Biological Signal Analysis
Data acquisition and analysis for life scientists

Didier A Depireux

October 23, 2011

This accompanies the teaching of NACS/MPHY 615 Biological Signal Analysis (3). The course was meant to cover the origin and analysis of various biological signals, especially those arising from the nervous system. Emphasis is on the measurement and interpretation of these signals by techniques such as spectrum analysis, average evoked responses, single-unit histograms, and pattern recognition.

Contents

List of Figures

List of Tables

Part I

Data acquisition

1.BasicSignalProcessing.tex

Chapter 1

Basic Signal Processing

Some Properties of Biological Signals

1.1 Introduction: Signal, noise and data

One scientist's noise is often another scientist's signal.

Is all data generated by the brain signal?

Speaking in a somewhat general way, we say that all biological data can be considered to be signals. Obviously, however, some data are more signallike than others. The dividing line between data that can be profitably considered to be signallike and data that cannot depends upon both the origin of the data and how we propose to process it and analyze it conceptually. A discussion of the many facets of this idea in the light of modern computer data processing methods is one of the major purposes of this book. marking in this direction requires that we first establish some of the major concepts and properties of signals insofar as they relate to biological processes. The properties of these signals influence, guide, and sometimes determine the ways in which computer programs are developed to perform signal analysis.

Signal: A variation in the amplitude and polarity of an observed physical quantity produced by a process whose mechanisms we desire to understand by experimental investigation. The requirement that the variation be produced by a mechanism we are interested in is of basic importance and brings us to consider at once, noise, the inseparable companion of signal.

Noise: A variation in the size of an observed physical quantity we are investigating produced by a process or an aspect of a process that we have no present interest in.

Data: Some combination, often additive, of signal and noise. The additive

situations are easiest to deal with in terms of analysis and interpretation of results. In much of what follows we will assume it applies. In general, however, additivity should not be taken for granted.

The errant course of scientific progress is such that often what is considered to be a signal in one investigation turns out to be noise in another. Or more colloquially, one man's signal is another man's noise.

The variations in the size of a physical quantity are often time-dependent. When they are, the data is said to be a function of time and written $x(t)$. Temporal data variation is most convenient for us to consider and also most appropriate since a realtime computer generally accepts data in time sequential form. However, we may also profitably consider data which are functions of such variables as distances or angle, for it is usually a simple matter to convert them into functions of time by a signal transducer. As an example, a scanning densitometer converts the spatially varying density of a translucent object into a function of time as the densitometer is moved over the scanned object. An oscilloscope screen is an example of the process in reverse for there the time-varying data is converted into a function of distance along the horizontal axis of the oscilloscope screen. Hereafter, when we mention data signals and noise, we will consider them to be temporally varying.

We are interested in establishing the basic principles of a wide assortment of procedures by which we analyze the signal-like data of neurobiological investigations. Temporally generated signals and noises exhibit a wide variety of waveform features or parameters, and it is essential to classify them according to such features, for the validity of much of the subsequent data processing depends upon the presence or magnitude of these features. The following pages contain a discussion of some of the properties of signals to serve as the basis of understanding the signal analysis procedures and techniques to be described in later chapters.

1.2 Continuous Signals and Their Discrete Counterparts

Let us begin with data which consist only of signals. A signal is said to be continuous if it is defined at all instants of time during which it occurs. A continuous signal may, however, possess discontinuities or sudden changes in amplitude at certain instants of time. At these instants the slope of the signal is infinite. At other times the signal amplitude changes gradually so that by choosing an interval short enough, the corresponding change in amplitude can be made as small as we like. While continuous signals without discontinuities are the rule in such biological phenomena as the EEG, deliberately generated discontinuous signals may be generated by the instrumentation associated with neurobiological investigations.

As an example, the signal produced by a rat when it pushes a switch to obtain food is discontinuous. This type of signal is referred to as a step function. Illustrations of continuous and discontinuous signals are shown in Fig.1.1. It is also to be noted that whether continuous or not, the signals are always single valued: they have only one value at any particular instant in time. A particularly interesting and important discontinuous signal is the unit step signal of Fig.1.1(c).

$$u(t) = 0 \text{ when } t \leq t_d, \tag{1.1}$$

$$1 \text{ when } t > t_d. \tag{1.2}$$

t_d is the instant of discontinuity. The equation indicates that the signal jumps to 1 as soon as t becomes greater than t_d. The unit step is used, among other purposes, to describe a stimulus that has a sudden onset.

Besides speaking of a continuous signal, $x(t)$, we will also have occasion to speak of its time derivatives, the first derivative being written $dx(t)/dt$ or, alternatively, $x'(t)$. The first derivative is, of course, the time rate of change of the variable. When it is zero, the variable itself is at a local maximum or minimum value or, less frequently, at an inflection point. (The derivative of a constant signal is always zero.) This property is often used in determining when a spike-like waveform reaches a maximum or minimum. A peak detection device which essentially takes the time derivative of the waveform is commonly employed for this. When its output, the waveform time derivative, goes through zero in a negative direction, a positive maximum has occurred; when it goes through zero in a positive direction, a negative maximum has occurred. Figure XX1.2(a) illustrates the situation for the former case. The first derivative is also important in indicating when the signal is changing most rapidly because it has its greatest value at that time. A positive maximum in the first derivative indicates the time when the signal is increasing most rapidly; a negative maximum, when it is decreasing most rapidly. Just as a continuous signal may exhibit discontinuities, so may its derivatives. A discontinuity in the first derivative occurs when there is a cusp in the original signal. An example is the sawtooth signal of Fig. XX1.2 (b) . When it is at its maximum and minimum values, discontinuities occur in its first derivative, a square wave.

The derivative operation is not without practical difficulties since noise contributions tend to corrupt the derivative measurement. In computer analysis of data, the derivative operation is approximated by comparing successive sampled values of the signal with one another to see when maximum and minimum rates of change occur. Although this is an approximation, the results are often more than adequate. It is worth noting here that approximation is different from estimation, the latter being a statistical procedure whose meaning will be made clear in the subsequent

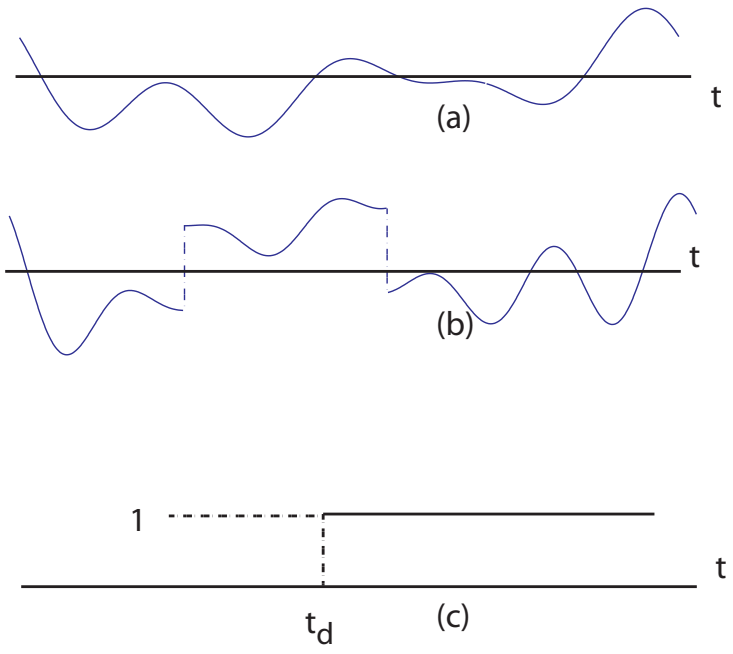

Figure 1.1: (a) A continuous signal; (b) a discontinuous signal; (c) the unit step $\mu(t)$, showing step onset at $t = t_d$

pages. In contrast to temporally continuous (T-continuous) signals are the temporally discrete (T-discrete) signals. These are signals which exist only at discrete instants in time. For our purposes the most important discrete signals are those which occur when a continuous signal has its amplitude measured or sampled at discrete instants of time that are usually equally spaced. A T-discrete signal is thus a sequence of measurements $x_1, x_2, \ldots, x_T$ lasting for the duration of the time the signal is observed. In digital data processing it is furthermore usually quantized in amplitude by an analog-to-digital converter. This gives it the property of being amplitude discrete (A-discrete). The result is a signal, T- and A-discrete, which provides the basic data thereafter for all subsequent computer analyses of the original signal.

Having introduced the continuous signal and its sampled T-discrete representation, it is useful to establish here a form of notation which permits us to distinguish between them with a minimum amount of confusion. We will use the symbol $^\circ$ to distinguish a sampled T-discrete signal from its continuous source signal. We will drop the $^\circ$ when no confusion seems possible. Similarly, we will use t to represent continuous time and $t^\circ\Delta$ to represent those instants that a signal is sampled at a

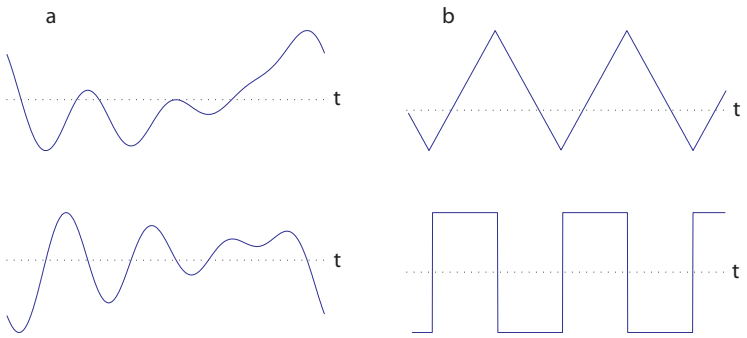

Figure 1.2: (a) Above, a continuous signal; below, its time derivative. The negative and positive going zero crossings of the derivative correspond to positive and negative peaks in the signal. (b) Above, a periodic sawtooth signal; below, its time derivative which is a periodic discontinuous square wave.

uniform rate. Δ is the interval between neighboring samples, and t° is an integer-valued index: $1, 2, 3, \ldots$, etc. Signal analyses are often most easy to describe when $\Delta = 1$. This results in no loss of generality. When there is no possibility of confusion, the Δ will be dropped. The signals or data handled by a digital computer are discrete not only in time but also in amplitude. This arises from the fact that the amplitude of a signal at a particular sampling instant is represented as a number within the computer, a number containing a limited number of digits or bits depending upon the computer's structure. To arrive at this numeric representation a continuous signal is first transformed into its A-discrete amplitude version by quantization in an analog-to-digital (A-D) converter. At each sampling time the quantization procedure assigns to the signal amplitude one of a finite number of levels. This level has a numeric value which represents the sample in subsequent data analysis computations. The subject of A-D conversion, or quantization, is discussed more thoroughly in Chapter 2.

Perhaps the simplest way of reconstructing a continuous signal from a set of its samples is shown in Fig. 1.3. Here the signal is assumed to remain constant at its sampled value for the time interval between the present and the next sample time. It is important to recognize that the sampling and interpolation process I can produce severe alterations of the signal depending upon the interrelationships between signal and sampling parameters. Two of the simplest errors are seen in Eig. 1.3 where in (a) a discontinuity is lost and in (b) a rapidly fluctuating component is suppressed because the sampling rate is too low. This type of error occurs regardless of how the interpolation between sampling instants is performed. A more thorough

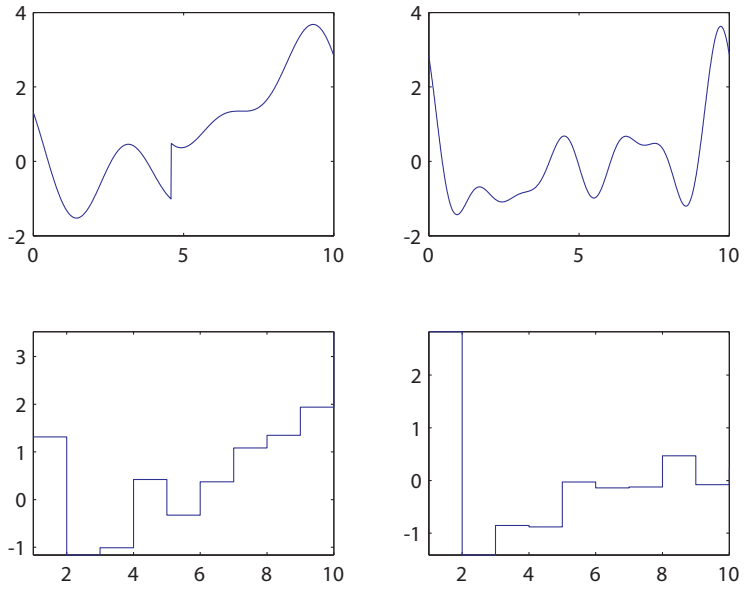

Figure 1.3: (a) Above, a signal with a discontinuity between t = 4 and 5; below, a reconstruction of that signal by interpolation with a constant value between sampling instants. (b) Above, another continuous signal fluctuating rapidly between the 4th and 5th sampling instants. Note how the same type of sampling reconstruction totally lacks evidence of the rapid fluctuation of the original.

discussion of sampling problems is also presented in Chapter 3. In some cases a signal is intrinsically T-discrete as for example is the count of the number of events occurring within an interval of time, such as the number of times an EEG waveform has a zero-crossing (a transition through zero amplitude) in one second. A second example is a list of measurements characterizing the structure of an object. It is important to note, however, that in the latter example the order in which the measurements are placed into a sequence may be of little or no importance. In temporal measurements or in measurements that are functions of a scanning process, the measurements follow one another in an order which must not be tampered with. The T-discrete or sampled version of a continuous signal is often used to construct estimates of parameters of the original continuous waveform, while an inherently discrete signal can, of course, never be meaningfully analyzed in this way. Thus far we have spoken of a signal as a one-dimensional quantity or variable. This is unduly restricting to many biological variables that can rightfully be called signals. For example, the amplitude of the EEG as measured at three different locations on the scalp is described by three coordinates. The net signal describing the observed EEG is therefore three-dimensional. If there were six recording locations, the observed EEG would be a six-dimensional signal. Each of the components of a multidimensional signal is distinguishable from a unidimensional signal and can be processed as such. There is an unavoidable burden placed upon a data processor employed to handle rapidly fluctuating multidimensional signals and keep up with these fluctuations, a burden .that increases with the dimensionality of the signal. Basically, the data processor must be able to sample each signal coordinate sequentially at a rate which preserves the information content in the signal as it is being processed. We will have more to say about this in Chapter 2.

1.3 Repetitive and Periodic Signals

Of considerable importance to biological signal analysis are repetition and periodicity. A signal is said to be repetitive if it has a particular waveform which recurs for as long as the signal persists. If, furthermore, this repetition occurs at uniformly spaced intervals in time, the signal is said to be periodic. Exact periodicity does not exist in biological signals unless external periodic stimulation is supplied to the preparation as is frequently done in the study of evoked responses from the nervous system. The periodicity of the stimulus is then looked for in the biological response. The EKG is an example of a biological signal which comes close to being periodic. Periodicity is important not only because it lends itself to relatively easily analyzable data, but also because it leads to the spectral concept of a signal. In this concept, to be discussed later in this chapter and throughout the

book, the signal is represented as the sum of sine waves of different frequencies and amplitudes. The periodic signal of duration T is of greatest interest to us here. It is represented by the equation

$$x(t) = x(t + mT), m = 0, \pm l, \pm 2, \ldots \tag{1.3}$$

with T being the period of the signal. The sawtooth wave of fig. 1.2 is an example of such a signal. Note that the signal persists from the infinite past to the infinite future.

1.4 Sampled Representation of a Signal

Let us assume that we sample a signal $x(t)$ without error once every Δ seconds throughout all time. We represent the signal by the discrete sequence of its sampled values, ignoring the behavior of the signal between sample times. The important question that arises is, how useful a representation of the signal is this set of ordered samples? We shall show here that the goodness of the discrete representation depends upon what is called the spectrum of the signal and its relation to the sampling rate. If sampling is done at the proper rate, it happens that this representation contains all the structure of the original signal. first, let us re-examine the signal reconstruction illustrated in fig. 1.3(b), where the signal amplitude is assumed to stay constant during the interval between successive samples. Such a reconstruction is useful when data are being inspected as they are received although, obviously, it almost always distorts the signal. The signal reconstruction that we shall discuss now is one that cannot be performed until all the signal data has been obtained. It therefore is not of practical value in the same sense that the previous method is; but it does demonstrate the degree to which the sampled data represent the original process.

The sequence of data samples obtained by the sampling process is:
$\ldots, x(-\Delta), x(0), x(\Delta), x(2\Delta), \ldots$. We now multiply each sample value $x(t^\circ\Delta)$ by the so-called 'sinc' function,

$$sinc(\frac{t - t^\circ\Delta}{\Delta}) = \frac{\sin[\pi(t - t^\circ\Delta)/\Delta]}{\pi(t - t^\circ\Delta)/\Delta} \tag{1.4}$$

This function is shown in Fig.1.4. It has the value of unity at $t = t^\circ\Delta$ and zero whenever t is any other integer multiple of Δ, t° being, as before, an integer. (It has the further important property that its Fourier transform, to be discussed later, has amplitude Δ when f is between $-1/2\Delta$ and $1/2\Delta$ and is zero for all other values of f.) The sinc function whose amplitude is $x(t^\circ\Delta)$ at $t = t^\circ\Delta$, and is zero at all other integer multiples of Δ, $u^\circ\Delta$.

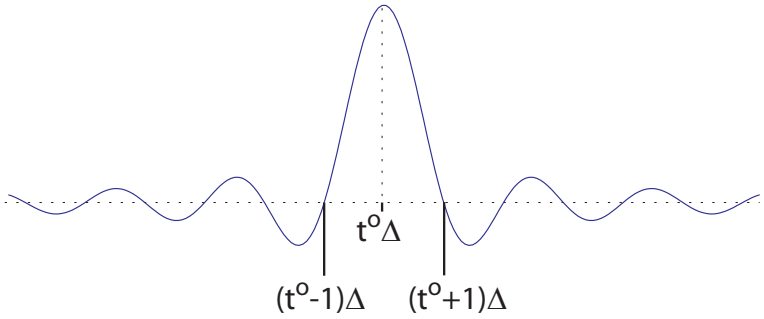

Figure 1.4: Fig. 1.4. The sinc function. The function is unity at $t = t\Delta$ and 0 at all other integer multiples of Δ.

That is,

$$x(t^\circ\Delta)sinc(\frac{t-t^\circ\Delta}{\Delta}) = \begin{cases} x(t^\circ\Delta), & t = t^\circ\Delta \\ 0, & t = u^\circ\Delta \end{cases} \tag{1.5}$$

Because of this, when all the sinc functions representing the signal at integer multiples of Δ are added together, we obtain the sum

$$x_s(t) = \sum_{t^\circ=-\infty}^{\infty} x(t^\circ\Delta)sinc(\frac{t-t^\circ\Delta}{\Delta}) \tag{1.6}$$

The value of $x_s(t)$ at each sample time $t^\circ\Delta$ is just the amplitude of the original sample obtained at that time, i.e., there is no interaction of samples at the sampling points. There is interaction, however, at all times between the sample points. In a sense, the sinc function provides a method of interpolating a smooth curve between the sample points $x(t^\circ\Delta)$. Now, it is possible to prove that if $x(t)$, the original function, has what is called its spectral bandwidth, F, smaller than $1/2\Delta$, the sum of the individual weighted sinc functions of Eq.XX1.5 will yield exactly $x(t)$ at all points in time, not just at the sample points. On the other hand, if the spectral bandwidth of $x(t)$ exceeds $1/2\Delta$, the reconstruction will not be perfect, the amount of error between $x(t)$ and $x_s(t)$ being related to the amount by which the bandwidth F exceeds $1/2\Delta$. When the sampling rate l/Δ is related to the bandwidth by $F = 1/2\Delta$, the rate is said to be the Nyquist sampling rate, a rate that is twice the signal bandwidth.

1.5 Fourier Series Representation of a Signal

Having pointed out the adequacy of sample values as a representation of a signal in terms of the relationship between sample rate and bandwidth, we must now put meaning into the term bandwidth. This can be done in the following way. Let us consider that we only know the behavior of $x(t)$ over a T second interval of time starting at $t = 0$. This is typical of what occurs in real situations. Since we have no knowledge of what $x(t)$ has done earlier than 0 or later than T, we assume that it repeats itself periodically with period T indefinitely. This is an artifice, but a valid one as long as our interest in confined only to what $x(t)$ does between 0 and T. $x(t)$ can be represented by the sum of a set of sine and cosine waves of different amplitudes and harmonically related frequencies infinite in number. This is its Fourier series representation which is given by

$$x(t) = \frac{1}{2}A_T(0) + \sum_{n=1}^{\infty}\left[A_T(n)\cos\frac{2\pi nt}{T} + B_T(n)\sin\frac{2\pi nt}{T}\right] \tag{1.7}$$

The lowest or fundamental frequency of the series is $1/T$. The amplitudes $A_T(n)$ and $B_T(n)$ of the components of this series are obtained from $x(t)$ by the equations

$$A_T(n) = \frac{2}{T}\int_o^T x(t)\cos\frac{2\pi nt}{T}\,dt$$

and

$$B_T(n) = \frac{2}{T}\int_o^T x(t)\sin\frac{2\pi nt}{T}\,dt\,. \tag{1.8}$$

The frequency of the nth component is n/T Hz. The nth frequency component of x(t) is defined by the two coefficients $A_T(n)$ and $B_T(n)$. If the amplitudes $A_T(n)$ and $B_T(n)$ are 0 whenever $f > F$, $x(t)$ is said to be bandlimited to the frequencies extending from 0 to F Hz.

From Eq.XX (1.7) it can be seen that the waveform of the observed segment of the signal determines the values of the Fourier coefficients uniquely. To obtain these coefficients, the amplitude of $x(t)$ must be processed at all values of time within the observation interval. When this is done the complete Fburier series so obtained will reconstruct the original waveform, if it is continuous, without error. Continuity generally prevails in biological signals although there are signals, such as those representing neuronal spike sequences, where continuity does not apply. These will be discussed later. Here we ignore continuous signals with discontinuities in them. For any signal, we can construct a curve relating the amplitudes of its $A_T(n)$ and $B_T(n)$ to the frequency $f_n = n/T$. Moreover, since $A_T(n)$ and $B_T(n)$ both pertain to the same frequency, we shall see that a more economical

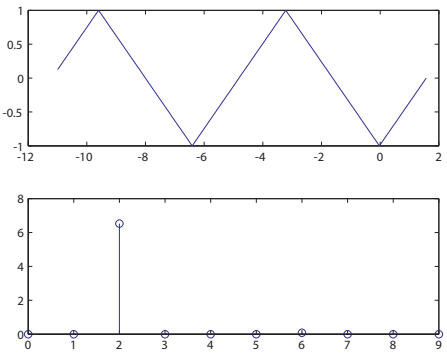

Figure 1.5: Fig. 1.5 (Up to a scale...) The spectrum $|X_T(n)|^2$ of the sawtooth wave from Fig. 1.2 (b) when the peak-peak amplitude is 2. The special coefficients are 0 for all odd values of n.

and significant plot is that of $|X_T(n)| = \sqrt{A_T^2(n) + B_T^2(n)}$ against frequency. An example of such a plot is shown in Fig. 1.5 for the sawtooth wave of Fig. 1.2(b). The period T of the wave is taken to be 1 sec. The value of the spectral coefficient $|X_T(n)| = 8/(\pi^2 n^2)$ for n even, and is 0 when n is odd or zero. Vertical lines are drawn with a height equal to the magnitude of $X_T(n)$. The existence of terms out to indefinitely large values of n is caused principally by the sudden changes (discontinuities) in the slope of the sawtooth at its peaks. $|X_T(n)|$ is referred to as the amplitude spectrum of $x(t)$, and $|X_T(n)|^2$ as the power spectrum, often shortened to spectrum. They are usually plotted as a function of frequency, n/T or n. The term power is employed because it is the square of an amplitude related to force (often voltage) and this is proportional to power. Because $|X_T(n)|^2$ is defined only for discrete frequencies corresponding to integer values of n, it is also called a line spectrum. More will be said of the power spectrum later.

