n]$, based on samples $x[n]$, $0 \le n < N$, from one realization of the random process. The method involves taking the DFT of the N samples which are available, squaring the DFT magnitudes, and dividing by N. Obviously, a computationally efficient implementation may be obtained by employing the FFT algorithm. The result is known as the **"periodogram"**.

3.3 Problems with the Periodogram

We have already seen that the periodogram is a biased estimate of the true power spectrum, where the nature of the bias is that $S_{XX}(\omega)$ has been convolved by the Bartlett window' Fourier transform, $\hat{w}_B(\omega)$. This convolution limits the inherent resolution of the estimated power spectrum, in the frequency domain. Another way to see that there is limited spectral resolution is to observe that all of the information in $S_{xx}^{\text{biased}}(\omega)$ is captured by its samples, at the locations $\omega = \frac{2\pi}{N}k$ – these are the normalized squared magnitudes of the N-point DFT.

Unfortunately, the periodogram has a much more serious problem. Even though the biased autocorrelation estimate behind this spectral estimator has lower variance than an unbiased autocorrelation estimate, the variance of $S_{xx}^{\text{biased}}(\omega)$ is still very large. In fact, the variance of $S_{xx}^{\text{biased}}(\omega)$ does not decrease, as the number of available samples, N, increases!! For Gaussian random processes, it can be shown that (see §12.1.2 in Proakis and Manolakis)

$$\begin{aligned} \text{var}\left[S_{XX}^{\text{biased}}(\omega)\right] &= S_{XX}^2(\omega)\left[1 + \left(\frac{\sin\omega N}{N\sin\omega}\right)^2\right] \\ &\longrightarrow_{N\to\infty} S_{XX}^2(\omega) \end{aligned}$$

Thus, the root-mean-square deviation in the power spectral estimate is as large as the estimate itself!

A reasonable conclusion from the above is that taking the DFT of a finite collection of samples does not produce a reliable estimate of the underlying power spectrum at all. Moreover, the reliability is not improved by increasing the number of samples (i.e., the number of points in the DFT). This is rather counter-intuitive, so it is easy to fall into the trap of using the DFT to estimate power spectra and assuming (wrongly) that having a long enough sequence will fix up any problems.

3.4 The Bartlett Method

The real source of the periodogram problem described above is that we are trying to use all of the available samples to get as much spectral resolution as possible. If the number of available samples, N, is doubled, the periodogram effectively applies a $2N$-point DFT, which samples the spectrum twice as finely, but each spectral sample is no more reliable than before. Another way to see this same thing is that as N doubles, we double the width of the Bartlett window, hence halving the width of $\hat{w}_B(\omega)$ which is "blurring" the spectral estimate.

In order to overcome the total lack of reliability of the periodogram's spectral estimates, we will have to sacrifice resolution in order to reduce the variance of the estimates. Many strategies have been proposed for doing this, but in this and the next section we outline two basic strategies on which the others build. By and large, it should be sufficient to have these two strategies in your arsenal.

The Bartlett method is perhaps the most obvious approach to trading spectral resolution for accuracy. Rather than devoting all N samples to improving the spectral resolution, we divide the N samples into disjoint segments of length L. Suppose there are P such segments, so that $N = LP$. Let $x_p[n]$ denote the sequence of L samples belonging to the p^{th} segment and let $S_p(\omega)$ denote its periodogram. The spectral estimate is then taken to be the average of all P periodograms. That is,

$$S_{xx}^{\text{bartlett}}(\omega) = \frac{1}{P}\sum_{p=0}^{P-1} S_p(\omega)$$

In the sampled frequency domain, this means that

$$\begin{aligned} S_{xx}^{\text{bartlett}}(\omega)\big|_{\omega=\frac{2\pi}{L}k \bmod 2\pi} &= \frac{1}{P}\sum_{p=0}^{P-1}\frac{1}{L}\left|X_k[p]\right|^2, \quad 0 \le k < L \\ &= \frac{1}{N}\sum_{p=0}^{P-1}\left|X_k[p]\right|^2, \quad 0 \le k < L \end{aligned}$$

Not surprisingly, the variance of the Bartlett spectral estimates is proportional to $\frac{1}{P} = \frac{L}{N}$. Thus, if we fix the value of L, using L-point DFT's, no matter how large the actual number of samples, N, then the variance of the spectral

estimates will obviously decrease in proportion to N, while the bias (the width of the Bartlett window) will be independent of N.

3.5 The Blackman-Tukey Method

While the Bartlett method averages periodograms from disjoint segments of the available data, the Blackman-Tukey method takes quite a different approach to improving the reliability of spectral estimates. In this case, the autocorrelation sequence is first estimated in the usual way as

$$r_{xx}^{\text{biased}}[m] = \frac{1}{N}\sum_{n} y[n]\, y[n-m]$$

where $y[n]$ is the windowed sequence,

$$y[n] = \begin{cases} x[n] & 0 \le n < N \\ 0 & \text{otherwise} \end{cases}$$

but the autocorrelation estimate is then windowed prior to taking its Fourier transform.

Let $w_{\text{BT}}[m]$ be the autocorrelation window. Then the spectral estimate is formed from

$$S_{xx}^{\text{BT}}(\omega) = \sum_{m=-L}^{L} \left(w_{\text{BT}}[m] \cdot r_{xx}^{\text{biased}}[m] \right) e^{-j\omega m}$$

where the L determines the support of the symmetric autocorrelation window, $w_{\text{BT}}[m]$, and is generally much smaller than N. Evidently, $S_{xx}^{\text{BT}}(\omega)$ is obtained by convolving (blurring) $S_{XX}^{\text{biased}}(\omega)$ with $\hat{w}_{\text{BT}}(\omega)$. Equivalently, the mean value of this spectral estimate, $E\left[S_{xx}^{\text{BT}}(\omega)\right]$, is obtained by convolving the true spectrum, $S_{XX}(\omega)$, by $\hat{w}_1(\omega)$ where

$$w_1[m] = \hat{w}_B[m] \cdot w_{\text{BT}}[m]$$

is the product of the Bartlett and autocorrelation window functions.

This blurring process reduces the spectral resolution, but also reduces the variance of the spectral estimate. When $\hat{w}_{\text{BT}}(\omega)$ decays rapidly with frequency and has a main lobe which is narrow compared with the spectrum, $S_{XX}(\omega)$, the spectral variance is given roughly by (see §12.2.3 in Proakis and Manolakis)

$$\text{var}\left[S_{XX}^{\text{BT}}(\omega)\right] = S_{XX}^2(\omega)\left[\frac{1}{N}\sum_{m=-L}^{L} w_{\text{BT}}^2[m]\right]$$

By fixing L, we ensure (as for the Bartlett method) that the variance decreases as $\frac{1}{N}$, while the spectral resolution (amount of blur) remains fixed (depends on L and the shape of the window).

The Bartlett and Blackman-Tukey methods have similar behaviour in regard to the way they trade spectral resolution for reliability of the estimates. As a rough rule of thumb, the Blackman-Tukey method yields estimates with roughly half the variance of those yielded by the Bartlett method, for the same number of samples and the same spectral resolution. However, the Bartlett method has obvious computational advantages, since it can be implemented using FFT's.