Although we have thus far restricted the lower frequency limit of the spectrum to 0 Hz, it is useful to talk about the negative frequencies of a spectrum. Doing so introduces some simplifications into our dealings with signal spectra. Negative frequencies can be introduced by an alternative way of writing a Fourier series for $x(t)$, one that employs complex notation:

$$x(t) = \sum_{n=-\infty}^{\infty} X_T(n) \exp \frac{2\pi jnt}{T} \tag{1.9}$$

where the $x_T(n)$ are complex numbers given by

$$X_T(n) = \frac{1}{T}\int_0^T x(t)\exp\frac{-2\pi jnt}{T}d/,t \tag{1.10}$$

That this series is equivalent to the original expression is seen by considering the sum of the pair of terms corresponding to the integers $-n$ and n:

$$X_T(-n)\exp(-2\pi jnt/T) + X_T(n)\exp(2\pi jnt/T) \tag{1.11}$$

By employing the Euler formula, $exp(j\theta) = \cos(\theta) + j\sin(\theta)$, we obtain for this pair of terms,

$$[X_T(n) + X_T(-n)]\cos(2\pi nt/T) \quad + j[X_T(n) - X_T(-n)]sin(2\pi nt/T) \tag{1.12}$$

This has the same form as the right-hand side of Eq. (1.6). If $x(t)$ is real, as it is for the kinds of signals we consider, Eq. (1.11) must be real regardless of the integer value of n. Then, by algebraic manipulation of the real and imaginary quantities we find that

$$X_T(n) = [A_T(n) - jB_T(n)]/2 = X_T^*(-n), n = 1, 2, \ldots$$

and

$$X_T(0) = A_T(0)/2 \tag{1.13}$$

The asterisk denotes the conjugate complex value. Thus the complex series is a simple rearrangement of the real Fourier series. It has the advantage of introducing negative frequencies, those frequencies in the expansion corresponding to negative values of n. Although it is not necessary to employ negative frequencies in dealing with spectral properties of signals, a certain amount of simplicity and overall clarity results when this is done. This becomes even more apparent when extended to the more general Fourier integral treatment of signals. The complex series has also justified more fully our use of $X_T(n)$ to indicate the magnitude of the spectral component corresponding to frequency n/T. We call the frequencies associated with the Fourier series of Eq. (1.6) real frequencies. They are always positive. We call the frequencies associated with the complex Fourier series of Eq. (1.8) complex frequencies. The average power of a unit amplitude cosine wave of real frequency n/T is $1/2$, since $A_T(n) = 1$, $B_T(n) = 0$. This is equally divided between the component at complex frequencies $-n/T$ and n/T. Generally, from Eq. (1.12)

$$|X_T(n)|^2 = A_T^2(n) + B_T^2(n) = |X_T(-n)|^2 \tag{1.14}$$

The amplitude spectrum of a real x(t) is the plot of $|X_T(n)|$ as a function frequency n/T and it is symmetric about the origin. Each frequency component of the signal contributes a power equal to the square of its magnitude, $|X_T(n)|$. The sum of all these powers is the total power in the signal. There is also informatiox about signal structure in the phase of $X_T(n)$. This is given by

$$\theta_X(n) = \arctan[-B_T(n)/A_T(n)] \tag{1.15}$$

$\theta(n)$, when plotted as a function of n or frequency n/T, yields the phase spectrum of the signal.

1.6 Bandwidth Limited Signals

Consider now the special case where the signal is bandwidth (or band) limited to frequencies between 0 and $1/2\Delta$. The total time of observation of the signal is T seconds and is such that $T = N\Delta$. Starting at time 0, N samples of the signal are taken at Δ second intervals yielding the values $x(0), x(\Delta), \ldots x(t^{\circ}\Delta), \ldots x[(N-1)\Delta]$. This is sampling at the Nyquist rate. For simplicity, let N be even; it is not a confining restriction. Because of the bandwidth limitation, the Fourier series representation for the continuous $x(t)$ during this time span is given by

$$x(t) = \frac{A_T(0)}{2} + \sum_{n=1}^{N/2}[A_T(n)\cos\frac{2\pi nt}{T} + B_T(n)\sin\frac{2\pi nt}{T}] \tag{1.16}$$

which is the same as Eq. (1.6) except that the upper limit for the index n is now $N/2$ instead of infinity. The integer $N/2$ is associated with frequency $N/(2T) = 1/2\Delta = F$, the highest frequency present in $x(t)$. It is also true that $whenx(t)$ is bandlimited and periodic the coefficient $B_T(N/2) = 0$ (cf., Hamming, 1973). We can thus see from Eq. (1.15) that we need to evaluate a total of N coefficients for a complete representation of $x(t)$. These can be evaluated by means of Eq. (1.7) or they can be obtained from the N sample value in a manner to be discussed in Chapter 3. What should be emphasized here is that sampling at the Nyquist rate provides just as many time samples as are necessary to evaluate the Fourier coefficients uniquely. That is, the sample values in themselves contain all the information necessary to completely reconstruct $x(t)$ provided its band limit is related to the sampling interval Δ by the equation $F = 1/2\Delta$. Restating this in a slightly different way, as long as the T sec sample of signal $x(t)$ can be represented by a Fourier series, all of whose coefficients are 0 for the terms higher in frequency than $1/2\Delta$, the sampling process is guaranteed to represent all the information or structure in the signal.

When the signal bandwidth is larger than $1/2\Delta$, a number of difficulties ensue in the processing of the data. These we discuss in more detail in Chapters 2 and 3. Briefly however, the data processing proceeds just as though the signal were band limited. But when the Fourier coefficients are determined from the sample values, each may suffer an error whose size depends upon the amount by which the signal bandwidth F exceeds $1/(2\Delta)$. The greater this excess is, the greater the errors and the less adequate are the samples as a representation of the signal. Many rather serious misinterpretations are likely to arise if this situation goes unrecognized. It is up to the investigator to see to it that the bandwidth of the signal is suitable for the sampling rate employed. Caution is required. Preliminary spectrum analysis performed by instruments specifically designed for this is often called for.

Thus far we have not been concerned with whether the T sec signal segment would periodically recur during an arbitrarily long observation time, or whether the segment is one glimpse of an infinite number of possible manifestations of signal activity, none of which ever recur. The sampling process is indifferent to these alternatives. Nonetheless, there is a distinction to be made in the kinds of spectra associated with each. When the signal is band limited and truly periodic with period T, doubling the time of observation to $2T$ would yield a Fourier series in which the $A_T(n)$, $B_T(n)$ and $X_T(n)$ are different from 0 only when $n/2$ is an integer, where now $0 \leq n \leq T/\Delta$. If a $3T$ segment of signal were used and a Fourier series obtained from it, the nonzero terms would correspond to integer values of $n/3$, etc. That is, the frequency spectrum would exhibit components only at a discrete set of frequencies corresponding to harmonics of the basic period T.

Periodic signals thus possess discrete or line spectra, related to the repetition period, not the time of observation.

1.7 Autocovariance Functions and Power Spectra of Periodic Signals

The nature of a signal's temporal structure is often investigated by means of autocovariance function analysis. It is a method of comparing or correlating the signal with a replica of itself delayed in time. The autocovariance function takes on a continuous form for continuous signals and a discrete form for discrete signals. It provides an indication of the degree to which a signal's amplitude at one time relates to or can be inferred from its amplitude at another time. There are other measures of signal temporal variability but correlation thus far has proved to be the most useful though it is not without flaws. The autocovariance function receives its name by being an extension of the statistical covariance measures for random variables x and y. From statistics, the covariance of x and y, written $Cov(x, y)$, is

the average or expectation value of the product $(x-\mu_x)*(y-\mu_y)$ where μ_x and μ_y are the average values of x and y. Suppose now that $x = x(t)$ and $y = x(t+\tau)$. The covariance of $x(t)$ and $x(t+\tau)$ is seen to be a function of their time separation, τ . Because the covariance is that of an individual signal, it is called an autocovariance. If, in addition, $x(t)$ is periodic with period T and has zero mean, we can define the autocovariance function (acvf) for $x(t)$ as the average of $x(t)\cdot x^*(t+\tau)$:

$$x_{xx}(\tau) = \frac{1}{T}\int_0^T x(t)\cdot x^*(t+\tau)\,dt \tag{1.17}$$

Though $x(t)$ is real, its complex conjugate is used here to facilitate later consideration of the power spectrum associated with $x(t)$. If the mean of $x(t)$ is not zero, the right hand side of Eq. (1.16) is referred to as an autocorrelation function. It can be converted to an acvf by subtracting out the mean, μ_x. The product being averaged would then be $[x(t)-\mu]\cdot[x^*(t+\tau)-\mu]$. Unless otherwise stated, the time functions we deal with will be considered to have zero mean or to have had their nonzero means removed first. Because the signal here is periodic, averaging needs to be carried out only over the time interval T. A longer averaging time than this is of no value since it only repeats measurements of amplitude products already obtained. For this reason $c_{xx}(\tau)$ is itself periodic with the same period T as that of the signal. The time required to measure the acvf for one value of time separation is the period T. This means that as T increases, so does the computation time.

The data processing operation called for in Eq. (1.16) is a continuous one utilizing the signal amplitudes at all times during One of its periods. The computation must also be carried out for all possible time lags up to T and so it can be seen that unless some type of sampling procedure involving τ is possible, the total time required to estimate the acvf is infinite. Fortunately, when the signal is band limited, there is a valid sampling procedure that makes the computation feasible. It uses the sequence of signal samples Δ seconds apart that were shown to fully represent a band limited signal. The acvf of the sampled signal T sec in duration is formed by the products of each sample and a second sample delayed in time from it by τ° sample intervals. The $N(=T/\Delta)$ individual products are then summed and divided by the total rider of sample products to obtain

$$c_{xx}(\tau^\circ\Delta) = \frac{1}{N}\sum_{t^\circ=0} N-1 x(t^\circ\Delta)x^*[(t^\circ+\tau^\circ)\Delta] \tag{1.18}$$

If we now substitute for $x(t^\circ\Delta)$ and $x^*[(t^\circ+\tau^\circ)\Delta]$ their Fourier series representations as given by Eq. (1.8) and then interchange the order of summations, we

find that

$$c_{xx}(\tau^\circ\Delta) = \sum_{n=-N/2}^{N/2-1} |X_T(n)|^2 \exp \frac{j2\pi n\tau^\circ\Delta}{T} \tag{1.19}$$

Here, in substituting for $x^*[(t^\circ + \tau^\circ)\Delta]$ we have used the complex conjugate of the series in Eq. (1.8). This shows that the acvf of the sampled signal can be expressed by a Fourier series whose coefficients $|X_T(n)|^2$ are completely determined by those of the original signal. Henceforth we represent $|X_T(n)|^2$ by $C_{xx}(n)$. On occasion we will also use the notation $C_{xx}(f_n)$ where $f_n = n/T$, so as to relate this more easily to the spectrum of aperiodic signals. $C_{xx}(f_n)$ is the power spectrum of $x(t)$, the distribution of signal power or variance at the harmonically related frequencies f_n.

Now let us return to the definition of the acvf of a continuous periodic signal as given in Eq. (1.16). Here we also substitute the Fourier series representation of Eq. (1.8) for $x(t)$. Performance of integration and then summation yields

$$c_{xx}(\tau) = \sum_{n=-N/2}^{N/2-1} C_{xx} \exp \frac{j2\pi n\tau}{T} \tag{1.20}$$

When $\tau = \tau^\circ\Delta$, this is the same as Eq. (1.18). This shows that the acvf of the sampled band limited signal has the same values at the sample times as the acvf of the original signal. It can be shown further that $c_{xx}(\tau)$ is itself a band limited signal in the τ domain and therefore that it can be completely reconstructed at all values of τ by using the coefficients $C_{xx}(n)$. Thus the acvf of the sampled signal completely represents the acvf of the continuous signal. Note that $C_{xx}(n)$ is the previously defined power spectrum of $x(t)$ and is given by the inverse relationship

$$C_{xx}(n) = \frac{1}{T}\int_0^T c_{xx} \exp \frac{-j2\pi n\tau}{T}\, d\tau \tag{1.21}$$

This is an important relationship between the acvf and the power spectral density of the signal. We also point out that $C_{xx}(n) = C_{xx}(-n)$ and consequently that $c_{xx}(\tau) = c_{xx}(-\tau)$. Another important relation that applies to band limited periodic signals is

$$C_{xx}(n) = \frac{1}{N}\sum_{\tau^\circ=0}^{N-1} c_{xx}(\tau^\circ\Delta) \exp \frac{-j2\pi n\tau^\circ}{N} \tag{1.22}$$

This shows how the Fourier coefficients are related to the N values of the acvf at the times $\tau^\circ\Delta$. It will be discussed in more detail in Chapter 3.

Since $C_{xx}(n) = C_{xx}(-n)$, the true autocovariance function is defined by $N/2$ parameters in distinction to the N required for $x(t)$. What has happened is that the autocovariance procedure has removed the phase structure properties given by the $A_T(n)$ and $B_T(n)$ and left only the $C_{xx}(n)$ terms measuring the power of the individual frequency components that describe $x(t)$. It is important to note that the absence of phase information in the autocovariance function makes it impossible to deduce from the acvf the waveform of the signal that produced it. Thus an individual autocovariance function or power spectrum can be obtained from an infinite number of signals differing only in their phase structure.

1.8 Aperiodic Signals

In contrast to the periodic signal, the aperiodic signal would, when the observation time is increased to 2T, then 3T, etc., yield nonzero values for the $A_T(n)$, $B_T(n)$ and $X_T(n)$ regardless of the value of n. By making the observation time large enough, we can make the frequencies at which we measure the spectral intensity as close as we like. In the limit, as T becomes infinite, the lines merge to a continuous spectrum that is characteristic of aperiodic signals. Aperiodic signals are treated by means of a generalization of the Fourier spectrum, the Fourier transform,

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp j2\pi ft \, dt \tag{1.23}$$

$X(f)$ is referred to as the Fourier transform of the signal $x(t)$. $x(t)$ can be recovered from its transform by the inverse Fourier transfom,

$$x(t) = \int_{-\infty}^{\infty} X(f) \exp j2\pi ft \, df \tag{1.24}$$

The Fourier transform is useful not only with aperiodic signals, as for example the EEG where we deal with its power Spectral density, but also with transitory signals which exist for only a short period of time, such as the nerve impulse and the impulse response of signal filters to be discussed in Chapter 2. In this case the energy of the response is more important than its power and we deal with the energy spectral density.

1.9 Autocovariance Functions and Power Spectra of Aperiodic Signals

When we pass from the periodic signal to the aperiodic (by extending to infinity the period of repetition), the expression for $c_{xx}(\tau)$ becomes

$$c_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)x^*(t+\tau)\, dt \tag{1.25}$$

In the situation of the infinite interval, the Fourier power spectral density for the signal passes from a series to an integral representation similar to that given in Eq. (1.18). As a result, the relationships between autocovariance function and power spectral density for the aperiodic signal become (Jenkins and Watts, 1968)

$$C_{xx}(f) = \int_{-\infty}^{\infty} c_{xx}(\tau) \exp(-j2\pi f\tau)\, d\tau \tag{1.26}$$

while the inverse relationship is

$$c_{xx}(\tau) = \int_{-\infty}^{\infty} C_{xx}(f) \exp(-j2\pi f\tau)\, df \tag{1.27}$$

Both f and T can range from plus to minus infinity. Here, $C_{xx}(f)$ is the power spectral density of the signal x(t), the amount of signal power in the small frequency band from f to $f + df$. This pair of equations is referred to as a Fourier transform pair. The knowledge of either function permits unique determination of the other.

An idealized spectrum whose shape is somewhat typical of continuous signals is shown in Fig. 1.6. It has significant components below F, the cut-off frequency. As the frequency increases above F, the spectral intensity falls rather sharply. The width of the region below F in which $C_{xx}(f)$ is near its maximum value is the bandwidth of the signal. As with periodic signals, if the frequency components of the signal actually vanish at all frequencies above F, the signal is said to be band limited with bandwidth F.

Aperiodic signals that are band limited to $f = 1/2\Delta$ also can be represented exactly by their sample values at times A sec apart and these sample amplitudes permit estimation of the covariance function and the spectrum of the signal. The distinction between an estimate of a function and the function itself is made in Section 1.12. Some difficulties are encountered when a T sec segment of an aperiodic signal is considered. These difficulties affect the adequacy of the representation of the signal by its T-discrete version near the beginning and end of the segment. They arise when we consider an aperiodic signal to be one period of a periodic wave that

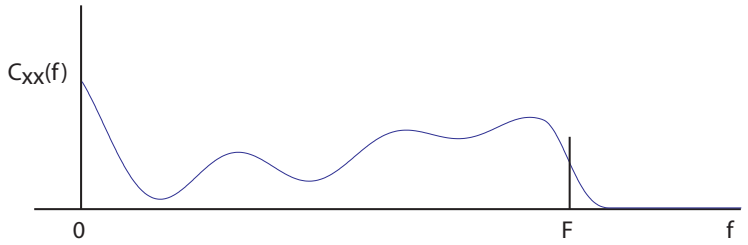

Figure 1.6: Fig. 1.6. A hypothetical spectrum $C_{xx}(f)$ of an aperiodic signal. F is the cutoff frequency.

repeats itself continually outside the time of observation. This artifice, comnly employed in the analysis of finite lengths of data, yields a discrete or line spectrum with components at integer multiples of $1/T$. The original signal, of course, has a continuous spectrum. Finally, since we have only a finite the to accumulate data, we can never obtain the precise autocovariance function and spectrum of the aperiodic signal regardless of whether there is noise interference. What we do obtain is estimates of them. The goodness of the estimates varies with the time available for observing the data. These are matters of great importance that are to be discussed in Chapter 3.

1.10 Cross covariance functions and cross spectra for a pair of periodic signals

There are many circumstances in which the data to be analyzed consist of two or more signals whose interrelationships are interesting. The relationship between an external stimulus and the several responses it gives rise to is also of considerable interest. The autocovariance function of a signal cannot cope with these matters because it deals only with the internal structure of an isolated signal. The analysis of signal interrelationships is a more complex affair. One approach to this general problem is via the use of the cross covariance function. A cross covariance function (ccvf) differs from the autocovariance function only in that the delayed signal $x(t+\tau)$ is replaced by $y(t+\tau)$, the delayed version of the second of the two signals being analyzed. The two signals are now denoted as $x(t)$ and $y(t)$. The cross covariance function is therefore an indication of the degree to which one signal's amplitude at one time relates to or can be inferred from a second signal's amplitude at another

time. If both signals have period T, the cross covariance function also will have the same period and can be written

$$c_{xy}(\tau) = \frac{1}{T}\int_0^T x(t)y^*(t+\tau)\,dt \tag{1.28}$$

The ccvf is obtained by continuous processing of the two signal waveforms.

For the ccvf there is a spectral counterpart, the cross spectrum which has a relationship to the ccvf similar to that which the spectrum has to the acvf. To see this we express the periodic $c_{xy}(\tau)$ of Eq. (1.27) in terms of the complex Fourier series:

$$c_{xy}(\tau) = \sum_{n=-N/2}^{N/2-1} C_{xy}(n)\exp\frac{j2\pi n\tau}{T} \tag{1.29}$$

It is the set of coefficients which we call the cross spectrum. $C_{xy}(n)$ is given by

$$C_{xy}(n) = \frac{1}{T}\int_0^T c_{xy}(\tau)\exp\frac{-j2\pi n\tau}{T}\,d\tau \tag{1.30}$$

If we then substitute for the ccvf the right-hand side of this equation and replace both $x(t)$ and $y^*(t+\tau)$ by their Fourier expansions, we obtain, after carrying out the indicated integrations,

$$C_{xy}(n) = X_T(n)Y_T^*(n) \tag{1.31}$$

$X_T(n)$ and $Y_T(n)$ are the Fourier coefficients for signals x and y. Thus the cross spectrum is the complex conjugate product of the Fourier series for each of the constituent signals. If we now substitute Eq. (1.30) in Eq. (1.28), we obtain

$$c_{xy}(\tau) = \sum_{n=N/2}^{N/2-1} X_T(n)Y_T^*(n)\exp j2\pi n\tau T \tag{1.32}$$

This is to be compared with Eq. (1.21) which relates the ccvf to its spectrum.

The ccvf and cross spectrum can be extended as well to aperiodic signals. This involves the same limiting procedure as T in Eq. (1.27) that was used with the autocovariance function. In this instance we have the Fourier transform pair relating the ccvf and cross spectrum:

$$C_{xy}(f) = \int_{-\infty}^{\infty} c_{xy}(\tau)\exp(-j2\pi f\tau)\,d\tau \tag{1.33}$$

$$c_{xy}(\tau) = \int_{-\infty}^{\infty} C_{xy}(f)\exp(j2\pi f\tau)\,df \tag{1.34}$$

Cross covariance functions and cross spectra can be defined for sampled signals. The ccvf of two sampled periodic signals is defined by

$$c_{xy}(\tau^\circ\Delta) = \frac{1}{N}\sum_{t^\circ=0}^{N-1} x(t^\circ\Delta)y^*[(t^\circ+\tau^\circ)\Delta] \tag{1.35}$$

We can proceed as before to show that

$$c_{xy}(\tau^\circ\Delta) = \sum_{n=-N/2}^{N/2-1} X_T(n)Y_T^*(n)\exp\frac{j2\pi n\tau^\circ\Delta}{T} \tag{1.36}$$

The cross spectrum, $C_{xy}(n)$, originally defined in Eq. (1.29), is also given by

$$c_{xy}(n) = \frac{1}{N}\sum_{\tau^\circ=-N/2}^{N/2-1} c_{xy}(\tau^\circ\Delta)\exp\frac{-j2\pi n\tau^\circ\Delta}{T} \tag{1.37}$$

This is the cross spectral counterpart of Eq. (1.23). Equation (1.35) shows that the ccvf is defined by N coefficients of the form $X_T(n)Y_T^*(n)$. The original signals each require N coefficients, the set of $X_T(n)$ and $Y_T(n)$ to describe them. Since the coefficients appear together in Eq. (1.35) as products, there is no way of separating them unless either $x(t)$ or $y(t)$ is also known. Thus the ccvf and its companion cross spectral density do not by themselves preserve all of the information in the two signal waveforms. Remember that the same statement was made of the acvf and spectrum of a single signal.

1.11 Summary: properties of covariance functions & spectra

There are several properties of covariance functions spectra that are worthwhile noting here. They are stated in terms of the continuous covariance functions but, except for A3, apply equally well to covariance functions and spectra of sampled signals. No proof of these properties is given here. They are easy to derive and more will be said of them in Chapter 3.

1.11.1 Autocovariance functions and power spectra

1. $c_{xx}(T)$ is an even function of time, i.e., $c_{xx}(\tau) = c_{xx}(-\tau)$.
2. The maximum value of $cxx(\tau)$ occurs at $\tau = 0$.

3. If $x(t)$ is continuous, $c_{xx}(\tau)$ is continuous also.

4. The power spectral density of $x(t)$ is real and an even function of frequency: $C_{xx}(f) = C_{xx}(-f)$.

1.11.2 Cross covariance functions and cross spectra

1. $c_{xy}(\tau)$ is not necessarily an even function of time. In general, $c_{xy}(\tau) = c_{yx}(-\tau)$.

2. The maximum value of $c_{xy}(\tau)$ does not necessarily occur at $\tau = 0$.

3. If $x(t)$ and $y(t)$ are continuous, $c_{xy}(\tau)$ is continuous also.

4. The cross spectral density of $x(t)$ and $y(t)$ is complex Tnd $_{xy}(f) = C^*_{xy}(-f)$.

Though these by no means exhaust the interesting properties of covariance functions and spectra, they are the most important in terms of a working knowledge useful for ordinary signal analysis problems.

1.12 Random or probabilistic signals

In the previous section we have considered aperiodic signals and the manner of representing them in terms of their covariance functions and spectral densities. To do this we employed the artifice of considering such signals to be an extension of the more simple periodic signals with the period of these signals becoming infinite. This is a useful approach to deterministic but aperiodic signals since it provides a straightforward frequency domain description .

A simple example of a deterministic aperiodic signal is i

$$x(t) = \sin 2\pi f t + \sin 2\pi\sqrt{(2)} f t \tag{1.38}$$

Although it is impossible to find a finite value of t corresponding to a repetition period for this signal, it does possess a power spectrum and an autocovariance function, both of which can be found easily. The signal is bandwidth limited, nonrandom and aperiodic, although its two components individually are periodic. Its behavior for all time is known from its functional form. We can infallibly predict its future behavior and also state how it behaved in the remote past. To see how this can be done, we recall from calculus that an explicit function of time can be described exactly for all time in terms of a Taylor series expansion provided that all its derivatives are known at some arbitrary time. Its past and future history are completely specified by the values of these derivatives at that time. If, on the

other hand, not all the higher derivatives exist (in the sense that they ”blow up”), or if they cannot be measured, and practically they cannot, then the past and future history of the function or signal cannot possibly be determined infallibly.

Nondeterministic signals, and neurobiological signals are generally in this category, cannot be inherently described by an explicit equation valid for all time either because (a) although it may be possible to determine one, we do not have all the information at hand to permit doing so, or (b) it is inherent in the nature of the signal that it cannot be so described. Both cases are of considerable biological interest, with the latter case being especially so for both theoretical and practical reasons. The principal examples of nondeterministic neurobiological signals are (a) the spike activity of a single neuron or a small group of isolated neurons, and (b) the electroencephalogram (EEG). The signals are nondeterministic because the mechanisms responsible for them are subject to internal and external influences which can never be completely described. Signals that have these properties are spoken of as being random for their behavior follows or seems to follow probabilistic rather than deterministic laws. They are manifestations of random processes, biological or nonbiological, that are themselves governed by probabilistic laws. In describing random signals we speak of their probability density functions, their means, variances, covariance functions, power spectra, and other statistical measures, and not about their functional descriptions except where it is occasionally useful to do so, as with the alpha bursts of the EEG. The probabilistic nature of a biological signal may not be entirely due to its generating process. Quite often, such a signal is observed in a background of other electrical activity unrelated to it. This activity is considered to be noise and may arise from other sources within the nervous system or it may arise from the measuring instrumentation itself. In either case its presence obscures the signal of interest making it more difficult to detect and analyze. Even though the signal being observed may be deterministic or nearly so, as in the case of the cochlear microphonic, its combination with the interfering noise makes the resulting mixture also qualify as a random signal. The reason is that noise is itself an example of a random process, different from the signal mainly in that the process that gives rise to it is not the one being studied. we exclude from our discussion of noise such phenomena as power line interference and stray electroma etic emissions from radio sources. Troublesome though these may be, they can be eliminated by careful laboratory practices. The noise which we are considering is an inherent basic constituent of an emerinent. It may be minimized to some extent but can never be eliminated. Sanewhat different but also noiselike are the data corrupting effects produced by signal quantization and by jitter in the time of smpling of the signal. These effects are inherent in the data processing operations and are treated in Chapter 2.

Random signals are best understood by considering the properties of a collec-

tion or ensemble of them as generated by their associated random process. This ensemble of signals characterizes the process. It can be, for exanple, a collection of all particular samples of signal of some given length T generated by the process. Each member of the collection is a unique function of time different from the others and is referred to generally as a sample function or as a realization. !tb avoid confusion with the physical sampling process, we will adopt an alternate designation, specimen function. The specimen functions of an ensemble might be, for example, a collection of ten-minute parietal lobe EEGs obtained from awake normal males between the ages of 20 and 30. A wide variety of EEG wavefoms may be included in the ensemble, each being a different realization generated by the underlying process. The fact that there can be substantial differences within an observed collection of specimen functions often leads to the question of whether or not there is but a single process responsible for the observations. Sometimes the specimen functions are so different as to make it obvious that they come from different processes; in other cases the differences are more subtle and considerable uncertainty arises as to whether more than one process is at work. It is possible to test the hypothesis of a single process being the source of the collected specimen functions although we shall not explore this problem. Some aspects of hypothesis testing are discussed in Section 1.17.

The existence of an ensemble of specimen functions makes it possible to describe the generating process in terms of the statistical measures of the ensemble. These measures are taken across the ensemble, each specimen function being but one member of the population. This concept of the statistical measure of an ensemble is to be distinguished from the previously employed time averages that are performed on a single member of an ensemble. Time averages tell us only of the properties of the individual ensemble member. They are rather easy to perform experimentally. When the averaging time is taken to the limit, infinite time, we have the definitions of the mean, variance and covariance function of the signal. Ensemble averages, in contrast, describe the overall properties of the process in terms of the ensemble. Thus, we can speak of the expected value of the ensemble of functions $x(t)$ at a particular time, $E[x(t)]$, or the variance $var[x(t)]$ or the autocovariance function $E[x(t)x(t+\tau)]$. But they are more difficult to deal with experimentally because they require measurements of many ensemble members. However, the value of the ensemble approach is that it leads to a far more penetrating understanding of the random process. The concept of a random process and its associated ensemble of specimen functions also applies not just to continuous siqals but so to the sampled version of a continuous random signal, to a screte signal, or to some function derived from originally connuous or discrete signals. An example of the last is the number of alpha bursts per ten-second interval as observed in each of the ten-minute EEGs previously described.

Suppose we now consider the value of a specimen function $x(t)$ at a particular time t'. Each member of the ensemble $x(t)$ will range over a set of permissible values in a probabilistic way that is determined by the random process itself. $x(t')$ is thus said to be a random variable. It is in fact a function whose value depends upon the many underlying events associated with the random process. To illustrate, an observer monitors the EEG tracing of a normal sleeping adult through the night in order to study its fluctuating patterns. He classifies the sleep status of the individual at any time into one of six different states: awake, stages 1 through 4, and rapid eye movement (REM) sleep. He does this by analyzing the fluctuations in the EEG during the minute preceding each classification. He proceeds to do so for a number of subjects for each of which he constructs a chart of the type shown in Fig. 1.7. The sleep states are assigned values 0 to 5. The $x(t)$ resulting is a specimen function of the sleep process. The value of $x(t)$, at $1AM$, say, $x(0100)$ is a random variable which can take on one of six different values for each of the subjects. The frequency with which the different values occur is determined by the sleep process and the observer's judgments of the EEGs associated with it. Assuming the validity of the observer's procedures, the result is a new, A-discrete, T-continuous signal derived (or filtered) from the EEG. Each of the six possible levels of the signal corresponds to an event, the sleep state of the subject, and the six possible events cover all the possible states that the subject can be in at any time. We can refer to these events in terms of an event or sample space. The sample space we have used here has a single dimension, depth of sleep, and it has a finite number of events in it, six. More generally, event spaces can be multidimensional and they can also be continuous with an uncountably infinite number of events possible. An example of a single-dimensional sample space with an infinite number of events in it is the temperature of a particular location of the body. If sleep were defined in terms of the original six states of the EEG and temperature of the hypothalamus, say, the event space of interest would be twodimensional, one dimension being discrete and the other continuous. An example of a five-dimensional continuous event space is the possible set of EEG voltages from a particular electrode at five instants that are one second apart. Sample spaces are used to describe specimen functions and the processes they arise from, and can be discrete or continuous according to whether the specimen functions are A-discrete or continuous. This is regardless of whether the specimen functions are T-discrete or continuous. In the sleep example the two-hour records of sleep states are A-discrete, T-continuous.

By making a small change in the manner of performing the previous experiment we can obtain a T-discrete specimen function instead. This would be accomplished by periodic examination of one-minute segments of the EEG at fifteen-minute intervals, say, and classification of them into sleep stages at those times. Under these circumstances the number of random variates is equal to the number

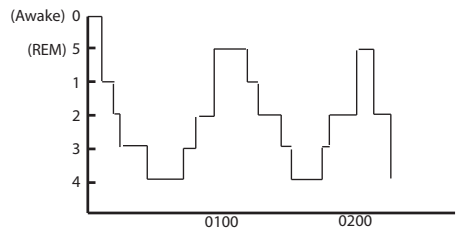

Figure 1.7: Fig. 1.7. Sleep as classified by the ongoing EEG into one of six possible states existing between the hours of midnight (0) and 2 AM (0200).

of samples obtained for each specimen function while the properties of the individual random variables are the same as under the continuous case.

The two kinds of spaces, specimen space for the specimen functions and sample space for the random variables, should not be confused. Since a specimen function is composed of the sequence of particular values assumed by a random variable, it may be useful to note that a point in specimen space can be considered to represent a particular trajectory traveled by the random variable in sample space as the sampling proceeds.

1.13 Some important probability distributions

1.13.1 Probabilistic description of dynamic processes

Basic to the study of dynamic probabilistic processes is a knowledge of the statistical properties of their random variables. This applies, first of all, to the derivation of the theoretical behavior of these processes, for, in most situations, they can be analyzed only after the probabilistic laws have been specified. It applies, secondly, to the validity of experimental data analyses which are generally dependent upon how the data conform to these laws.

In the case of continuous processes, the Gaussian probability distribution has been found to be by far the most applicable to them. Most of the work done on dynamic processes has been centered upon those that have Gaussian properties insofar as the amplitude fluctuations of the random variables are concerned. It is worth noting, moreover, that many of the statistical techniques that have been developed for testing Gaussian processes have also been found applicable to the study of some non-Gaussian processes. Such tests have been labeled ”robust” for this reason and in recent years extensive efforts have been devoted to the development of such statistical techniques. Nonetheless, it is the Gaussian probability

distribution which is basic to the understanding of continuous processes. We shall, therefore, summarize some of its basic properties.

Another distribution of great importance to both continuous and point process analysis is the chi-squared distribution. It arises not because it is a description of either continuous or point process random variables but because it gives an effective way of dealing with the sums of Gaussian or exponential random variables that are encountered in the statistical analyses of long records of data. The chi-squared distribution offers a compact representation of such data and also is a help in developing insights into the strengths and weaknesses of a variety of statistical tests.

A third probability distribution that is encountered extensively in neurophysiological work is the exponential distribution. It finds its application in the study of sequences of action potentials generated by individual neurons. These are sequences in which the times of occurrence of events are the only data of importance. The processes generating the events are referred to as point processes and we shall have more to say about them in Chapters 6 through 8. In the remaining part of this section we shall briefly summarize some of the basic properties of the Gaussian, chi-squared and exponential distributions. A more detailed exposition of them may be found in such standard texts as Mood (1950) and Cram (1946).

1.13.2 The Gaussian distribution

A random variable X is said to be Gaussianly or normally distributed if its probability density function is

$$prob\{x < X < x + dx\}/dx = p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \qquad (1.39)$$

where x can take on any positive value or negative value. The mean of the random variable is μ and its variance is σ^2. These follow from the definition of the nth moment and nth central moment of a random variable:

$$E[x^n] = \int_{-\infty}^{\infty} x^n p(x)\, dx \qquad E[(x-\mu)^n] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx \qquad (1.40)$$

Also, the second central moment $E[(x-\mu_x)^2] = var(x)$. The Gaussian or normal distributlon occurs so frequencly as to warrant a special notation. A Gaussian random variable with mean μ and standard deviation σ is said to be "normal (μ, σ)".

All the odd order central moments of a normal random variable are zero. This follows from the fact that the Gaussian density function is symmetrical about its

mean. The even moments of the normal $(0, \sigma)$ random variable are given by

$$E[x^n] = 1 \cdot 3 \cdot \ldots (2n-1)\sigma^{2n} \tag{1.41}$$

If we have a sum of N random variables, x_i, the sum of their means is the mean of their sum. If the random variables are independent, the sum of their variances is the variance of their sum.

If the random variables are also normal $(\mu_{x_i}, \sigma_{x_i})$, then their sum y is also normal with mean and variance given by

$$\mu_y = \sum_{i=1}^{N} \mu_{x_i} \qquad \text{and} \qquad \sigma_y^2 = \sum_{i=1}^{N} \sigma_{y_i}^2 \tag{1.42}$$

When the x are normal and identically distributed (μ, σ), then y is normal $(N\mu, N\sigma)$. Should the x_i not be independent, their sum will still be normal but the sum of their variances will not be equal to the variance of their sum.

A useful characterization of a random variable is its coefficient of variation (cvar), the ratio of its standard deviation to its mean. For the sum of N identically distributed normal random variables we have

$$\text{cvar}[y] = \frac{\sigma_y}{\mu_y} = \frac{\sigma}{N1/2\mu} \tag{1.43}$$

This result will be applied to signal averaging as discussed in Chapter 4.

1.13.3 The Chi-Squared Distribution

The chi-squared distribution describes a family of probability distributions. The first member of the family is the distribution of the random variable χ^2 which is the square of a normal random variable (0,l). χ^2 is restricted to values that are 0 or greater.

In this case we have the probability distribution

$$\frac{\text{prob}[x < \chi^2 \leq x + dx]}{dx} = p(x) = \frac{x^{-1/2}\exp(-x/2)}{\sqrt{2\pi}} \tag{1.44}$$

The mean and variance of this distribution are $E[\chi^2] = 1$ and $var[\chi^2] = 2$. More generally, the random variable χ_N^2 is defined as the sum of the squares of N independent and identically distributed normal random variables $(0, l)$. The subscript N, since it represents the number of independent normal random variables

composing χ^2_N, is often referred to as the number of degrees of freedom $(d.f.)$ of the chi-squared random variable.

$$\chi^2_N = \sum_{i=1}^{N} x_i^2 \tag{1.45}$$

The probability density function for χ^2_N is given by

$$p(x) = \frac{x^{n/2-1}\exp(-x/2)}{2^{N/2}\Gamma(N/2)} \tag{1.46}$$

where $\Gamma(\cdot)$ is the well-known gamma function.

Since χ^2_N is just the sum of N random variables, its mean and variance are seen to be given by

$$E[\chi^2_N] = N, \qquad var[\chi^2_N] = 2N \tag{1.47}$$

When the x_i's are independent and normal (O, σ), the sum of N of them is distributed according to $\sigma^2\chi^2_N$. Hence, $E[\sum x_i^2] = N\sigma^2$ and $var[\sum x_i^2] = 2N\sigma^4$. It can be shown that as N becomes large, the distribution of χ^2_N approaches the normal with mean and variance given by Eq. (Just above). This is often a useful approximation when $N \leq 30$.

The sum S_K of the squares of K independent normally distributed random variables is not distributed according to chi-squared when the random variables do not have identical distributions. Nonetheless, the distribution of this sum is sufficiently close to such a chi-squared distribution as to justify its approximation as such. To arrive at the approximation, one merely finds the hypothetical χ^2 random variable which has the same mean and variance. Thus one sets

$$E[S_K] = (d.f.)\,\sigma^2 \qquad \text{and} \qquad var[S_K] = 2(d.f.)\sigma^4 \tag{1.48}$$

d.f. here takes the place of N in the preceding paragraph.

This yields

$$d.f. = 2\frac{E^2[S_K]}{var[S_K]}, \qquad \text{and} \qquad \sigma^2 = \frac{var[S_K]}{2E[S_K]} \tag{1.49}$$

The sum of the K squared normal random variables is said to possess the number of "equivalent" degrees of freedom given by Eq. (1.47). A similar technique can be used with the sum of $K = \chi^2_2$ random variables that do not have the same mean and variance. If they were identically distributed, their sum would be chi-squared with $2K$ of freedom; if not, the number of equivalent degrees of freedom is always less. Note also from Eq. (1.47) that

$$(d.f.)cvar^2[S_K] = 2 \tag{1.50}$$

These aspects of the chi-squared distribution have important applications to spectral smoothing discussed in Chapter 3.

1.13.4 The Exponential Distribution

The probability density function for the exponentially distributed random variable is given by

$$p(x) = \begin{cases} \nu \exp(-\nu x), & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{1.51}$$

The mean and variance of the exponential random variable can easily be shown to be

$$E[X] = 1/\nu \qquad var[x] = E[x^2] - E^2[x] = 1/\nu^2 \tag{1.52}$$

From this it follows that $cvar[x] = 1$. The sum of M independently and identically distributed exponential random variables is distributed according to what is known as a gamma distribution:

$$p(x) = \frac{\nu(\nu x)^{N-1}}{(N-1)!} \exp(-\nu x) \tag{1.53}$$

The mean and variance of this sum are, as might be anticipated

$$E[x] = N/\nu \qquad var[x] = N^2/\nu^2 \tag{1.54}$$

We show the form of the gamma distribution because it is closely related to the chi-squared distribution. If we consider a random variable which is 2ν times the random variable in Eq. (1.49), i.e., a random variable whose mean is 2 , we find that the sum of such random variables has a χ^2_{2N} distribution. This property of sums of exponentially distributed random variables will be discussed more in Chapter 6. It is also true that as N becomes large, the sum of N exponentially distributed random variables becomes nearly normal with mean $2N$ and variance $4N$.

1.13.5 Ensemble Autocovariance Functions

Statistical measures of ensembles of signals generated by random processes are defined in terms of the probability density functions (PDFs) or cumulative distribution functions (CDFs) of the ensemble. The mean of the ensemble x(t) is given by

$$E[x_t] = \int x_t p(x_t)\, dx_t \tag{1.55}$$

where for convenience we have written $x(t)$ as x_t. $p(x_t)$ is the PDF of the ensemble of signals at time t and the integration is over all values of x_t. Here x_t represents either a continuous or a sampled random variable. $p(x_t)dt$ is the probability that at time t a member of the signal ensemble will take on some value between x_t and $x_t + dx_t$ Similarly, the variance of x_t is defined by

$$var[x_t] = E[x_t - \mu_{xt}] = \int (x_t - m_{xt})^2 p(x_t)\, dx_t \tag{1.56}$$

where $\mu_{xt} = E[x_t]$ The ensemble autocovariance function (acvf) is defined in terms of the expected value of the product of the amplitudes of the signal at t and u:

$$cov[x_t, x_u] = c_{xx}(t, u) = E[(x_t - \mu_{xt})(x_u - \mu_{x_u})^*] \tag{1.57}$$

The asterisk denotes the complex conjugate. Even though x is real (making $x^* = x$), its presence facilitates later consideration of the spectrum of the process. In most cases the average value of x is of no interest and can be subtracted from the data. This makes $\mu_{xt} = 0$ for all t. The covariance function is then just the product $E[x_t x_u^*]$. Determination of the ensemble acvf requires knowledge of the ensemble at times t and u:

$$c_{xx}(t, u) = \int (x_t - \mu_{xt})(x_u - \mu_{x_u})^* p(x_t, x_u)\, dx_t\, dx_u \tag{1.58}$$

The integration is over all values of x_t and x_u. While the ensemble acvf superficially bears little resemblance to the time acvf of an individual signal, Eq. (1.24), it will be shown later that the two are in fact closely related and in many important instances are equal. The acvf is a measure of how the fluctuations of the signal amplitude at two different times are related to one another. A positive covariance indicate that when one is greater (or less) than its mean value, the other tends to be also. A negative covariance, on the other hand, indicates that when one is greater than its mean, the other tends to be lower than its mean. The normalized autocovariance function is

$$\rho_x(t, u) = \frac{cov[x_t, x_u^*]}{var[x_t] var[x_u]} \tag{1.59}$$

It ranges in value from -1 to $+1$.

If a pair of variables, herex_t and x_u, are statistically independent, it is generally true that

$$E[x_t, x_u] = E[x_t] E[x_u] \tag{1.60}$$

It follows that there is 0 covariance between statistically independent random variables. Random variables which have 0 covariance are said to be uncorrelated.

However, it is not necessary for a pair of random variables to be independent in order for the mean of their product to be equal to the product of their means. Thus, lack of correlation does not imply statistical independence except in special instances. Gaussianly distributed random variables are particularly interesting in this regard. If x_t and x_u are Gaussian random variables and uncorrelated, then they are also statistically independent. This is the most commonly encountered case in which zero correlation implies statistical independence. Serious errors can often result if one assumes statistical independence only on the basis of lack of correlation. An example of two random variables that are uncorrelated but statistically independent is $x = sin\theta$ and $y = sin2\theta$, where θ is some arbitrary random variable.

1.14 Ensemble Autocovariance and Cross Covariance Functions, and Stationarity

Some basic properties of the ensemble autocovariance function deserve special mention:

- It is an even function of time:

$$c_{xx}(t,u) = c_{xx}(u,t) (1.59)$$

- Its maximum value occurs when $t \approx u$ and is the variance of the random variable:

$$c_{xx}(t,t) = var[x_t] \geq c_{xx}(t,u)$$

If a process possesses an autocovariance function in which times t and u always appear in the form of a time difference $t-u \approx \tau$, the process is said to be stationary. The covariance properties are then dependent only upon relative times, not upon any absolute value of time. Thus in terms of covariance properties, the process is the same throughout all time. The covariance function notation for a stationary process can be written as $c_{xx}(\tau$ and we shall generally do so. For covariance stationary processes $c_{xx}(0) = var[x_t] = \sigma_x^2$ is the average power of the process contributed by the fluctuations. An example of a covariance stationary process is the white noise $n(t)$ generated in electrical resistors. The amplitude of white noise at any one time is statistically independent from its value at any other time. Its covariance function can be shown to be:

$$c_{nn}(t,u) = c_{nn}(\tau) = \sigma_n^2 \rho(t) \quad (1.61)$$

This means the noise amplitudes at two different times are without correlation.

when we are interested in correlating the behavior of specimen functions belonging to two different ensembles, we employ the ensemble cross-covariance function (ccvf). Thus, for specimen functions $x(t)$ and $y(t)$

$$c_{xy}(t,u) = E[(x_t - \mu_{x_t})(y_u - \mu_{y_u})^*](1.62) \tag{1.62}$$

As with the autocovariance function, average values are often of little interest and can be removed from the data. This makes the ccvf equal to $E[x_t y_u^*]$. In contrast to the autocovariance function, the cross-covariance function is not an even function of time nor does it always have a maximum at $t = u$ or at $\tau = 0$ when the processes are stationary.

An example of the dependency of the autocovariance function upon the times t and u is the autocovariance function that would obtained from the membrane potentials of a population of cells that have been damaged in exactly the same way at $t = 0$ and then gradually healed. First there is a sudden depolarization in membrane potential at the instant of injury. Then, during the healing process there is a slow recovery of the resting membrane potential to its pre-injury level. Different cells have different initial changes in membrane potential and different rates of recovery or ghearling. For extreme but useful simplicity, let the membrane voltage during early stages of recovery be represented by the equation

$$v(t) = V_0 + kt \tag{1.63}$$

V_0 is the initial value of membrane potential immediatly after injury and k the initial rate of recovery. Both are assumed independent random variables with means $\overline{V}_0$ and $\overline{k}$ respectively. The autocovariance function for v is given by

$$c_{vv}(t,u) = E[(v_t - \overline{v}_t)(v_u - \overline{v}_u)^*](1.64) \tag{1.64}$$

After some straightforward manipulations, this becomes

$$c_{vv}(t,u) = var[V_0] + t.u.var[k] \tag{1.65}$$

This is a reasonable approximation to the covariance function in the tne early stages of recovery where linear Eq. (1.63) is a valid representation of the membrane potential. Notice (a) that the covariance function was obtained from the variances and without the knowledge of the probability distributions of the variables; and (b) that the covariance function is a function of both t and u. Only variances were required. A process of this type is said to be evolutionary.

Covariance stationary processes form a broad class of starlonary random processes, processes in which the underlying probabilistic mechanisms up to the second order joint density function do not change with the passage of time. If the

nth order joint distribution function of some ensemble were obtained at times $t_l, \ldots, t_m$ and compared with a similar joint distribution function obtained at later times, $t_l + \tau, \ldots, t_m + \tau$ and found to be the same regardless of how high m is or how large τ was chosen, the process would be said to be strictly stationary. This is a rather stringent condition for stationarity, if for no other reason than that it is seldom feasible to measure joint statistics beyond second or third order. Covariance stationarity, also called wide sense or second order stationarity, is more practical to deal with. Only the second order joint distributions need be independent of time. Many of the random signals involved in the study of biological processes are, or can effectively be considered to be, covariance stationary. Unless there is possibility for confusion, henceforth we abbreviate covariance stationary to stationary.

In the stationary situation. it is possible to show (Davenport and Root 1958), that the ensemble autocovariance function and the power spectrum of a random process are related to each other by the Fourier transform pair:

$$C_{XX} = \int_{-\infty}^{\infty} c_{xx}(\tau) \exp(-j2\pi f\tau)\, d\tau \qquad c_{xx}(\tau) = \int_{-\infty}^{\infty} c_{xx}(f) \exp(j2\pi f\tau)\, df \tag{1.66}$$

This relation was spoken of previously in terms of individual signals; here it is stated in terms of the ensemble autocovariance function and power spectrum of a process.

Some indication of how this relation arises can be obtained by representing the zero mean random periodic signal $x(t)$ and its delayed complex conjugate $x^*(t+\tau)$ by their complex Fourier series. The Fourier coefficients $X_T(n)$ and $X_T(m)$ are uncorrelated (Jenkins and Watts, 1968). That is, $E[X_T(n)X_T^*(m)1 = |X_T(n)|^2$ when $m = n$ and 0 otherwise. The Fourier transform of $E[x(t)x^*(t+\tau)]$, the acvf of $x(t)$, is then found to be given by the expression $\sum_{n=infty}^{\infty} |X_T(n)|^2 \delta(f - k/T)$. This is the power spectrum of the random periodic signal. Of particular interest is the situation when $\tau = 0$. Here we have

$$c_{xx}(0) = \sum_{\infty}^{\infty} C_{cc}(f)\, df \tag{1.67}$$

That is, the average power in the process is the sum of the powers of all the frequency components in the spectrum. We can also define a cross power spectral density in terms of the cross-covariance function. For this we have the Fourier transform pair:

$$C_{xy}(f) = \int_{-\infty}^{\infty} c_{xy}(\tau) \exp(-j2\pi f\tau)\, d\tau \qquad c_{xy}(f) = \int_{-\infty}^{\infty} C_{xy}(f) \exp(j2\pi f\tau)\, df \tag{1.68}$$

This is important in that it expresses a general relationship between processes and not just particular specimens of the processes. Unfortunately, the interpretation of the shape of a ccvf can be difficult and there is no simple interpretation to be attached to the relation between $c_{xy}(0)$ and the integral of the cross power spectral density.

1.15 The Relationship between Ensemble and Time Statistics

In biological work directed toward the study of dynamic processes, one is often restricted to studying relatively few ensemble members and for only limited amounts of time. A question then arises as to whether one may legitimately infer the statistical properties of the ensemble from the behavior of just a few, or even one, of its members. This can be done if the stationary process satisfies what is called the ergodic hypothesis. According to it the behavior of one member of an ergodic process, if observed long enough, will be characteristic of all other members. Another way of stating this is that a stationary process is ergodic if time averaging of a single specimen function is equivalent to averaging over the entire ensemble. The frequency of occurrence of events in a single realization or specimen function converges to the ensemble distribution. Furthermore, if a process is ergodic, there is zero probability of getting "stuck" on a particular realization which does not have the long run property. Let us consider the relationship between time and ensemble statistics.

For a continuous random process the time average of a (real) specimen function $x(t)$ is given by

$$< x(t) >= \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)\, dt \tag{1.69}$$

and its autocovariance function by

$$< x(t)x(t+\tau) >= \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)\, dt \tag{1.70}$$

As indicated, the duration of the averaging interval T is made arbitrarily long. The brackets on the left hand side of the equations indicate temporal averaging. Ergodicity then implies that

$$< x(t) >= E[x(t)] \tag{1.71}$$

$$< x(t)x(t+\tau) >= c_{xx}(\tau) \tag{1.72}$$

Equations similar can be written for discrete random processes:

$$< x(t^{\circ}\Delta >= \lim_{N\to\infty} \frac{1}{N} \sum_{t^{\circ}}^{N-1} x(t^{\circ}\Delta) \tag{1.73}$$

and

$$< x(t^o\Delta)x[(t^o+\tau^o)\Delta] >= c_{xx}(\tau^o\Delta) = \lim_{N\to\infty} \frac{1}{N} \sum_{t^o=0}^{N-1} x(t^o\Delta)x[(t^o+\tau^o)\Delta] \tag{1.76}$$

(1.74)

The ergodicity relations (1.73) and (1.74) hold here as well. Eq. (1.76) differs from Eq. (1.17) only in that it indicates a limiting process with N becoming indefinitely large. When N is finite and $x(t)$ is a specimen function of an aperiodic process, Eq. (1.76) is an estimate of the autocovariance function of the process. The goodness of this estimate improves as N increases. Estimation problems are considered more fully in Chapter 3.

Another property of ergodic processes is that the joint probability distributions estimated from a single ensemble member approach those of the ensemble as the length of time the specimen is observed increases. Thus a long and detailed enough examination of one member of the ensemble can reveal all the statistical properties of the process, not just the second order ones.

The problem in biology, and neurobiology in particular, is that it is difficult to define at the outset of an investigation whether the process being studied is ergodic. It is often taken for granted, but such a definition can require an extensive examination of many specimen functions and even then there may be no clear-cut indication of the process's ergodicity. It is not possible to demonstrate conclusively that a biological process is ergodic since to do so one needs to observe all members of the ensemble for all time. At best we can only examine the available specimen functions and infer from them that the process is ergodic. Stationarity and ergodicity are often appealed to, sometimes implicitly, as justifications for studying but a few specimen functions of a process. They are assumptions that need to be considered carefully. For example, it is clear that the EEG process observed at any time by an electrode located anywhere on the scalp of a normal human is not ergodic because there are numerous differences between the EEGs recorded from anterior and posterior sites. The relative prevalence of the alpha rhythm is one particular instance of these differences. On the other hand, if only a single recording location is employed, then a particular specimen function of the EEG may quite possibly arise from an ergodic process describing activity over an ensemble of individuals. This can be more precisely defined by restricting the conditions of observation to normal alert adults. Similar illustrations can also be made for the spontaneous

spike activity of individual neurons. When properly defined, their activity can be described as arising from ergodic processes.

1.16 Mixtures of Signal and Noise

We have now come to the point where we must deal with mixtures of signals and noise. To rephrase our original definition at the beginning of the chapter, a signal is that constituent of the data which we are interested in and noise is whatever else is present. It is inherent in the observation of electrophysiological phenomena that some noise be present. What we are concerned with is the means to extract the signal from the noise. And since one or both of these data constituents is random in nature, it follows that the two must be separated by statistical means. Here we give a brief introduction to this problem.

The simplest method for representing the structure of observed electrophysiological data is as an additive mixture of signal and noise. If the observed data is $x(t)$, then

$$x(t) = s(t) + n(t)(1.77) \tag{1.75}$$

where $n(t)$ is the noise. It is in part biological and partly instrumentational in origin. Often in what follows, the signal $s(t)$ will be synonymous with a response. In the latter usage we refer to the electrophysiological response elicited by a stimulus. The response process, as already noted, may be random in structure. We always distinguish the signal or response from the net observed data $x(t)$. The additive nature of signal and noise is an assumption which, while valid in many situations, is frequently open to question. The alternative assumption is that the signal and noise interact in some nonlinear fashion, especially where the biological component of $n(t)$ is concerned.

The ease with which signal and noise can be separated depends in large part upon how large the signal is with respect to the noise. Obviously, the stronger the signal is, the easier it is to detect and analyze. When signal and noise are comparable in strength, problems in the reliability of signal detection and analysis procedures arise. Generally then, the goodness of signal analysis depends upon the ratio of signal to noise strength, the signal-to-noise ratio (SNR). The higher this ratio is, the more reliable are the estimates of signal structure. Several measures of SNR are in use. We mention two of them. (a) RMS signal-to-RMS noise ratio. When the signal component of the data originates from a continuous ongoing process, its rms level is a useful characterization of its strength. In this situation we measure SNR in terms of the rms values of signal and noise. The rms is a time measure of the standard deviation of a process. It is the square root of average power. (b) Peak signal-to-rms noise ratio. When the signal has a pulse or spikelike

waveform of limited duration, the most distinguishing feature of the waveform is its peak amplitude. In this case the convenient measure of SNR is the peak value of the signal to the rms value of the noise.

1.17 Response Detection and Classification – Hypothesis Testing

The randomness of neurobiological signals coupled with the background noise they are immersed in causes signal analysis procedures to involve statistical decision making in uncertain situations. For example, we may desire to ascertain whether a particular stimulus is effective in evoking a response from the test subject. Does the observed data contain a response, however obscure, or does it contain only noise? A related problem is when we know that a stimulus evokes a response and are interested in determining how changes in the stimulus parameters affect the response. To what extent are the observed differences in the data due to stimulus changes and to what extent to the interfering noise? The first of these problems is the signal detection problem and the second the signal classification problem. The signal here is the response to the stimulus. They possess considerable similarity in their theoretical formulations and in their solutions. Detection involves only the determination that the data do or do not have signals in them. Classification involves a quantitative description of what is already accepted to be a response in the data. This description is given in terms of such signal defining parameters as, for example, the amplitude, frequency, and phase of a sine wave. These parameters are estimated from the data with their goodness being affected by the amount of noise present. From these estimates signal classification is performed. Different segments of the data are judged to contain the same or different responses. according to how similar or different the corresponding parameter estimates are. The classification can involve as many groupings as one has reason to suspect exist in the data, perhaps one for each type of stimulus employed.

Solutions to detection and classification problems involve the concepts of hypothesis testing. In detection there are two mutually exclusive hypotheses: $H1$, that a signal is present and $H0$ that it is absent. In classification there are as many mutually exclusive hypotheses as there are signal classes to distinguish among. In either detection or classification the data are processed according to an algorithm determined by the experimental design and by the properties of the signals and noise, insofar as they are known. The algorithm yields a number whose magnitude then determines which of the hypotheses to accept. Associated with the acceptance or rejection of a hypothesis is the fact that decision errors are inevitable. Minimization of these errors is in fact a critical ingredient that goes into the choice of the

data processing algorithm.

Two types of errors can occur in signal detection. If $H1$ is the hypothesis that a signal is present and $H0$ is the hypothesis that it is absent, there is the possibility that $H1$ will be mistakenly accepted when $H0$ is actually true, and another possibility that $H0$ will be accepted when $H1$ is actually true. The first error is often referred to as a false alarm (error of the first kind) and the second error as a false dismissal (error of the second kind). Similar types of errors occur in the classification problem. If there are three different signals to choose amongst, six different misclassification errors are possible.

Let us illustrate a simple signal detection problem in terms of a three component signal contained in noise. The components are its amplitudes at three consecutive sampling instants: $s(t), s(t+1), s(t+2)$. The signal is present in background noise, a combination of background biological activity and instrument noise The noise is Gaussian with mean 0 and standard deviation σ. The signal and noise combine additively to yield the response data, $x(t) = s(t) + n(t)$. Successive samples of the noise are uncorrelated with one another. On the basis of these properties of the response data it is desired to construct a test to examine the hypothesis that in any particular three-sample sequence, there is a signal of arbitrary structure present in the data. As before, $H0$ is the hypothesis that a signal is absent from the data samples and $H1$ is the hypothesis that it is present. $H0$ is known as a simple hypothesis because it is concerned with only one possible value for the response vector in response space, here 0. $H1$, in contrast, is referred to as a composite hypothesis because it is concerned with the signal parameters such as amplitude and latency that have any non-zero value. So long as at least one of these parameters is different from 0 a signal is present. It is possible to have $H1$, a simple hypothesis stating, for example, that all sample values of the signal are unity. Usually, however, it is composite hypotheses covering a range of signal parameters that are of greatest interest.

An example of this is the hypothesis that some stimulus-related response of arbitrary waveshape is present in the data vector that we are examining. If Gaussian noise alone is present and its samples at consecutive sample times are independent, the data vector will tend to be found in a spherical region surrounding the origin. We can, from the three-dimensional Gaussian distribution, compute the radius of the sphere that a noise vector falls within some given percent of the time, say 99%. This radius is X_0. Let us then set up the test that we will accept hypothesis $H0$ if the observed data vector is within that sphere. That is, we accept $H0$ if the observed data, samples of the additive combination of signal

$$x^2(0) + x^2(1) + x^2(2) = X_0(1.78) \tag{1.76}$$

and otherwise we accept $H1$, the hypothesis that an arbitrary signal is present. In

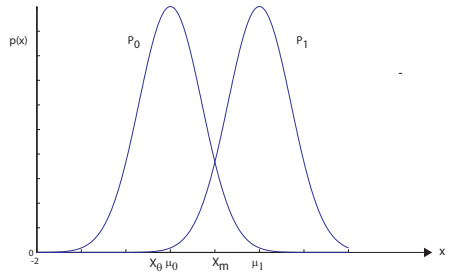

Figure 1.8: Fig. 1.8. Amplitude probability densities P_0 and P_1 for the data amplitude x, under hypothesis $H0$ and $H1$. X_θ is a threshold value for choosing $H0$ or $H1$ on the basis of experimental observation. X_m is the threshold for a maximum likelihood test.

choosing this and the threshold X_0 we have fixed at 1% the probability of false alarm. The probability of false alarm is called the level of the test. The probability of accepting $H1$ when it is true is called the power of the test. It is 1 minus the false dismissal probability and depends upon the strength of the response relative to the noise. Since the test is one involving the squares of the sample amplitudes, the power of the test will be low when the signal energy is small compared to the noise energy, and high when it is large, regardless of how it is apportioned among the three samples.

Fig. 1.8 illustrates these definitions. In it are shown two amplitude probability density functions of similar shape but differing means. They correspond to the two hypotheses being tested by a signal measurement of the data x. The left-hand density function is associated with hypothesis $H0$ and the other density function with the hypothesis $H1$. In this case the strength of the signal is the difference between the two means , $\mu_1 - \mu_0$. Let us somewhat arbitrarily select a decision threshold value along the abscissa such that if a measurement of x yields a value greater than X_θ, hypothesis $H1$ will be accepted; if not $H0$ will be accepted. Since the two probability densities overlap the threshold, some possibility of error is clearly to be expected. If noise only is present and a measurement exceeds X_θ an error of the first kind is made. The probability of this happening is the level of the test and is measured by the area under that part of the $H0$ curve to the right of X_θ. If $H1$ is true and the measurement exceeds X_theta, $H1$ is correctly accepted. The probability of this occurring is given by the area of the $H1$ curve to the right of X_θ, the power of the test. The area of the $H1$ curve to the left of X_θ measures the probability of making an error of the second kind, falsely rejecting $H1$.

While we have chosen an arbitrary threshold in this illustration, there is one

particular location for it that is usually selected, in a test such as this, that point where the two density curves intersect. Selection of the threshold at this point X_m results in a so-called maximum likelihood test. The value of the likelihood ratio (see below) at X_m is 1. Any measurement to the left of X_m is more likely to have resulted from a situation in which $H0$ was correct. But if it fell to the right of X_m, $H1$ is more likely to have been correct.

An optimum choice of boundaries in a multidimensional data space to use in accepting one of the hypotheses can be determined by the use of what is called the likelihood ratio. This is a ratio of two conditional probabilities. The one in the numerator, $P_1(X|H1)$, expresses the probability of having obtained the observed data if $H1$ were true (signal present); the one in the denominator is the probability $P_0(x|H0)$ of having obtained the observed data if only noise were present. Whenever the likelihood ratio equals or exceeds a preset threshold value, hypothesis $H1$ is accepted; otherwise H_0 is accepted. The two conditional probabilities are referred to as likelihood functions. The reason is that a conditional probability density is evaluated by inserting into it the observed data values. This yields an expression in which the parameters of the density, its mean and variance, say, are expressed as functions of the data. It is then possible to find values for the density parameters that maximize the conditional probability.

The values so obtained are the maximum likelihood estimates, the parameter values which are most likely to have produced the observed data. In the case of the likelihood ratio, the hypotheses being tested are associated with particular values for unknown parameters of the distribution. Then when the observed data values are inserted into each of these likelihood functions, the one which is the more likely will have the larger value. The value of the likelihood ratio can be computed for each observed set of data points only if the form of the conditional probability distributions governing the situation is known. In the particular three dimensional example of signal detection that we have chosen, one involving the detection of a signal in known Gaussian noise, the likelihood ratio is constant on the surface of a sphere centered on the origin, as is shown in Fig. 1.9. A simpler case is the situation in which we test the simple hypothesis that a three-sample signal is absent against another simple hypothesis that the signal has a constant value A. In this case the likelihood ratio is constant on a plane as shown in Fig. 1.9. The plane is oriented perpendicularly to the line joining the origin to the point (A, A, A) and its distance from the origin is determined by the choice of the level of the test.

When the decision involves choosing one of several possible signals in Gaussian noise, the data space is partitioned into planes, hyperplanes if there are more than three data samples per data vector, whose orientations and locations are determined by the statistics of the noise and the parameters of the different signals. The hyperplanes are the geometric embodiment of the likelihood ratio equations.

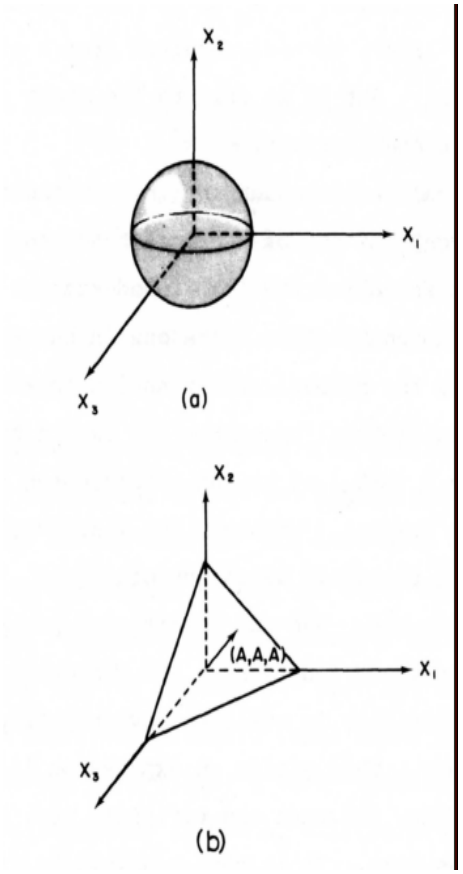

Figure 1.9: Fig. 1.9. (a) A sphere has constant likelihood ratio for testing for an arbitrary signal in noise. (b) A plane normal to the data vector (A, A, A) has constant likelihood ratio in the test for the presence of that particular signal vector in noise.

Signals which vary from one another in less simple ways can also be separated by data space partitions which are no longer hyperplanes but more complex surfaces. Nonetheless, the likelihood ratio concept still applies. The acceptance of one hypothesis in preference to the others is dictated by the location of the data vector in the data space.

Hypothesis testing as described here is performed by the establishment of decision rules with which to test the observed data. These rules are established by knowledge of the properties of the anticipated signals and the interfering noise. The more comprehensive is this knowledge, the more effective the tests can be. But as long as there is noise to contend with, the decisions can never be error-free. Once a decision rule has been adopted for a particular experiment, the error probabilities are determined by the properties of the response and noise processes. It is important to understand that the choice of the decision rule may be crucial to the success or failure of an experiment, for there are good decision rules and bad. A bad one will obviously have associated with it high probabilities of errors. But there are also other aspects to the choice of decision rules to be concerned with. The first is that there is generally an optimum decision rule for an experiment, one which minimizes the error probabilities. No other means of processing the data can improve upon this decision rule. In some cases, that optimum decision rule is known or can be calculated and then relatively simply instrumented; in others, it can be calculated and then instrumented only at great cost. When the latter is true, it often leads to the search for suboptimm decision rules, rules which are almost as good theoretically but have the advantage of being practical to employ. Biological data processing problems are commonly solved by the application of ad hoc suboptimum techniques. Great care is advised in considering the use of such techniques for there are often no satisfactory methods for dealing with them analytically. To prove their value in comparison with other techniques it may often be necessary to test them with computer simulated data and trial analyses on pilot data. It often turns out that what seemed on first inspection to be an effective analysis procedure is no better than the method it is meant to replace and sometimes worse. The appealing simplicity of the suboptimum technique must be accompanied by verified adequate performance if it is to be accepted as useful for data processing.

REFERENCES

- Cramer, H., "Mathematical Methods of Statistics," Princeton University Press, Princeton, 1946.
- Davenport, W. B., Jr. and Root, W. L., "An Introduction to the Theory of Random Signals and Noise," McGraw-Hill, New York, 1958.
- Hamming, R. W., "Numerical Methods for Scientists and Engineers, 'I 2nd

ed., McGraw-Hill, New York, 1973.

- Jenkins, G. M. and Watts, D. G., "Spectral Analysis and its Applications," Holden-Day, San Francisco, 1968.
- Mood, A. M., "Introduction to the Theory of Statistics," McGraw-Hill, New York, 1950.

Chapter 2

BASICS OF SIGNAL PROCESSING

2.1 Introduction

The data arising from an electrophysiological experiment on the nervous system initially consist of records in continuous analog form of stimulus events and the responses that they give rise to. If these data are to be analyzed in more than a qualitative way, digital computation techniques are usually called for. This means that the analog data have first to be converted to digital, sampled form. Then the full range of analysis techniques that have been developed to study dynamic processes can be brought to bear. These include filtering, averaging, spectral analysis, and covariance analysis. In this chapter we discuss first the properties of the analog-to-digital conversion processes with particular regard to their effect on the experimental data, and the subsequent tests the data are subjected to. Then we move to a discussion of filtering operations, analog and digital, with emphasis on the latter and how it fits into computer data analysis procedures. From time to time we consider some of the hardware aspects of filtering since familiarity with them is quite useful for a fuller comprehension of filtering procedures.

2.2 Analog-to-digital conversion

An analog-to-digital converter (ADC) converts a continuous signal into a sequence of T- and A-discrete measurements. The two steps of time sampling and amplitude quantizing are usually performed in a combined procedure. The ADC is first given the command to sample by the computer and then holds the amplitude of this sample briefly while quantizing it. We illustrate the ADC in Fig. 2.1 as performing

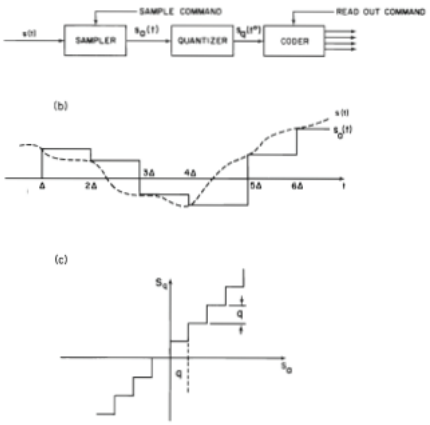

Figure 2.1: (a) Block diagram structure of an A-D converter. Sampling is initiated periodically. Quantization of the sample is followed by coding it into digital format. When this is complete a read-out command causes delivery of the converted signal to the data processor. (b) The signal $s(t)$ before sampling and its sampled version $s_a(t)$. (c) The input-output relation for the quantizer. The step size is q.

its operations in the sequence indicated there. The organization of the converter is not intended to describe a particular type of ADC, but to illustrate the function of such a device. In addition, the data analysis problems we are concerned with do not depend upon the detailed circuitry linking the computer to the ADC or upon the structural features of the converter itself. The sampled version of the signal is $x_a(t)$, a sequence of maintained voltage levels lasting the duration between sampling times, Fig. 2.1(b). The amplitude of each level is the signal amplitude at the sampling instant $t^o\Delta$. In what follows, we assume Δ to be unity so that $t^o\Delta$ can be replaced by the integer valued time variable t^o. Sampling devices are often referred to as sample-and-hold circuits because of their ability to hold the sampled value without significant decay until quantization has been completed–a time duration that is often considerably shorter than the interval between samples.

In a number of experimental situations in which a response to a stimulus is being analyzed, the instrumentation is organized so that the stimulator is triggered by the same pulse that initiates A-D conversion of the data. This insures that there will be no jitter (random variation in time) or asynchrony between the onset of the stimulus and the data sampling instants. That is, sampling always occurs at fixed delays from stimulus onset. If, on the other hand, the stimulator is driven independently of the ADC and notifies that device when to initiate sampling, jitter

of the sampling instants can occur and tend to result in some temporal smearing of the digitized data. The jitter effect will be small when the cycle time of the computer is small compared with the sampling interval. Here we ignore the effects of jitter in A-D conversion.

The sampled signal $x_a(t)$ is then quantized to yield an output $x_q(t^o)$ which can take on only a limited number of, usually, uniformly spaced values. The input-output relationship for the quantizer is shown in Fig. l(c). The quantization step is q volts in amplitude. The output is 0 as long as the input is greater than 0 and no larger than q; it is q as long as the input is greater than q and no larger than $2q$ and so on. In equation form, the input-output relationship is, at integral values of $t = t^o$ (with $\Delta = 1$)

$$x_q(t^o) = \begin{cases} Mq, & for x_a(t^o) \geq Mq = Q \\ mq & mq < x_a(t^o) \leq (m+1)q, \quad |m| \leq M \\ -Mq, & x_a(t^o) < -Mq = -Q \end{cases} \tag{2.1}$$

The maximum and minimum voltage levels that can be handled without saturation are Q and $-Q$ and the total number of levels $2M$ that the output signal can take on is usually some integer power L of 2:

$$2M = 2^L \tag{2.2}$$

The degree of precision of an A-D conversion is referred to in terms of the number of bits in the output word of the converter. A lO-bit converter will quantize voltages between -1 and +1 Volt into one of 1024 levels each of whose magnitude is 1.952 mV.

The final step in the conversion is to code $x_q(t^o)$ (only the values of x_q at the sampling times are important) into a form acceptable for use by the computer. Most often this means that $x_q(t^o)$, whether positive or negative, is represented in binary q form, L binary digits being adequate for this. Typically, one coded output line is assigned to each binary digit and the value of the voltage on this line at the read-out time indicates whether that binary digit is a 1 or a O. The time for both sampling and readout are determined by a clock contained within the computer. ”Interrupt” features of the computer assure that the incoming data are accepted after each quantization has been performed.

2.3 Quantization Noise

Each conversion has associated with it a discrepancy between the quantized and the true value of the signal. It is useful to consider this error as a form of noise,

called quantizing noise, $z_q(t^0\Delta)$. We can then write

$$x_q(t^o\Delta) = x(t^o\Delta) + z_q(t^o\Delta) \tag{2.3}$$

(2.3) z_q is limited in absolute value to 1/2 the size of the quantizing step q. (The properties of quantizing noise in the uppermost and lowermost quantizing levels are different but do not substantially alter this analysis.) We assume the incoming signal to be a random one that is band limited to $F = 1/2$ such that $\Delta = 1$. This means that sampling is done at the Nyquist rate. We also assume that the signal's amplitude is large compared to the size of a quantizing step but small enough not to produce peak value limiting at any time in the converter. Under these reasonable assumptions the following statements hold reasonably well: (1) the quantizing error of a sample is uncorrelated with that of its sequential neighbors; (2) the probability density function for the error z_q of a sample is uniformly distributed over the interval 0 to q. That is, it is equally likely that the magnitude of the error be anywhere in this range. From assumption (2) and the quantization rule of Eq. (2.1), the mean value of the quantizing noise is $q/2$. This is a bias term.

$$var[z_q] = \int_{-q/2}^{q/2} z_q^2\, dz = \frac{q^2}{12} \tag{2.4}$$

The lack of correlation between sample errors implies that the autocovariance function for the noise is given by

$$c_{z_q z_q}(t^o) = \begin{cases} \frac{q^2}{12}, \text{ for } \tau^o = 0 \\ 0 \quad \text{otherwise} \end{cases} \tag{2.5}$$

The power spectrum of the noise, excluding the dc bias term, is flat to $F = 1/2$. To see this, suppose the data consist of N samples of the signal and that we assume the combination of signal and noise to be periodic with period $T = N\Delta = N$. The substitution of Eq. (2.5) into Eq. (1.23) results in spectral terms $C_{z_q z_q}(n)$, which are all equal and independent of n. This is because $c_{z_q z_q}(\tau^o)$ is different from 0 only when $\tau^o = O$. Thus the quantizing noise is equally divided among all the $N/2$ frequency components between 0 and $N/2$:

$$C_{z_q z_q} = \frac{q^2}{12N}, \qquad 0 \le n < \frac{N}{2} \tag{2.6}$$

The ADC converter thus adds noise of its own to the incoming signal, a noise whose covariance and spectral properties are determined solely by the sampling rate and the fineness of quantization. Although quantizing noise has the appearance of being random, it is best to remember that this is not entirely so. To illustrate this

point, suppose the incoming signal were a repetitive wave synchronized exactly to some multiple of the sampling period. Samples taken of the waveform during each period at the same time relative to the beginning of a period will always produce the same quantizing error and this would not be removable by the process of averaging over successive waveform repetitions. However, as soon as some background noise is added to the fixed waveform, the situation changes. The quantizing noise then takes on many of the characteristics of random noise. In a sense, the uncorrelated quantizing noise is induced into the quantized signal whenever the incoming signal has a fluctuating random component. Thus, if the input noise bandwidth were very low relative to 1/2, the quantizing noise would still exhibit the flat power spectrum indicated by Eq. (2.6). This induced noise can only be removed by numerical or digital filtering of the digital data subsequent to the A-D conversion operation, a topic covered later in this chapter. Since there are many situations in which one is interested in signal peaks which may be small compared to the largest one present, the existence of quantizing noise must not be ignored, for it tends to make the small peaks less detectable. It can, for example, become an important factor when the biological noise contains a significant amount of lowfrequency components giving rise to what is referred to as baseline drift in the received data. When this occurs, it is common practice to reduce the amplification of the signal so as to prevent too frequent saturation of the signal amplifiers or peak limiting in the ADC. It is then quite possible that lesser peaks in the signal will be no larger than a few quantizing intervals, making the quantizing noise a factor of importance.

The fineness of A-D quantization is of importance in still another way. It affects the ability to reconstruct from the quantized output data, the amplitude probability distribution of the input data. This issue is somewhat different from that of detecting by response averaging a weak but constant response in a background of noise (Chapter 4). There, one is not interested in determining the nature of the amplitude distribution of the data. Here, detection of such subtleties in the data is the desideratum, with response detection being secondary. To find how well this can be done, it is necessary to know how fine, relative to the peaks in the amplitude distribution, the quantization steps must be. When a large number of quantized samples of the input signal are available, the answer, as Tou (1959) has shown, can be arrived at by considering the signal amplitude distribution as itself a waveform which is to be represented by a set of uniformly spaced samples along the amplitude axis. In this approach, the amplitude axis is analgous to the time axis of conventional waveform sampling. One can then apply the sampling theorem that states that for perfect reconstruction of a band limited wave whose highest frequency is F, sampling should be performed at a rate no lower than $2F/sec.$ In practice, when the experimenter examines the sampled version of the waveform on an oscilloscope, the Nyquist rate is usually inadequate to permit satisfactory visual reconstruction of

the waveform. Sampling rates for this purpose should be no lower than $3F/sec$ to $5F/sec$. Although probability distributions of amplitude are not truly band limited in terms of their Fourier transforms (called characteristic functions), it is possible to arrive at a convenient rule-of-thumb in determining what an adequate quantization step or sampling interval should be. Thus, suppose the narrowest peak in the amplitude probability distribution of the data is normal in shape, with variance σ^2. The Fourier transform of this distribution is also Gaussian and has more than 99% of its area confined to "frequencies" less than $1/3\sigma$. Considering this to be an adequate approximation to the "bandwidth" of the distribution, simple computations indicate the size of the quantizing step should then be very nearly σ. Note that though quantizing noise is present, its variance, $\sigma^2/12$, is small compared to the variance of the smallest peak in the input distribution. Our rule can now be stated in terms of the distance D between points three standard deviations away from the narrowest peak: a sampling width D/6 volts is adequate to represent peaks in the amplitude distribution which are D volts or more in width. The result holds for overlapping peaks as long as no component peak is narrower than D. If the peaks are sharper, the rule stated here will produce some distortion of their shapes which will be further contaminated by quantization noise. Sharp peaks therefore require some decrease in the quantization step.

2.4 Multiplexing: monitoring data sources simultaneously

Multiplexing is the process whereby several data sources have their information transmitted to the data processor over the same channel. Here the channel is the ADC and the multiplexing is performed by a process of switching the input of the ADC from one signal source to another. The rate at which the switching is performed and the choice of the source to be selected are determined by the data processor which accepts the data from the converter output. Both are constrained, of course, by the data handling capabilities built into the converter. When multiplexing is performed, an additional amount of time is required to perform a data conversion. The additional time arises because the process of switching the data converter from one source to another introduces a brief electrical transient into the signal and it is necessary to wait for this transient to subside before performing a conversion. The multiplexing time can increase the total conversion time by about 10%. Multiplexing of different data sources is performed most commonly at a uniform rate proceeding from source 1, to source 2, to source 3, etc., and back to source 1 in a recurrent, cyclic fashion. This is the mode of operation when the data from the different sources are signals of comparable bandwidths and whose temporal fluctuations are judged to be of equal interest and importance. When equal

sharing of the ADC by the different sources occurs, the minimum period between samples of anyone source is increased by a factor equal to the total number of multiplexed channels. As a consequence, the maximum bandwidth which each signal can have without introducing spectral aliasing is $1/2M\Delta$ where M is the number of equally multiplexed sources and $l/M\Delta$ is the effective sampling rate. In addition to being certain that the effective sampling rate is adequate to preserve signal structure, one must also consider the effects of noise in the input data and quantization noise. Ideally, prior to A-D conversion, filtering should be performed to remove from the input data all frequency components higher than $1/2\Delta$. If this is not done, the higher frequency noise components in the data will, after digitizing, be aliased with the lower frequency ones. Aliasing means that signal components at frequencies greater than 1/2 of the sampling rate will be misinterpreted as components at frequencies less than half the sampling rate. This falsifies the interpretation of signal structure made from the sampled data. Aliasing is discussed more thoroughly in Chapter 3. The net result is a decrease in the signalto- noise ratio of the digitized data. Suppose, for example, that the sampling rate of the ADC were 1000 samples/sec and that five data channels were being multiplexed. Suppose also that the prefilter had a high frequency cutoff at 500 Hz corresponding to the resolvable bandwidth if only one channel were being digitized. Now, five data sources are being multiplexed. The effective sampling rate of each source is 200/sec and the corresponding resolvable bandwidth is 100 Hz. Even if the response components of the input data have bandwidths less than 100 Hz, all the instrument noise between 100 Hz and the filter cutoff at 500 Hz will be aliased into the spectral region below 100 Hz. producing a degradation of the quantized data from the ADC. This degradation can be eliminated only by reducing the input data bandwidth to 100 Hz. For this reason it is highly desirable when background noise is an important factor to use a prefilter whose cutoff frequency is 1/2 the effective sampling rate.

The total quantizing noise remains unchanged during multiplexing since the quantizing error in each conversion is the same. However, the bandwidth of the digitized output has been reduced so that the spectral intensity of the quantizing noise is increased by the factor M. Filtering prior to A-D conversion cannot reduce this. As basic communications theory shows, this means that when sampling is done at the Nyquist rate, narrow bandwidth data are more affected by quantization noise than are broad bandwidth data.

In some situations, the monitored data sources have widely different bandwidths making it possible to sample the narrow bandwidth signals less frequently than the broad. This often results in a nonuniform rate of sampling of the broader bandwidth signals, there being occasional intervals in which they are not sampled. Usually no serious deterioration in the data analysis results. Infrequent interruptions in sampling can be further minimized by post A-D conversion digital filtering,

discussed later in this chapter, which has the effect of interpolating the missed data points in addition to smoothing the data.

If one considers only the spectral properties of the data sources and the sampling rate of the ADC, the problems associated with multiplexing are straightforward. However, another factor, the size of the computer storage area, needs also to be considered when real time data analysis is being performed. As discussed previously, in single channel A-D conversion all real-time data processors have a limited memory capacity in terms of the number of registers available to store data. When multiplexing is employed, these registers are parceled out to the different data sources so that over a given observation epoch, it is never possible to attain the same temporal resolution in each of the several multiplexed channels as it is with just one. The decision to resort to multiplexing must take this into account.

2.5 Data filtering

The operation of data filtering is one in which certain attributes of the data are selected for preservation in preference to others which are ”filtered out.” To design a satisfactory filtering device or program we must have some knowledge of the structure of both the signal and the noise. In the classical approach to filtering, the spectral components of the signal and noise are of major interest and filters are designed to select or ”pass” some spectral components, those containing primarily signal information, and reject or ”stop” others, those consisting mostly of noise. While any filtering operation can be described in terms of how it treats the different spectral components of the data, we shall see that this is not the only satisfactory way of dealing with filtration. Prior to the advent of computers, the filtering was concerned primarily with continuous electrical signals and was performed by networks consisting of passive elements (resistors, capacitors, inductors) and active devices (vacuum tubes, transistors). A network of this type is referred to as an analog filter. It operates on continuous data and yields a continuous filtered output. With the advent of the digital computer, it was recognized that analog filters performed computation on their input data which could be carried out equally well and sometimes better by computations on digitized sampled data without the need for constructing specific analog filter devices. In the following sections we discuss some basic attributes of filtering principally from the standpoint of the digital filter but, inasmuch as the continuous analog filter is still of great value in biological data processing, we also consider it to some extent.

2.6 The digital filter

The history of the digitized version of a signal $x(t)$ over T seconds is represented by N samples of it from $t^o = 0$ to $t^o = (N-1)\Delta$. They form the set $\{x(t^0\Delta)\}$. Once again we assume that the signal is band limited to $F = 1/2$ and that $\Delta = 1$. A filter, digital or otherwise, is a device with input $x(t)$ and output $r(t)$. If the device is digital, it stores a sequence of the past samples of the signal $x(t^0)$ in digitized form and operates upon them according to a filtering rule or algorithm to yield the output sequence, $r(t^o)$. If the rule does not vary with time, the filter is a time invariant one and if, in addition, the rule involves only the computation of weighted sums of the stored data samples, the filter is linear. The output of a linear, time invariant digital filter can be written as

$$r(t^o) = \sum_{t^o=0}^{N-1} h(\tau^o)x(t^o - \tau^o)(2.7) \tag{2.7}$$

In this chapter we are concerned mainly with such filters. Though linearity and time invariance are confining restrictions to put upon a filter's properties, the variety of filtering tasks that can be performed by linear filters is sufficiently rich to satisfy many of our data processing requirements. Linear filtration is easy to understand and perform in both its digital and analog forms. However, it does have deficiencies that limit its ability to deal with time varying processes, and some of them will be made apparent in the discussion.

As Eq. (2.7) indicates, the linear filter consists of a set of N fixed weighting terms $\{h(\tau^o)\}$. $h(\tau^o)$ multiplies the τ^oth most recent sample of the quantized signal and the products when summed yield the filter output. As each new sample of the signal is acquired, the filtered output has to be recomputed, since each of the past samples is now one sampling period older and must be multiplied by the weighting factor corresponding to its age.

An understanding of the nature of the operation of a digital filter can be obtained by a specific example. Let us consider that the sampled $x(t)$, in addition to arising from a band limited signal, is band limited and repetitive with integer valued period T. Its Fourier series representation, valid at the sample times, is then

$$x(t^o) = \sum_{n=-N/2}^{N/2-1} X_T(n) \exp \frac{j2\pi n t^o}{T} \tag{2.8}$$

We demonstrate how any one of the Fourier components can be filtered from the signal.

2.6.1 Filtering of the constant component

If the filter output is to be $X_T(0)$, it will have this value regardless of the value of time at which the output is examined. That is, for any integer value of t the output of a filter operating upion M consecutive signal samples is

$$r(t^o) = X_T(0) = \sum_{\tau^o}^{M-1} h(\tau^o)x(t^o - \tau^o) \tag{2.9}$$

Now, since $x(t)$ is periodic and band-limited, we need only to have M = N = T equally spaced samples represent the signal. Then Eq.XX becomes

$$r(t^o) = X_T(0) = \sum_{\tau^o}^{N-1} h(\tau^o)x(t^o - \tau^o) \tag{2.10}$$

What we seek is the set of values that the $h(\tau^o)$ should have to make this equation an identity. Replacing each sampled value $x(\tau^o)$ by its Fourier series representation, as given in Eq. (2.8), we have

$$r(t^o) = X_T(0) = \sum_{\tau^o}^{N-1} h(\tau^o) \sum_{n=-N/2}^{N/2-1} X_T(n) \exp \frac{j2\pi n(t^o - \tau^o)}{N} \tag{2.11}$$

Interchanging the order of summation gives

$$r(t^o) = \sum_{n=-N/2}^{N/2-1} X_T(n) \exp \frac{2\pi jnt^o}{N} \sum_{\tau^o}^{N-1} h(\tau^o) \exp \frac{-2\pi jn\tau^o}{N} \tag{2.12}$$

(2.12) If all the $h(\tau^o)$ are equal to a constant value h, the inner summation becomes

$$h(0) \sum_{\tau^o}^{N-1} \exp \frac{-j2\pi n\tau^o}{N} = \begin{cases} Nh(0), & n = kN \\ 0 \quad \text{otherwise} \end{cases} \tag{2.13}$$

k is 0 or any other integer. Then, substituting this result into Eq. (2.12) gives

$$r(t^o) = X_T(0)Nh(0) \tag{2.14}$$

If $h(0) = 1/N$, we obtain the desired identity $r(t^o) = X_T(0)$. This means that the digital filter which extracts $X_T(0)$, the average value of a periodic $x(t)$, operates on the N most recent signal samples of $x(t)$, adding them and dividing by N. This particular filter is called a digital integrator since it simply numerically integrates the previous signal sample values. This result holds regardless of the value of Δ.

2.6.2 Filtering the m^{th} frequency component

We now wish to filter from the same band limited, periodic signal one of its harmonic components, in particular the component having the frequency $m/T = m/N$. This component is expressed by the sum of two terms, either

$$A_T(m)\cos(2\pi mt/N) + B_T(m)\sin(2\pi mt/N) \tag{2.15}$$

or its identical counterpart

$$X_T(m)\exp(2\pi jmt/N) + X_T(-m)\exp(-2\pi jmt/N) \tag{2.16}$$

We desire to specify a digital filter whose output at each sample time is the same as the amplitude of the m^{th} component of the signal, Eq. (2.16), at that time. Again, because of periodicity only N samples of the signal are necessary:

$$r(t^o) = \sum_{\tau^o=0}^{N-1} h(\tau^o)x(t^o-\tau^o) = X_T(m)\exp(j2\pi mt/N)+X_T(m)\exp(-j2\pi mt/N) \tag{2.17}$$

The problem is to find if there is a set of values of $h(\tau^o)$ which makes this relation an identity. To do this, we proceed as we did previously when determining the average value filter. We obtain the equation (which is identical to Eq. (2.12))

$$r(t^o) = \sum_{n=-N/2}^{N/2-1} X_T(n)\exp\frac{j2\pi nt^o}{N}\sum_{\tau^o=0}^{N-1} h(\tau^o)\exp\frac{-j2\pi n\tau^o}{N} \tag{2.18}$$

Inspection of this equation reveals that for it to be identical to Eq. (2.16) we need to have

$$\sum_{\tau^o}^{N-1} h(\tau^o)\exp\frac{-j2\pi n\tau^o}{N} = \begin{cases} 1, & n = \pm m \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

(2.19)

Let us employ some intuition and guess the form of the solution for $h(\tau^o)$:

$$h(\tau^o) = \frac{2}{N}\cos\frac{2\pi m\tau^o}{N} = \frac{1}{N}\left[\exp\frac{j2\pi m\tau^o}{N} + \exp\frac{-j2\pi m\tau^o}{N}\right] \tag{2.20}$$

We substitute this into the left-hand side of Eq. (2.19) and find that

$$\frac{1}{N}\sum_{\tau^o=0}^{N-1}\left[\exp\frac{j2\pi\tau^o(m-n)}{N} + \exp\frac{-j2\pi\tau^o(m+n)}{N}\right] = \begin{cases} 1, & n = \pm(m+kn) \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

(2.21) just as we needed. Since the signal is band limited, only the terms $n = \pm m$ are of interest. Substitution of Eq. (2.20) into Eq. (2.18) gives

$$r(t^o) = X_T(m) \exp \frac{j2\pi m t^o}{N} + X_T(m) \exp \frac{-j2\pi m t^o}{N} \tag{2.22}$$

the desired result. The right-hand side may also be expressed in terms of Eq. (2.15). We have thus obtained the digital filter which operates upon the N most recent samples of the signal to yield at its output the sampled sequence representing the mth component of the signal. The values of $A_T(m)$ and $B_T(m)$ can be obtained from the output of the filter, Eq. (2.22), by measuring both the peak value of the output and the time it is 0 and by employing the well-known identity

$$\begin{aligned} A_T(m)\cos(2\pi mt/N) + B_T(m)\sin(2\pi mt/N) = \\ [A_T^2(m) + B_T^2(m)]^{l/2} \cos 2\pi mt/N + arctan[A_T(m)/B_T(m)] \end{aligned} \tag{2.23}$$

(2.23)

$X_T(m)$ and $x_T(-m)$ can be obtained if desired by using Eq. (1.12). Now that we have seen that it is possible to design a digital filter that extracts the mth component of a periodic signal without error, it is possible to demonstrate, though we do not do it here, that any combination of components of such a signal can be filtered, by a single compound filter that combines the properties of the individual component filters. Furthermore, this combined filter can weigh the contribution of the individual components to the output. Thus, suppose we wish to filter the q^{th} and m^{th} component of the signal $x(t)$ and weight them $V(q)$ and $V(m)$ respectively.

The filter output will then be

$$\begin{aligned} r(t^o) =& V(q)[X_T(q)\exp(-j2\pi qt^o/N) + X_T(-q)\exp(j2\pi qt^o/N)] \\ &+ V(m)[X_T(m)\exp(-j2\pi mt^o/N) + X_T(-m)\exp(j2\pi mt^o/N)] \end{aligned} \tag{2.24}$$

Reference to Eq. (2.20) shows that this response can be obtained from a filter whose response is defined by

$$\begin{aligned} h(\tau^o) =& \frac{V(q)}{N}[\exp(j2\pi q\tau^o/N) + \exp(-j2\pi q\tau^o/N)] \\ &+ V(m)[\exp(j2\pi mtau^o/N) + \exp(-j2\pi m\tau^o/N)] \end{aligned} \tag{2.25}$$

(2.25)

The output of this filter contains only the q^{th} and m^{th} components of $x(t)$ and in the desired strengths. The result can be generalized to a filter operating upon all the frequency components of the signal.

2.7 Impulse response of a digital filter

A convenient way to represent the response of a digital filter is by means of its unit sample response or impulse response $h(t^o)$, its response to a unit amplitude signal sample or discrete time impulse at $t^o = O$. All other signal samples, before and after, are O. As an example, consider the impulse response of a filter which weights equally each of the previous M samples of a signal. This response is

$$h(t^o) = \begin{cases} K, & 0 \leq t^o \leq M-1 \\ 0, & \text{elsewhere} \end{cases} \tag{2.26}$$

(2.26)

A unit amplitude sample which appeared at the filter input less than M samples ago yields a filter output whose present value is K. If a unit sample at the filter input has occurred more than M samples in the past, then the present value of the output is O. Later in the chapter we shall see that the discrete time impulse response described here is closely related to the continuous time impulse response of an analog filter.

When the impulse response of a filter is known, its response to any input signal can be calculated in a straightforward way by taking advantage of the linearity properties of the filter. Thus, to obtain the response at $t^o = 0$, $h(l)$ weighs the signal amplitude at $t^o = -1$, $h(2)$ weights the signal amplitude at $t^o = -2$, and so on with all the weighted signal amplitudes then being summed to obtain

$$r(0) = \sum_{\tau^o}^{\infty} h(\tau^o)x(-\tau^o) \tag{2.27}$$

If we are interested in the value of the response at time t, the same procedure follows, each term in the sum being the product of $x(t^o - \tau^0)$ and $h(\tau^0)$. Thus,

$$r(\tau^o) = \sum_{\tau^o=0}^{\infty} h(\tau^o)x(t^o - \tau^o) \tag{2.28}$$

This is the same as Eq. (2.9) if we consider only the first N terms. What we did in the preceding sections, therefore, was to design a digital filter by specifying its impulse response. The computational procedure described in Eq. (2.28) is referred to as convolution of the impulse response with the signal. It is often symbolized mathematically by the notation

$$r(t^0) = h(t^0) * x(t^0) \tag{2.29}$$

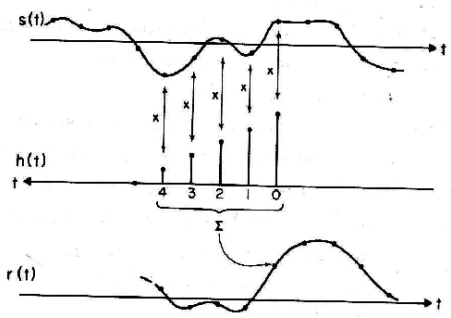

Figure 2.2: The convolution of a five sample filter with a sampled signal. The filter impulse response is shown in the middle trace with time reversed. The convolution computation is indicated at time t = 8. The continuous lines represent the band limited functions corresponding to the samples.

The convolution procedure is illustrated in Fig. 2.2 for a filter which has five terms in its impulse response. To make the procedure more visually comprehensible, we have reversed or folded over the time axis for the plot of $h(t^0)$. It tends to bring out the nature of the filter weighting procedure more clearly.

2.8 Spectral relations between filter input and output–the discrete fourier transform principles of neurobiological signal analysis

The relationship linking a filter's output to its input and its impulse response by the convolution process can also be expressed as a relationship between the corresponding Fourier coefficients. To see this, we again consider a periodic input signal that has period $T = N$ and is bandwidth limited to frequencies less than $1/2$. The signal at sample time $t^o - \tau^0$ can be represented, as before, by the Fourier series

$$x(t^o - \tau^o) = \sum_{n=-N/2}^{N/2-1} x_T(n) \exp \frac{j2\pi n(t^o - \tau^0)}{N} \tag{2.30}$$

This signal is passed through a digital filter with unit sample response $h(\tau^o)$. The output of the filter is given by Eq. (2.28) which has its own Fourier series expan-

sion, also with integer period N.

$$r(t^o) = \sum_{n=-N/2}^{N/2-1} R_T(n) \exp \frac{j2\pi n t^o}{N} \tag{2.31}$$

Note that the continuous counterpart $r(t)$ of the filter output $r(t^o)$ must also be band limited to the same frequency band since the filtering operation only modifies the amplitude and phase of the signal components that are present; it never introduces new frequencies. We need consider only filter responses which are less than or equal to the signal period since in the previous sections we have seen that no additional information is gained by making the memory longer than that. If the filter memory is less than the duration of the signal, this can be handled by setting to zero the values of $h(\tau^0)$ corresponding to $\tau^o \geq N$.

Let us now substitute Eq. (2.30) into Eq. (2.28). We obtain, after changing the upper limit in Eq. (2.28) to $N-1$,

$$r(t^o) = \sum_{\tau^o=0}^{N-1} h(\tau^o) \sum_{n=-N/2}^{N/2-1} x_T(n) \exp \frac{j2\pi n(t^o - \tau^0)}{N} \tag{2.32}$$

(2.32) We now interchange the order of the summations to obtain

$$r(t^o) = \sum_{n=-N/2}^{N/2-1} x_T(n) \exp \frac{j2\pi n(t^o)}{N} \sum_{\tau^o=0}^{N-1} h(\tau^o) \exp \frac{-j2\pi n(\tau^o)}{N} \tag{2.33}$$

(2.33) and then write the inner summation as

$$H_N(n) = \sum_{\tau^o=0}^{N-1} h(\tau^o) \exp \frac{-j2\pi n\tau^0)}{N} \tag{2.34}$$

(2.34) The right hand side of Eq. (2.34) is referred to as the discrete Fourier transform of the impulse response, $h(\tau^o)$. $H_N(n)$ and $h(\tau^o)$ are further related by the inverse discrete Fourier transform:

$$h(\tau^o) = \frac{1}{N} \sum_{n=-N/2}^{N/2-1} H_N(n) \exp \frac{j2\pi n(\tau^0)}{N} \tag{2.35}$$

(2.35) We call $H_N(n)$ the transfer function or system function of the filter. When Eq. (2.35) is substituted into Eq. (2.33) and the result compared to Eq. (2.31), the Fourier coefficients of the filter's output are seen to be given by

$$R_T(n) = H_N(n) X_T(n) \tag{2.36}$$

(2.36) Clearly, the output never has frequency components where the signal has none so that if the input is band limited, so is the output, and to the same bandwidth or less. The properties of the direct and inverse Fourier transform will be discussed more thoroughly in Chapter 3.

2.9 Filtering aperiodic signals

2.9.1 A. Short duration signals

Digital filtering has been discussed in terms of periodic signals that persist indefinitely. Under these circumstances a filter whose memory can store the waveform samples of a single period of a band limited signal will have an output which is also periodic and whose Fourier components are related, as shown in the previous sections, to those of the input signal by the weighting factors built into the filter. Often, however, we are confronted with signals that are transitory in nature, having a definite beginning and end. More often than not, in neurobiology, such signals occur in a background of noise and are of great interest. We would like therefore to build a filter which, essentially, passes th transitory signal but suppresses the noisy background. An example of such a situation arises when we wish to devote attention to alpha bursts in an ongoing EEG. The bursts occur infrequently and persist for relatively short periods of time while the EEG process is itself aperiodic. It persists indefinitely and may in this situation be considered noise. It also differs substantially from the alpha burst in its frequency content, and so a filtering operation on a sampled representation of the data can help to suppress the noise with respect to the alpha burst. Another example of a transitory potential is the action potential of an individual neuron. This is also often observed in a background of noise arising both from unrelated electrical activity of the nervous system and from the microelectrode and its associated amplifier. Here again, because the noise differs substantially in its frequency content from the signal, filtering can serve to enhance the size of the signal relative to the noise. In either example the goal is to preserve the structure of the signal and eliminate, as much as possible, the noise. However, a necessary consequence of filtering short duration signals is that there always occurs a certain amount of alteration in the structure of the signal, even when sampling is performed without distortion and at the Nyquist rate. The alteration is ascribable to the memory of the filter. Because of it, the filter's response to a brief signal has a longer onset than the signal and, likewise, a longer decay. The filter may be said to smear the signal out in time. This temporal smearing increases with the length of the filter's memory. Let us illustrate this with a digital filter designed to extract a brief burst of a $10Hz$ sine wave from background noise, an idealization of an alpha burst in the EEG. The filter memory consists of 20 sam-

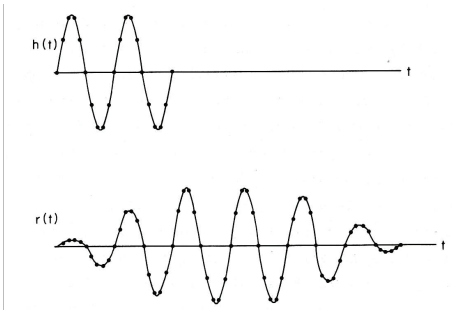

Figure 2.3: (a) The impulse response (dots) of a 10 Hz digital filter with a 20 sample memory two cycles in duration. (b) The response of this filter to a burst of four cycles of a 10 Hz sine wave. Note the response onset and decay. In both (a) and (b) a continuous line is drawn through the dots to aid visualization.

ples taken at a rate of $100/sec$. The filter's unit sample response illustrated in Fig. 2.3(a) is two cycles of a

sampled sine wave. If the filter were operating upon a periodic signal whose period corresponded to the memory duration of the filter, $0.2sec$, $10Hz$ would correspond to the second harmonic, $n = 2$, of the $0.2sec$ interval. The filter passes this component of a periodic signal and rejects all the other harmonically related ones. The filter's output when the actual s ignal is a burst of four cycle s of a 10 Hz sine wave is shown in Fig.2. 3(b). It can be seen that the filter output exhibits transient behavior over the first two cycles of the response and then is transient-free for the next two cycles. This is followed by a transient decay to zero output for the time lasting from the end of the burst until 0.2 sec later, the duration of the filter's memory. Note that the original four-cycle burst has been stretched by the filter into one of six cycles duration with slower onsets and offsets than were exhibited by the original signal.

2.9.2 B. maintained signals

Let us move from the transient, short duration signals to signals which last indefinitely. Periodic signals, as previously pointed out, are of this class and we showed how we could analyze them completely with a restricted number of samples. For a band limited signal of T seconds and bandwidth $1/2$, $N = T$ samples are necessary for this. Suppose we now consider the filter memory to be limited to these N samples and let the period of the signal gradually increase beyond T seconds to T' seconds while still maintaining its bandwidth limitation. It is clear, first of all, that

one result of lengthening the period is that the signal acquires an increased number of signal components equally spaced between 0 and $1/2$ Hz, $T'/2$ to be exact, and that these are not harmonically related to the original fundamental frequency $1/T$ (except when the extended period is a multiple of T). In the limit as T' becomes indefinitely large, the signal becomes aperiodic and has its frequency components spread throughout the continuum of frequencies between 0 and $1/2$. If we examine how the N sample filter operates upon an arbitrary frequency $f = 1/T'$ in this frequency region, we find that it passes frequencies other than those which are integral multiples of the fundamental frequency $1/T$. To see this, let us consider a filter designed to pass only the mth harmonic of a periodic signal with period $N = T$. We know from Eq. (2.20) that its impulse response is given by

$$h(\tau^o) = (2/N)\cos(2\pi m\tau^o/N) \tag{2.37}$$

(2.37) Let the input to this filter be the single frequency signal $x(t) = \cos 2\pi ft$. Then the output of the filter is, by Eq. (2.7)

$$r(t^o) = \frac{2}{N}\sum_{\tau^o=0}^{N-1}\cos\frac{2\pi m\tau^o}{N}\cos[2\pi f(t^o-\tau^o)] \tag{2.38}$$

(2.38) This equation is easiest to deal with when complex notation is used for the cosine terms. We can then simplify the right hand side in a straightforward manner that is, however, somewhat tedious. The simplification makes use of the identity arising from the summation of a geometric series,

$$\sum_{tau^o}^{N-1}\exp j2\pi g\tau = \exp[j\pi(N-1)g]\frac{\sin\pi Ng}{\sin\pi g} \tag{2.39}$$

(2.39) The result is that the output of the filter is found to be

$$\begin{aligned} r(t^o) = &\frac{\sin(\pi N(f-m/N))}{N\sin(\pi(f-m/N))}\cos 2\pi\left[ft+\frac{N-1}{2}f-\frac{m}{N}\right] \\ &\frac{\sin(\pi N(f+m/N))}{N\sin(\pi(f+m/N))}\cos 2\pi\left[ft+\frac{N-1}{2}f+\frac{m}{N}\right] \end{aligned} \tag{2.40}$$

(2.40) In this formidable looking equation it is the first term that is of major importance since it makes the principal contribution to the filter output in most circumstances. This can be seen by referring to our previously discussed 10 Hz filter. That filter has a 20-sample impulse response ($\Delta = .01sec$) that covers two cycles of a 10 Hz wave. Thus $m = 2$ and $N = 20$. To evaluate the filter's performance let us plot the expressions

$$\frac{\sin(\pi N(f-m/N))}{N\sin(\pi(f-m/N))} \quad \text{and} \quad \frac{\sin(\pi N(f+m/N))}{N\sin(\pi(f+m/N))} \tag{2.41}$$

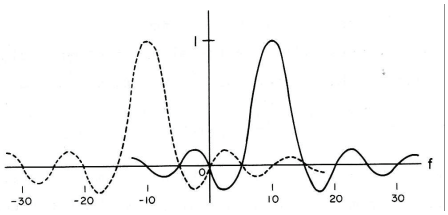

Figure 2.4: The two expressions of Eq. (2.41) plotted as a function of frequency. They show how the amplitudes of the two response components of Eq. (2.40) vary with signal frequency.

(2.41) as a function of frequency f for the chosen values of m and N. This is done in Fig. 2.4. (The phase shift terms in Eq. (2.40) are of minor importance.) There are several important properties to note:

1. 1. The filter has unity transmission at the single frequency $f = m/N = 1/10$ (This corresponds to a data frequency of $m/N\Delta$ = 10 Hz.)

2. 2. The filter has no transmission (infinite attenuation) at other frequency multiples of $l/N = 1/20$.

3. 3. There is significant transmission of other signal frequencies that are located in the frequency band between $(m - l)/N$ and $(m + l)/N$, here $1/20$ and $3/20$. (This corresponds to data frequencies of 5 and 15 Hz.)

4. 4. There exist other frequency bands on either side of this central band or main lobe where signal components can also be passed though with significant attenuation. These bands are often referred to as the side lobes of the filter. Their size is an important consideration in the design of filters that are used in spectral analysis.

The second term defining the filter response in Eq. (2.40) represents a contribution to the output effect that arises ' from the fact that a cosine wave is the sum of two complex frequency terms: exp j2rrft and exp(-j2rrft) that are equal in magnitude. The filter's cosine unit sample response has the same representation. Note that when the signal frequency is in the filter's main lobe, (m/N) - f is small while (m/N) + f tends to be large. This means, as already noted, that the contribution of the second term to the filter output is usually negligible except in filters that are designed to pass frequencies near O.

Suppose now that the memory of the filter of Fig. 3 were increased in duration by a factor of 5. Its unit sample response would then be 10 cycles of a 10 Hz

sine wave. We then have $M = 100$ and $m = 10$ (since we are interested in that harmonic of the fundamental period). If we examine expression (2.41) we find that it still has unit amplitude at $f = m/N$ but that now the zero transmission frequencies defining the main lobe are at $f = 9/100$ and $11/100$ (corresponding to data frequencies of 9 and 11 Hz). We have thus narrowed the pass band of the filter by a factor of 5. From this it can be surmised that there is an inverse relationship between the length of a filter's unit sample response (at a given sample rate) and its bandwidth. This general relationship between temporal and frequency properties always needs to be kept in mind when data filtering is employed.

2.10 Data smoothing by digital filtering

Let us now consider a slightly different filtering situation. Here the incoming data to the sampler is bandwidth limited to 1/2 Hz by a preamplifier and sampling is performed at a l/sec rate. We know, however, that the bandwidth of the response we are interested in is somewhat less than 1/2 Hz. Without adjustment of either the preamplifier or the sampling rate, we would like to design a digital filter that passes only the lower frequencies present in the response and rejects the higher frequencies as much as possible. We would like to do this using as little memory as possible and without introducing phase distortion which alters the response waveform. A filter which does this is called a smoothing filter and it has broadly useful properties. Its smoothing action results from its weighted averaging of a usually short sequence of consecutive signal samples.

A simple smoothing filter that computes the average of three consecutive samples of the signal has its response given by

$$r(t^o) = \sum_{\tau^o=0}^{2} h(\tau^0)\, x(t^o + 1 - \tau^o) \tag{2.42}$$

(2.42)

This way of representing the signal samples simplifies the analysis. $t^o + 1$ is the most recent sample time and $r(t^o)$ is the smoothed version of the signal one second ago . For simplicity let $x(t)$ have period $T >> 1$. We then substitute its complex Fourier series representation for it and interchange the order of summation to obtain

$$r(t^o) = \sum_{n=-N/2}^{N/2-1} X_T(n) \exp \frac{j2\pi n t^o}{N} \sum_{\tau^o=0}^{2} h(\tau^0) \exp \frac{-j2\pi n(\tau^o - 1)}{N} \tag{2.43}$$

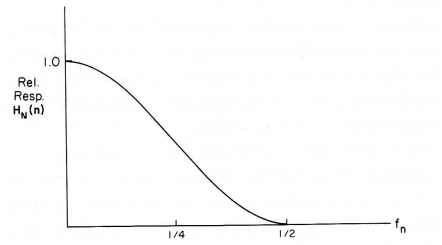

Figure 2.5: The frequency response $H_N(n)$ of the three component filter with weights $1/4$, $1/2$, $1/4$. The frequency axis is $f_n = n/N$.

(2.43) When the inner sum is expanded into its three terms, we have

$$h(0)\exp(-j2\pi n/N) + h(1) + h(2)\exp(j2\pi n/N) = H_N(n) \tag{2.44}$$

(2.44) See Eqs. (2.34) - (2.36) and the accompanying text. $H_N(n)$ is a spectral component weighting factor, possibly complex, for each term in the Fourier series representation for $r(t^0)$:

$$r(t^o) = \sum_{n=-N/2}^{N/2-1} X_T(n)H_N(n)\exp\frac{j2\pi nt}{N} \tag{2.45}$$

(2.45) Now let us choose the values of the $h(\tau^o)$ so that (a) there is unity gain at 0 frequency and (b), the frequency component in $x(t)$ at $f = 1/2$ is completely removed from $r(t^0)$. A simple filter which meets these constraints has $h(l) = 1/2, h(O) = h(2) = 1/4$, and, from Eq. (2.44)

$$H_N(n) = \frac{1}{2}[1 + \cos(2\pi n/N)] \tag{2.46}$$

(2.46)

This weighting filter $H_N(n)$ is plotted in Fig. 2.5 as a function of $f = n/N$. The nth frequency component in $r(t^o)$ is $H_N(n)$ times the nth component in $x(t)$. Note that the weighting factor is always real and positive so that none of the frequency components in r(tO) is different in phase from those of the original signal. Also, the low frequency components near 0 are passed almost without attenuation and the higher frequency components, those above $1/4$, are highly attenuated, meaning that if there is noise present and it is uniformly distributed in the spectrum, a large fraction of it has been removed without adversely affecting the low frequencies in the signal waveform. The noise reducing effects can be exemplified

by referring to the statistics of the original noise. If its bandwidth were $1/2$ and its variance unity, the variance of the noise at the filter output would be

$$var[r(t^o)] = (1/4)^2 + (1/2)^2 + (1/4)^2 = 3/8 \tag{2.47}$$

(2.47) This reduction in noise has been obtained without changing the bandwidth of the preamplifier preceding the digital filter. It is also important to remember that if the sampling rate had been reduced to be commensurate with the bandwidth of the response component of the data, there would have been no change in the variance of the noise per output sample. Instead, the noise at the output would have had its bandwidth compressed (by aliasing of the higher frequency components) to produce increased contamination in the spectral region, occupied by the signal. Thus, when signal bandwidth is considerably less than noise bandwidth, digital filtering can yield significant reductions in noise, including that introduced by quantization, that would not be obtainable by alterations in sampling rate.

Although the preceding discussion was based upon $x(t)$ being periodic, the results obtained also apply to stochastic band limited signals in general. The response properties of the filter are such as to inherently pass low frequencies and attenuate high frequencies regardless of the signal structure.

The digital filter described here is one example of a type encountered often in problems dealing with signal smoothing (Hamming, 1973). A certain amount of improvement in removing the higher frequencies can be obtained by increasing the filter memory to 5, 7, 9, etc., samples. Such filters can be designed to have a variety of transfer functions with different filtering characteristics, to have no phase shift, to pass the 0 frequency component without loss, and to attenuate the 1/2 Hz component completely. Such filters can be quite useful, but an adequate discussion of them is beyond the scope of this book. The interested reader is referred to Oppenheim and Schafer (1975). We note, however, that as the memory and complexity of the filter increases, the amount of time required to perform the computations also increases. When the data are being processed in real time, the amount of time available to compute each filtered data value can be no greater than the length of the sampling interval minus the time required to sample and perform A/D conversion. The computation time required for any given filter will depend upon the speed and structure of the computer employed. For these reasons, there can be no hard and fast formula relating sampling rate to allowable filter characteristics. In addition to being used for spectral filtering of a response from the background noise contained in the data, digital filters may at times be useful for interpolating purposes. This use arises in situations in which the data occasionally contain brief spikelike transients that are perhaps artifactual in origin and unrelated to the response of interest. In this case an interpolating filter can act to minimize the effect of the transients by

discarding the data sample amplitude at the time of the transient. An example of a three-sample interpolating filter is one whose weighting coefficients are 1/2, 0, 1/2. Postexperimental use of this filter upon those data points where the transients are suspected can be of value. But it is not a filter to be used with abandon. The reason is that its transfer function, evaluated according to the methods employed above, is

$$H_N(n) = \cos(2\pi n/N) \tag{2.48}$$

(2.48) Note that there is complete attenuation of frequencies near 1/4 Hz and that all frequency components above that are phase shifted by 180^o at the filter output. This can produce severe distortion in the filtered version of the data. A five-sample interpolating filter with weighting coefficients 1/4, 1/4, 0, 1/4, 1/4 can reduce this distortion somewhat but it also produces 180^0 phase shifts over half the frequency band and weights negatively the frequency components near $f = 1/2$.

Digital filters have been applied with good success to spectral analysis of random processes such as the EEG. In this type of application which is discussed in more detail in Chapter 3, the spectral characteristics of the filter are of equal importance to its temporal response properties. The filter is spoken of as providing a spectral window through which to view the random process. The spectral shape of this window is one of the factors that determine how well one can estimate the spectrum of the random process from a sequence of its sampled values. There are a number of window filters which have been proposed and used for this type of analysis. Their predominant use is off-line with recorded data. This means that the filters are not constrained to operate upon a small temporal segment of the data and can have substantially large memories in order to compute the estimates necessary to a spectral analysis.

2.11 Digital filters with feedback-recursive filters

The filters discussed above have used the N most recent samples of the signal. Their finite impulse response means that they possess no memory of data that occurred more than N samples ago. This situation can be modified without increasing the physical size of memory of the filter by the use of data feedback from the filter's output. A filter employing feedback is often called a recursive filter. It turns out to have an infinitely long response to a unit sample. An example is given in Fig. 2.6.

Here the filter has a single storage element or shift register, D_l. The output, $r(t^o)$, of the register is the value its input, $e(t^o)$, had at the previous sampling time. The input to the register is the weighted sum of the signal and the register output:

$$e(t^o) = x(t^o) + Kr(t^o) \tag{2.49}$$

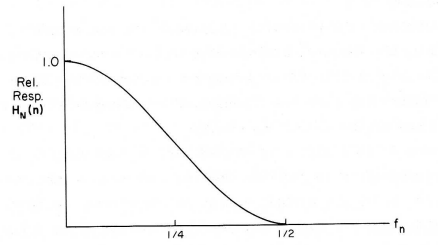

Figure 2.6: Fig. 2.6. A digital recursive filter with a single storage element D_l. The output of the storage element is equal to its input one time interval earlier.

(2.49) where K is the weighting factor applied to the output of the storage element. The output of the filter is just

$$r(t^o) = e(t^o - 1) \tag{2.50}$$

(2.50) If we substitute for $e(t^o - 1)$ in the above expression:

$$r(t^o) = x(t^o - 1) + Kr(t^o - 1) \tag{2.51}$$

(2.51) We then substitute $e(t^o - 2)$ for $r(t^o - 1)$ using Eq. (2.50) and substitute for $e(t^o - 2)$ using Eq. (2.49). We find that

$$r(t^0) = x(t^0 - 1) + Kx(t^0 - 2) + K^2 r(t^0 - 2) \tag{2.52}$$

(2.52) Continuing on in this manner for earlier and earlier values of r, we see that

$$r(t^o) = \sum_{\tau^o=0}^{\infty} \cdots \tag{2.53}$$

(2.53) The impulse response of the filter yielding this response is

$$h(\tau^o) = K^{\tau^o} \tag{2.54}$$

(2.54) The single shift register filter is thus theoretically equivalent to a filter storing all the past signal samples each of which it weights according to the power of K that corresponds to the age of the sample. Great flexibility can be obtained in feedback filters by using several shift registers in various feedback configurations. The theoretical response of such configurations can be obtained without undue difficulty , although we shall not do so here. For more details, see Oppenheim and Schafer (1975). In practice there are distinct limitations to the amount of memory

realizable with such a filter. Suppose, for example, we let $K = 1/2$. As the age of the sample increases, the value of the corresponding power of K decreases until it becomes so small that the weighted sample is not representable in a computer by a fixed point number. The smaller the size of the computer word, the smaller is the number of past samples that can be usefully represented. There are also problems encountered in the round-off errors of the products resulting from the multiplication by K called for in Eq. (2.49) and in the time requirements to perform them. These increase if floating point multiplication is employed to circumvent the limitations imposed by fixed point arithmetic. It can be seen that such problems require careful consideration when one is designing a feedback digital filter to meet a given response specification.

2.12 The linear analog filter

The continuous analog filter preceded the digital in its application to signal analysis and its use remains widespread, the growth in the use of the digital filtration notwithstanding. The reason is that solid state technology has made analog filters easy to design and apply, usually at a modest cost, to specific filtering problems. The concepts of the spectrum and impulse response of a filter and the relationship between them were first understood in terms of the linear, time-invariant analog filter. These were later adapted and extended to the linear, time-invariant digital filter.

The most common biological instrumentation application of the analog filter is in preamplifiers and amplifiers which link the biological preparation to the data analyzing system. As such, it produces the requisite signal amplification and some preliminary filtering, though at the unavoidable cost of adding instrument noise to the biological signal. Amplification is produced by the active power-producing elements. Filtering is produced by electrical circuits composed of passive resistors, capacitors, and, occasionally, inductors acting in conjunction with active amplifying elements. The configuration of these circuits and the relative sizes of the elements in them determine the characteristics of the filter. When the elements involved are linear, i.e., when their parameters are independent of the voltage across them or the current passing through them, the filters are referred to as linear. The relationship between the output and input of a linear filter is described by a linear ordinary differential equation with constant coefficients. An example of a linear filter is the low-pass filter illustrated in Fig. 2.7. The triangular symbol is commonly employed to indicate an active circuit element, usually an operational amplifier, which amplifies the incoming signal by a factor of K and inverts its phase. If we employ (a) Kirchhoff's rule that the sum of the currents entering a circuit junction

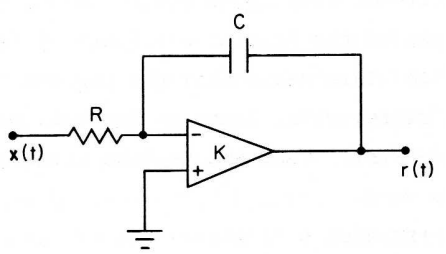

Figure 2.7: Fig. 2.7. A linear low-pass filter constructed from an operational amplifier, a resistor R, and a capacitor c.

is equal to the currents leaving, (b) Ohm's law, and (c) the current-voltage relation for a capacitor, $i = Cdv/dt$, and if we assume that there is no current entering the input terminals of the amplifier, it is a simple matter to show that the differential equation for this configuration is given by

$$[K/(K+1)]RC(d/dt)r(t) + (l/K)r(t) = x(t) \tag{2.55}$$

(2.55) The characteristics of this filter are obtained by solving the differential equation. As will be shown, this filter is a lowpass filter because it tends to pass the low frequency components of the data with little attenuation while, at the same time, it attenuates the high frequency components.

2.13 The laplace transform, the filter transfer function, and impulse response

In order to obtain the solution of the linear differential equation and, thereby, the explicit response of the filter to any arbitrary input signal, the most useful mathematical technique to employ is that of the Laplace transform. The Laplace transform when applied to a suitable function of time $f(t)$ yields a new function $F(s)$ defined by the transform equation

$$F(s) = \int_O^\infty f(t)\exp(-st)dt \tag{2.56}$$

(2.56)

There are many valuable properties of this transform that show up in the transformed function $F(s)$. Among them is the fact that the nth time derivative of $f(t)$ turns out to be $s^nF(s)$. [We ignore here consideration of the initial condition of

$f(t)$ and its derivatives at $t = 0$.] This means that the Laplace transform of a linear ordinary differential equation for $f(t)$ yields an algebraic equation in s and $F(s)$. Thus the Laplace transform of Eq. (2.55) is

$$[K/(K+l)]RC\ s\ R(s) + (l/K)R(s) = X(s) \tag{2.57}$$

(2.57) where $R(s)$ and $X(s)$ are Laplace transforms of $r(t)$ and $x(t)$ respectively. The solution for $R(s)$ in terms of $X(s)$ is

$$R(s) = \frac{X(s)}{[K/(K+l)]RCs + (l/K)} \tag{2.58}$$

(2.58) This equation can then be inverse transformed to yield the temporal response $r(t)$ as given by the inverse Laplace transform

$$R(s) = \frac{X(s)}{[K/(K+l)]RCs + (l/K)} r(t) = \int_{-\infty}^{\infty} R(S) \exp stds \tag{2.59}$$

(2.58)

Tables of the Laplace transform and its inverse have been compiled for the more common functions of t and s and they can often be used to advantage. See Abramowitz and Stegun (1965) for example. In practice, the application of Laplace transforms can be involved and direct solution of the differential equation may be preferable. One way to do this is by computer simulation methods, digital or analog. We shall not consider this further.

The Laplace variable s is a complex one with real and imaginary parts: $s = \sigma + j\omega = \sigma + j2\pi f$. Among other things this means that the integration in Eq. (2.59) is in the complex plane. It should also be noticed that if $\sigma = 0$, the Laplace transform resembles the Fourier transform closely. The resemblance is more than coincidental. The two are actually intimately related and this is important in deriving the properties of filters. A more complete discussion of the Laplace transform is beyond the scope of this book but can be found in many standard texts, Spiegel (1965) and Milsum (1966) for example. Simon (1972) provides an introduction to the Laplace transform at a more basic level.

If we know nothing about the properties of a filter except that it is linear, an examination of the relationship between its input signal and its output can reveal the filter's exact mathematical structure. In this regard, a particularly important input to the filter is the unit impulse, $\delta(t-T)$. This is an infinitesimally short pulse occurring at time τ whose amplitude is inversely proportional to its duration so that its area is 1:

$$\int_{-\infty}^{\infty} \delta(t-\tau)dt = 1 \tag{2.60}$$

(2.60) This impulse, or delta function, is also called the Dirac delta function. One of its important properties is that the time integral of its product with another time function $f(t)$ yields the value of that other function at time T:

$$\int_{-\infty}^{\infty} \delta(t-\tau) f(t) dt = f(\tau) \tag{2.61}$$

(2.60) The Laplace transform of $\delta(t-\tau)$ is easily shown by Eq. (2.56) to be $\exp(-s\tau)$. Note that when $\tau = 0$, this becomes unity.

If an impulse is applied at $t = 0$ to the input of a filter, the filter output at time t is, appropriately enough, its impulse response, $h(t)$. If the impulse is applied at time τ, the response of the filter at time t to this delayed impulse is $h(t-\tau)$. Suppose we are interested in the output of the filter at time t to some arbitrary input $x(t)$. We may determine this response by the following line of reasoning. Any signal can be considered to be composed of a steady stream of short pulses ΔT sec in duration, each of whose strength (area) at time $t-\tau$, τ sec earlier than t, is $x(t-\tau)\Delta\tau$. The response of the filter τ sec after such a pulse has been delivered is approximately $x(t-\tau)h(\tau)\Delta\tau$. Now, the linearity property of the filter assures us that at any time t the response of the filter to the entire past signal is the sum of its responses to the individual impulses of which that signal is composed. If we pass to the limiting situation by letting $\Delta\tau$, the pulse duration, become very small, we have

$$r(t) = \int_{0}^{\infty} x(t-\tau) h(\tau) d\tau \tag{2.62}$$

(2.62)

Note that the integration is over the past histor9 of the signal. The response of the filter $h(\tau)$ is thus 0 when τ is negative. This means that the filter cannot anticipate what its input will be in the future and this is a property of all real filters. It is not to be confused with the fact that, under suitable circumstances, a properly designed filter may predict future values of the signal on the basis of the signal's past behavior. That is another matter. Equation (2.62) shows that the output of the filter is the convolution of the input with the impulse response of the filter. This is the analog version of Eq. (2.28) obtained for the digital filter. Using the convolution notation of Eq. (2.29) we have

$$r(t) = h(t) * x(t) \tag{2.63}$$

(2.63) If we take the Laplace or Fourier transform of both sides of Eq. (2.62) we find that

$$R(f) = H(f)X(f) \tag{2.64}$$

(2.64) (It is not necessary here to distinguish between the two, beyond noting that the Laplace transform is evaluated for $s = j2\pi f$.) This is an analog counterpart of Eq. (2.36). An important facet of of these analog filter relationships is that there are no bandwidth restrictions on the incoming signal and as a result no aliasing problems to be concerned with. The relations are independent of both signal bandwidth and filter impulse response.

Let us now return to the filter described by the differential equation, Eq. (2.55). The solution of this equation indicates that the impulse response of the filter is given by

$$h(t) = \begin{cases} \exp(-t/RC), & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{2.65}$$

(2.65)

The quantity RC whose dimension in sec is the time constant of the filter. This filter is the analog of the digital feedback filter whose impulse response was given in Eq. (2.54). That this is so can be seen by considering $h(l)$ as obtained from Eq. (2.65).

$$h(l) = exp(-l/RC) = K \tag{2.66}$$

(2.66)

Then, for integer values t

$$h(t) = exp(-t/RC) = K^t \tag{2.67}$$

(2.67) which is the same as Eq. (2.54). Thus while the digital filter weights the past of the signal exponentially at the sample times t, the continuous analog filter does the same type of weighting for all values of time into the infinite past. In this sense the digital filter is a sampled version of the continuous one. At the sample times, it performs indistinguishably from the continuous filter provided the bandwidth of its input signal is properly limited.

The Fourier or Laplace transform (with $s = j2\pi f$) of Eq. (2.65) is the transfer function of the filter and is given by

$$H(f) = \frac{R\,C}{1 + j2\pi fRC} = |H(f)| \exp(j\theta) \tag{2.68}$$

(2.68) Note that if the incoming signal is a unit amplitude sine wave of frequency f, the output of the filter will be another sine wave of the same frequency:

$$r(t) = |H(f)| \sin(2\pi ft + \theta) = \sqrt{\frac{R\,C}{1 + (2\pi fRC)^2}} \sin(2\pi ft + \theta) \tag{2.69}$$

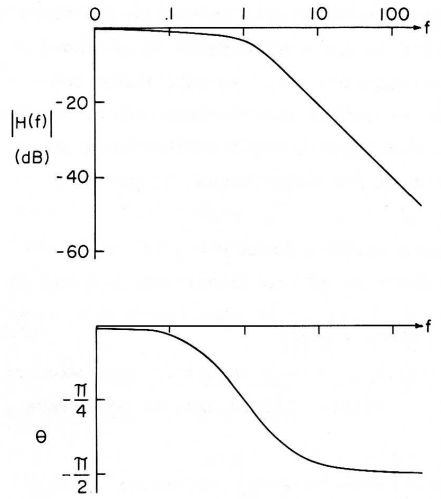

Figure 2.8: Fig. 2.8. Bode plots for the low pass filter of Eq. (2.69). RC = 1. The frequency axis is logarithmic. The upper diagram plots the gain in dB; the lower diagram, the phase shift in radians. The 3 dB cutoff frequency is at f = 1 Hz.

(2.69) The amplitude of the sine wave is the amplitude of $H(f)$ at the signal frequency f, and there is a shift θ in phase of the output relative to the input. In this case, the phase shift is given by

$$\theta = -\arctan(2\pi f RC) \tag{2.70}$$

(2.70)

The filter $H(f)$ has a pass band and a stop band. The pass band is defined as that band of frequencies in which a sine wave signal is attenuated by less than $sqrt2$. This amount of attenuation in decibels is $20log_{10}\sqrt{2} = -3dB$. The band of frequencies where attenuation is greater than 3 dB is defined as the stop band. The frequency marking the boundary between pass and stop bands is the cutoff frequency, here the 3 dB cutoff frequency. (Other attenuation levels are sometimes used to define the limits of a pass bana.) At the 3 dB cutoff frequency, the phase shift produced by the filter is 45 or $\pi/4$ radians. In the simple low-pass RC filter, the maximum phase shift occurs at very high frequencies and is -90. The characteristics of filter performance can be summarized in the pair of curves called Bode plots (see Fig. 2.8),

which relate its amplitude and phase properties to frequency. The upper curve plots $20log_{lO}|H(f)|$, the decibel value of $H(f)$, as a function of frequency while

the lower plots the phase angle θ. RC is taken to be $1/2\pi$. The decibel gain measure is preferred because when filters are cascaded, both their logarithmic gains and their phase shifts add. [This is true as long as the individual filter stages are properly isolated from one another (buffered) so that they do not interact.] As can be seen, the slope of the log gain curve in the region somewhat above $f = 1/2\pi$ is nearly linear. Thus, a simple low-pass, single time constant RC filter has a gain slope of -20 dB/decade of frequency. That is, each time the frequency increases tenfold, the gain is reduced one-tenth. Equivalently, each time the frequency doubles, the gain halves, a 6 dB change in gain per octave frequency change. This is characteristic of a filter described by a linear first order differential equation.

It is possible to increase the rate of attenuation of a filter in the stop band to -40 dB/decade by designing a filter represented by a second order differential equation, to -60 dB/ decade by a third order filter, etc. As mentioned above, one simple way of achieving the higher cutoff rates is to cascade filter units or stages. However, there are now far more elegant techniques for designing inexpensive filters that have sharp cutoff properties. The more prominent types of filters are of the Butterworth and Chebychev types. The principles behind their designs can be found in standard texts on filter design. See Brown et al. (1973) and the Federal Telephone and Radio Handbook (1963), for example.

Besides the low-pass filter, there are two other general types of filters which find wide application in studying dynamic processes. These are the high-pass and the bandpass filter. The former is characterized by being able to pass high frequencies with little or no attenuation while substantially attenuating low frequencies. It has a low frequency cutoff and a log gain-versusfrequency curve which is essentially a mirror image of the highpass filter. So is its phase-versus-frequency characteristic. High-pass filters are typically used to remove slow-wave activity from single unit records. They are also used to remove from the data spurious very low frequency components as might arise from electrode instabilities or from the amplifier components themselves. The bandpass filter, on the other hand, is characterized by being able to pass only a limited band of frequencies located between low and high cutoff frequencies. The attenuated frequencies are in the stop bands located beyond these cutoff frequencies. Filters of this type are employed when there is a more or less narrow range of frequencies of primary interest in the signal data as is the case, for example, in studying the alpha frequency component of the EEG.

The inverse of a bandpass filter is a stop band filter. One common application of it is to remove power line interference from recordings of EEG data. Although these filters have a very narrow stop band, their phase characteristics inherently introduce phase distortion of frequency components of the signal which may be relatively remote from the stop band. An inevitable consequence is that there is some

waveform distortion of the filbered signal. Thus, one should use such remedial filters with caution and only when other techniques for interference suppression at the source have failed.

2.14 The operational amplifier

Our brief discussion of the linear analog filter has presented only its essential properties in outline. Since the linear analog filter is so widespread in the prefiltering operations that precede A-D conversion, it is useful also to consider the analog filter from a more instrumentational point of view. Here we consider some of the properties of the active amplifying element that forms the heart of the analog filter, the operational amplifier. This device has simplified linear filter design in many instances to little more than cookbook complexity. Consequently an understanding of its basic properties will help the neurobiologist in applying these recipes to his own requirements. The operational amplifier's name derives from its original use in analog computers where it was developed to help perform the mathematical operations of summation, integration, and differentiation. These are accomplished by incorporating the amplifier into feedback networks which take advantage of the amplifier's most important property, extremely high gain or amplification. The applications of the operational amplifier have by now been greatly diversified so that it finds extensive use wherever analog filtering applications occur. At present, the most common configuration of the operational amplifier is the differential configuration shown in Fig. 2.7. The output of the amplifier is $-K$ times the voltage difference between the inverting (-) and the noninverting (+) inputs. Practical values for K range from 10,000 to considerably higher. Another basic property of the operational amplifier is that its input terminals draw negligible current from the electrical networks connected to them. In Fig. 2.9 the operational amplifier is shown in a four resistor network which makes it function as a differential amplifier with respect to the two signal sources e_l and e_2

Some simple network relations show how this comes about. First,

$$e_o = K(\Delta e) = -K(e_a - e_b)$$

The voltages e_a and e_b are derivable from the signals e_l and e_2 and the output e_O assuming that the resistances of these sources are very small compared to the network resistors. Thus,

$$e_a = e_0 \frac{R_1}{R_1 + R_F} + e_1 \frac{R_F}{R_1 + R_F} \tag{2.71}$$

(2.72)

$$e_b = e_2 \frac{R_G}{R_2 + R_G} \tag{2.72}$$

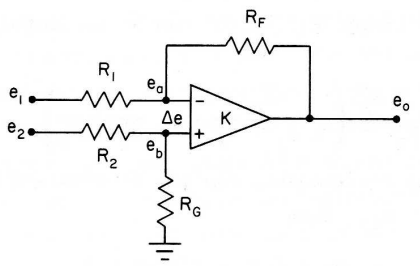

Figure 2.9: Fig. 2.9. An operational amplifier configured to function as a differential amplifier. The voltages at the inverting and non-inverting inputs are e_a and e_b, respectively. R_F is the feedback resistor.

(2.73) Substituting Eqs. (2.72) and (2.73) into Eq. (2.71) and simplifying somewhat, gives

$$e_0(1+\frac{K}{R_1+R_F}) = -K(e_1\frac{R_F}{R_1+R_F} - e_2\frac{R_G}{R_2+R_G}) \tag{2.73}$$

(2.74)

We now divide both sides by the coefficient of e_O and simplify it a little further

$$e_0 = \frac{-(R_1+R_F)}{[(1+K/K]R_1+(R_F/K)}(e_1\frac{R_F}{R_1+R_F} - e_2\frac{R_G}{R_2+R_G}) \tag{2.74}$$

(2.75) Since K is very large, Eq. (2.75) can be accurately approximated by

$$e_0 = \frac{-(R_1+R_F)}{R_1}(e_1\frac{R_F}{R_1+R_F} - e_2\frac{R_G}{R_2+R_G}) \tag{2.75}$$

(2.76) The differential relationship can now be obtained by setting $R_G = R_F$ and $R_2 = R_1$. Then

$$e_0 = \frac{-R_F}{R_1}(e_1 - e_2) \tag{2.76}$$

This is the defining relation for the differential amplifier. As long as K is large, the amplification is determined by the sizes of the resistors in the network and not by the gain of the operational amplifier. Changes in the sizes of R_2 and R_G from those selected here weight the contribution of e_2 differently from that of e_l. Note, however that the output is always in phase with e_2 and in phase opposition to e_l. Moderate variations in K, as inspection of Eq. (2.75) will indicate, do not materially alter the differential operation. This arises from the negative feedback from the output to the inverting input via R_F.

When a number of inputs are to be added in phase with one another and amplified, they may all be brought to the inverting input terminal of the amplifier of Fig. 2.9 through their own input resistors. The inverting input is thus also referred to as the summing junction. The contribution of each input to the amplifier output is in proportion to the Ohmic value of the resistor connecting it to the summing junction. Thus if e_l , e_3 and e_4 are connected to the summing junction by resistors R_l, R_3 and R_4, the amplifier output will be proportional to $R_l e_1 + R_3 e_3 + R_4 e_4$. It should also be noted that if the non-inverting input is connected directly to ground, the summing junction is at "virtual ground" potential in that its potential is the output voltage divided by the gain of the amplifier. Under normal circumstances the summing junction is never more than one or two millivolts from ground.

Another important use of the operational amplifier is as a buffer between signal and load . In this type of use, the amplifier is used to furnish more power to a load than the signal source can . The amplifier isolates the load from the signal and thereby prevents the load from distorting the signal waveform properties and from interfering with the properties of a biological preparation. Because of its low output impedance the buffer amplifier tends to suppress transient artifactual potentials which may be generated in the load. This can be the case where the load is an ADC. The switching transients which occur in these devices may, unless properly guarded against, corrupt the signal being digitized. A more complete discussion of the characteristics of operational amplifiers may be found in Brown et ai. (1973).

2.15 The amplitude comparator

It is common in the examination of single and multiple unit activities to have to assign an occurrence time or epoch to each individual waveform, be it a spike from a neuron or some particular feature of an EEG wave. The major difficulty in occurrence time measurement arises when the events are waveforms occurring amidst other activity such as background noise. Error-free estimation of the epoch is impossible; but the epochs of events whose waveforms are larger in amplitude than those of the background noise can be measured quite accurately with an amplitude comparator, a device which operates upon the instantaneous amplitude of the observed signal.

The output of the comparator changes rapidly from one voltage level to another when the amplitude of its input signal increases through some reference threshold level. The reverse transition in output occurs when the signal amplitude decreases through this threshold. Thus, both these time instants can be marked by the comparator. Many of the common amplitude comparators are based on the bistable

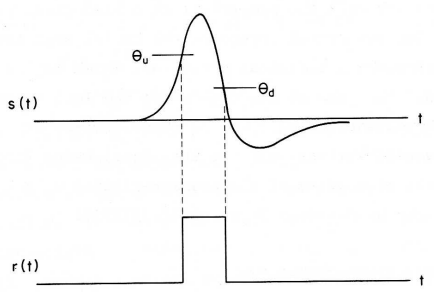

Figure 2.10: Fig. 2.10. Above, the input to a Schmitt trigger circuit. au is the upward going threshold; ed' the downward. The difference between the two, here exaggerated, is the hysteresis. Below, the output of the circuit. The pulse exists as long as the input exceeds au and has not gone below ed.

Schmitt trigger circuit and are further specialized so that they generate a brief pulse only at the time of the upward threshold crossing. Bistable means that the circuit has two stable states, one when the input is below threshold, th other when the input exceeds the threshold. The transition time is very rapid. In Fig. 2.10 is shown the action of such an amplitude comparator on a brief, pulse-like signal. The downward threshold crossing time is disregarded when only onset times are of interest. Also, comparators of the bistable Schmitt trigger type exhibit a hysteresis phenomenon: the threshold for a downward crossing is at a different lower level from the upward threshold. The hysteresis may be minimized by careful design and adjustment, but it can never be eliminated. Any measurement which requires accurate determination of upward and downward crossings of the same threshold must therefore employ amplitude comparators which effectively remove hysteresis.

The epoch time estimated by an amplitude comparator is subject to errors introduced by the presence of noise in the data and by inherent fluctuations in the observed spike waveform itself.

As the noise and waveform fluctuations increase, errors increase. The temporal distribution of the epoch estimate of a spike is dependent upon the properties of the noise and upon the spike shape and how it fluctuates from spike to spike.

Amplitude comparators can also be used to separate pulse-like waveforms of different amplitudes that are mixed together in the electrode signal. This often occurs in extracellular micro-electrode records. The activity of different neurons observed by the electrode will differ most obviously in the amplitudes of the action potentials from the different units. If the action potential peaks from each of these units occupy non-overlapping ranges in amplitude, the spikes from each unit can be

filtered from the others by the use of paired amplitude comparators, often referred to as an amplitude window circuit. One comparator is set at a low level A_1 and the other at a higher level A_2. The amplitude interval between them is the window and covers the range of amplitude variation exhibited by one unit. If the peak of a spike falls in the A_l to A_2 window, its waveform will ascend and descend through A_l without passing through A_2 in the intervening interval. Spikes which are larger than A_2 in amplitude will pass through both the A_l and A_2 levels, and those which are too low will pass through neither. The decision on whether a peak falls within the selected window must await the recrossing of the lower level. An output pulse is generated by the discriminator only if there has been no crossing of A_2 between the upward and downward crossings of A_1. The delay is generally not significant.

A pair of amplitude comparators can be used to filter out each spike amplitude range of interest. The window width of each pair is set experimentally to pass only those spikes thought to be associated with a particular single unit. Each pair operates independently of the others and their ranges must not overlap. In many experimental situations, however, overlap does exist in the amplitude ranges of the spikes generated by the different units. This is due to the intrinsic variability of the spike amplitudes themselves, as observed by the electrode, and to the presence of background noise. The overlap greatly reduces the utility of the filtering scheme since spikes arising from one neuron can be improperly attributed to another. A noise spike can also be mistakenly classified as arising from a unit and occasionally, the combination of noise and spike activity can cause a unit spike to be missed entirely. These standard misclassification errors can degrade the analysis of unit and inter-unit activity with a severity that depends upon the frequency of their occurrence. Except in exceptional circumstances, three units with non-overlapping amplitude ranges seem to be the practical limit that can be satisfactorily filtered by amplitude comparators. More powerful techniques are available for filtering spikes from one another on the basis of their waveform shapes. These techniques are discussed in Chapter 7.

2.16 Time-varying and nonlinear filters

With the exception of the amplitude comparator, the filters that have been discussed are linear. That is, the output is a weighted sum of the input signal, its time derivatives and integrals. The properties of a particular linear filter are determined by the assignment of weight to each of these terms. Once assigned, these weights are not changed during a filtering procedure. The filter is thus both linear and time invariant. These filters are usually referred to simply as linear filters in contrast to time-varying linear filters. We do the same here. The mathematics describing

the properties of the filter are those of linear difference equations, where computer operations on sampled signals are concerned, or upon linear differential equations, where operations on the original continuous signals are concerned. One of the key facts to keep in mind with respect to linear filters is that their best application is to situations in which the signal and the background noise are stationary. In certain instances of this type, mainly where the noise is Gaussian, it has been shown that a linear filter is the best filter that can be employed to extract response information. However, few of the neurological signals of interest can be said to fit completely the description of stationarity; nor is biological noise purely Gaussian. Linear filters perform data processing that is less than optimum in these situations. More satisfactory solutions to the problems encountered in dealing with non-stationarity and non-Gaussian processes require the application of a variety of filtering operations which may be either time-varying, nonlinear, or both.

A time-varying linear filter is one whose weights (or coefficients in the defining filter differential equation) may be systematically altered as a function of time according to some prescribed recipe. This is done in the case of adaptive or learning filters. These can be used to achieve better response estimates in the situation in which either the response or the noise properties vary with time. A nonlinear filter is one that performs operations which cannot be described by linear differential or difference equations. The amplitude comparator is an example of such a filter. Products, quotients, and powers of derivatives and integrals are among those that may be encountered in nonlinear filtration; so are logical operations. A moment's consideration will indicate that the class of nonlinear filters is vastly greater than that of linear filters. Nonlinear filtering operations are implicitly involved in a number of statistical data processing techniques that have been applied to response estimation. The consideration of these tests from the filtering point of view can have conceptual advantages for it helps provide a concise description of how signal data are processed.

REFERENCES

- Abramowitz, M. - and Stegun, 1. A., "Handbook of Mathematical Functions," Dover, New York, 1965.
- Brown, P. B., Maxfield, B. W. and Moraff, H., "Electronics for Neurobiologists, II MIT Press, Cambridge, 1973.
- Federal Telephone and Radio Corp., "Reference Data for Radio Engineers," American Book-Stratford Press, New York, 1963.
- Hamming, R. W., "Numerical Methods for Scientists and Engineers," 2nd ed., McGraw-Hill, New York, 1973.

- Milsum, J. H., "Biological Control Systems Analysis," McGraw-Hill, New York, 1966.
- Oppenheim, A. V. and Schafer, R. W., "Digital Siglial Processing," Prentice-Hall, Englewood Cliffs, 1975.
- Simon, w., "Mathematical Techniques for Physiology and Medicine," Academic Press, New York, 1972.
- Spiegel, M. R., "Laplace Transforms," Schaum, New York, 1965.
- Tou, J. T., "Digital and Sampled Data Control Systems," McGrawBill, New York, 1959.

Chapter 3

POWER SPECTRA AND COVARIANCE FUNCTIONS

3.1 Introduction

In the introductory chapter we pointed out the usefulness of covariance functions and spectral representations as ways of describing continuous data that are mixtures of signal and noise. These two ways of representing continuous dynamic processes lead to powerful methods of signal analysis. However, we indicated that the analysis procedures are generally performed not upon specimen functions of the original continuous processes, processes that are essentially infinite in duration, but upon finite segments of their sampled versions. The results and conclusions drawn from these analyses are then used to draw inferences about the original processes: the wave form of a response, its spectrum, its correlation with another response, its dependence upon a stimulus parameter, etc. The question is, how good are these inferences? Although we made some effort to point out the legitimacy of the procedures under many circumstances of practical interest, it is important that we establish their validity somewhat more securely. Once this is done we can examine specific applications of covariance f unctions and spectral analysis in more depth and detail. This will permit us in addition to move to related methods of signal analysis, such as coherence functions, which also have found applicability in studying the relationships between pairs of processes. Finally, these methods are applicable not only to the study of continuous processes such as the EEG but also form the bas is for the analysis of certain aspects of single and multiple unit activity. Thus an under standing of how continuous processes are analyzed forms a basis for studying unit activity.

3.2 Discrete Fourier Representations of Processes

At the outset it is important to state a basic attribute of band-limited signals that is of fundamental importance: Whether periodic or not, such signals must be infinite in duration. This fact follows directly from the properties of the Fourier transform for continuous signals. On the other hand, the properties of the Fourier transform also guarantee that the spectrum of a finite duration signal, such as a segment of an infinite duration signal, cannot be band-limited even when the infinite duration signal is. This means that there is an inherent contradiction built into our procedures for analyzing infinite duration signals from their finite segments. The contradiction is only resolved when the infinite duration signal is truly periodic. In all other cases we are forced to settle for errors of estimation. The sampling procedure does not alleviate these errors but introduces problems of its own, the kinds of problems we deal with here.

Although it may be a fiction, we have assumed that the processes we are studying are stationary mixtures of signals and noise with at least the noise being a random process. The analysis procedures are by necessity performed upon their finite duration segments. And here we invoke the next assumption, a true fiction. This is that the finite duration segment is a single period of a periodic specimen function. As objectionable as this might seem at first, it does no real harm since we have no knowledge of the specimen function's behavior outside this observed interval. Because of stationarity, the statistical behavior of the specimen function outside this observed interval is not likely to be much different. Thus we are not disregarding any information that we have concerning the specimen's behavior. In the first chapter we assumed that the repetition period was equal to the time of observation T. Other periodicity assumptions are also possible. If we want, we can consider the repetition period to be longer than T, T' say, by padding out the observed segment with a zero amplitude data segment lasting $T' - T$ sec. In a sense this is falsifying the data, but we know exactly how we have falsified it and we can take this into consideration in the subsequent analyses in order to avoid arriving at erroneous conclusions. Padding out the data with zero amplitude segments is a routine procedure when dealing with the estimation of the covariance functions. In this case, as we shall see, it is convenient to make $T' = 2T$. For the moment, however, let the repetition period be T. Let us keep in mind, then, the fact that we have forced periodicity upon the process and that for practical purposes we can make this periodicity length T or longer. Later on we shall use a $2T$ repetition period to deal with autocovariance function estimation.

In Chapter 1 we introduced the Fourier series representation for a $T-$ continuous periodic signal and showed that if the signal were band- limited, its waveform could be completely represented by a finite number of parameters. Specifically, if

the period of the signal is T and its bandwidth is F, then $N = 2FT$ terms are involved in either the real or complex Fourier series representation. It was also demonstrated that the signal could be completely represented by N consecutive sample amplitudes spaced Δ sec apart where $\Delta = 1/2F$ sec. The Fourier and time sample representations are closely related, the relationship between the two involving what is called the discrete Fourier transform. We introduce it here and show that in the bandwidthw limited situation, it leads to the same Fourier coefficients as would be obtained from a Fourier series representation of the original T-continuous signal.

We start with a single T sec segment of a band-limited signal which we consider to have period T. We obtain N samples of this signal at times δ sec apart starting at the beginning of the segment, $t = O$. Using these samples we can partially reconstruct the original signal by means of weighted sine functions. The partial representation is

$$x(t) \simeq \sum_{t^o=0}^{N-1} x(t^o\Delta)\frac{sin[\pi(t - t^o\Delta)/\Delta]}{\pi(t - t^o\Delta)/\Delta]} \tag{3.1}$$

The reason for the reconstruction being partial is that we have ignored the tails of the weighted sine functions outside the T sec segments. We can, however, insert them because of the assumed periodicity of $x(t)$. The complete reconstruction takes in all the weighted sine functions throughout time:

$$x(t) = \sum_{t^o=-\infty}^{\infty} x(t^o\Delta)\frac{sin[\pi(t - t^o\Delta)/\Delta]}{\pi(t - t^o\Delta)/\Delta]} \tag{3.2}$$

(3.2)

This now holds for all t. Now let us take the complex Fourier series representation of a single period from 0 to $T = N\Delta$:

$$\begin{cases} X_T(n) & = \frac{1}{T}\int_0^T \sum_{t^o=-\infty}^{\infty} x(t^o\Delta)\frac{sin[\pi(t-t^o\Delta)/\Delta]}{[\pi(t-t^o\Delta)/\Delta]}\exp(-2j\pi nt/T)dt \\ X_T(n) & = \frac{1}{T}\sum_{t^o=-\infty}^{\infty} x(t^o\Delta)\int_0^T \frac{sin[\pi(t-t^o\Delta)/\Delta]}{[\pi(t-t^o\Delta)/\Delta]}\exp(-2j\pi nt/T)dt \end{cases} \tag{3.3}$$

(3.3)

This is actually a simple equation to deal with, given the periodicity of $x(t)$. Because of periodicity we have $x[(N + t^o)\Delta = x(t^o\Delta)$. When this fact is taken into account for values of t^o outside the range 0 to $N-1$, Eq. (3.3) simplifies, after some elementary substitutions, to

$$X_T(n) = \frac{\Delta}{T}\sum_{t^o=0}^{N-1} x(t^o\Delta)\exp(-2j\pi nt^o\Delta)\int_{-\infty}^{\infty}\frac{\sin\pi x}{\pi x}\cos\frac{2\pi nx}{N}dx \tag{3.4}$$

(3.4)

As long as $n/N < 1$, which is true for the band- limited signal, this further reduces to

$$X_T(n) = \frac{1}{N} \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2j\pi n t^o/N) \tag{3.5}$$

(3.5)

The original integration operation upon $x(t)$ has thus been modified into a summation operation upon $x(t^o\Delta)$.

This is an opportune time to reconsider the steps that led us to Eq. (3.5). Our data specimen was a T sec segment of an ongoing process band-limited to real frequencies between 0 and $1/2\Delta$. We assumed, solely for the purpose of analysis, that this segment was one period of a periodic process. We then sampled the segment at interval Δ sec apart to obtain an N sample representation of it.

Because of the bandwidth limitation and the periodicity assumption, we need only N Fourier components at complex frequencies spaced equally from -1/2Δ to 1/2 Δ to represent the data completely. Now, in the majority of situations, the data do not arise from a periodic process but are specimens of an aperiodic process with power distributed at all frequencies up to 1/2Δ (or at all the complex frequencies between - 1/2Δ and 1/2Δ). Hence, our periodicity assumption has in a sense falsified the data. It has produced a representation of the signal requiring only N Fourier components. This is not a serious falsification, however. What it amounts to is saying that all the frequency components in the narrow frequency band between $(n - 1/2)/N\Delta$ and $(n + 1/2)/N\Delta$, a band $1/N\Delta$ wide, are considered to be concentrated at the single frequency $n/N\Delta$, and represented by $x_T(n)$. $x_T(n)$ is therefore essentially the product of the frequency density of the Fourier representation times the incremental bandwidth $1/N\Delta$. The density in that frequency region can then be obtained by dividing Xr(n) by I/N.6. This gives

$$N\Delta X_T = \Delta \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2j\pi n t^o/N) \tag{3.6}$$

(3.6) A is a constant independent of the duration of the specimen and plays only a minor role in the reconstruction of $x(t)$ from the Fourier representation. For this reason we define the discrete Fourier transform (DFT) of $x(t)$ as $x_N(n)$:

$$X_N(n) = NX_T(n) = \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2j\pi n t^o/N) \tag{3.7}$$

(3. 7) The elimination of the factor N that appeared in Eq. (3.5) means that in

order to recover $x(t^o\Delta)$ from $X_N(n)$, we must define the inverse DFT as

$$x(t^o\Delta) = \frac{1}{N}\sum_{n=-N/2}^{N/2-1} X_N(n)\exp(2j\pi nt^o/N) \tag{3.8}$$

(3.8) To see this, we multiply both sides of Eq. (3.7) by $\exp(2j\pi nu^o/N)$ and sum over all the values of n between $-N/2$ and $(N/2)-1$, the range of the complex Fourier expansion. We obtain

$$\sum_{n=-N/2}^{N/2-1} X_N(n)\exp(2\pi jnu^o/N) = \sum_{n=-N/2}^{N/2-1} X_N(n)\left[\sum_{t^o=1}^{N-1} x(t^o\Delta)\exp(-2\pi jnt^o/N)\right]\exp(2\pi jnu^o/N) \tag{3.9}$$

(3.9) We then interchange the order of the two summations on the righthand side and consider the summation with respect to n. This is

$$\sum_{n=-N/2}^{N/2-1}\exp[2\pi jn(u^o-t^o)/N]$$

For any value of t^o different from u^o, this summation is zero (as can be seen by using the summation formula for a geometric series). But when $t^o = u^o$, the summation is N. Thus, Eq. (3. 9) reduces to Eq. (3.8) which is what we wished to show.

Equations (3.7) and (3. 8) are a discrete Fourier transform pair and have been justified on a heuristic basis. Later in the chapter we shall establish the validity of the relation somewhat more carefully, paying closer attention to the properties of continuous processes. It is also worth noting that the definition of the DFT varies from author to author according to the handling of the factor N. The definition adopted here seems to be the most common one.

The cosine and sine versions of the DFT are given by

$$A_N(n) = 2\sum_{t^o=0}^{N-1} x(t^o\Delta)\cos(2\pi nt^o/N)$$

(3.9a)

$$B_N(n) = 2\sum_{t^o=0}^{N-1} x(t^o\Delta)\sin(2\pi nt^o/N)$$

(3.9b) These are associated with the complex relations $A_N(n) = X_N(n)+X_N(-n)$ and $B_N(n) = j[_XN(n) - X_N(-n)]$. It is worthwhile pointing again that $X_N(n)$,

the direct DFT, is a periodic function of n, period N, and its inverse $x(t^o\Delta)$ is a periodic function of time. That is, $X_N(-N+n) = X_N(N+n)$, etc. and $X[(-N+t^o)\Delta] = x[(N+t^o)\Delta]$, etc. In the previous chapters, we considered the index for the direct DFT to run from $-N/2$ to $(N/2)-1$. It is clear now that because of the periodicity it is equally satisfactory to consider n to range from 0 to $N-1$.

The periodicity of the direct and inverse DFT emphasizes the fact that when the DFT is applied to an N sample sequence of data points, it is done under the assumption that the data arise from a periodic process, period N. SOmetimes the period can be considered to be greater than N by appending or "padding" a sequence of zero amplitude samples, $N'-N$ of them so that the overall length of the resulting sequence is N'. This padding with zeros is a technique commonly employed in digital filtering and in the estimation of the acvf and spectrum of a specimen function, as we shall see later. The resulting sequence of sample values can be considered to arise from a periodic band-limited signal $\tilde{X}(t)$, period N', which is zero at $L = N'-N$ consecutive sample times. The DFT of this signal is

$$\tilde{X}_{N'}(m) = \sum_{t^o=0}^{N'-1} \tilde{x}(t) \exp(-2\pi j m t^o/N') \tag{3.10}$$

(3.10) Because of the fact that $\tilde{x}(t) = x(t)$ for values of t^o ranging from 0 to $N-1$ and is zero for values of t^o ranging from N to $N'-1$, we have

$$\tilde{X}_{N'}(m) = \sum_{t^o=0}^{N-1} x(t) \exp(-2\pi j m t^o/N') \tag{3.11}$$

(3. 11)

An especially important case is $N' = 2N$. Here we have

$$\tilde{X}_m = \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2j\pi m t^o/2N) \tag{3.12}$$

(3. 12) Because of the 2N periodicity of $\tilde{x}(t)$, the values of n range from $-N$ to $N-1$ instead of from $-N/2$ to $(N/2)-1$. If we examine Eqs. (3. 7) and (3.10), we see that when $m = 2n$, i.e. it is an even number or zero,

$$\tilde{X}_{2n} = \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2j\pi n t^o/N) = X_N(n) \tag{3.13}$$

(3. 13) This shows that the even index terms for $\tilde{x}_{2N}(n)$ are completely determined by the values of the $X_N(n)$. But what about the odd index terms? Some reflection

on this reveals that these terms arise solely because of the padding procedure. They are necessary to force $\tilde{x}(t)$ to be zero at the sample times between N and $N' - 1$. They provide no additional information about $x(t)$, but, interestingly enough, are an essential ingredient for obtaining an estimate of the acvf from the estimated spectrum. This point will be discussed later. Finally, it is easy to see that similar results would be obtained if N' were any other multiple value of N.

3.3 Aliasing

As we discussed in Chapter I, the necessity for sampling a signal at a rate compatible with its bandwidth, the Nyquist rate, is vital to a meaningful interpretation of a spectral analysis. Here we wish to establish this point somewhat more securely and show in what way improper sampling, sampling at too low a rate for a given bandwidth, obscures and falsifies spectral analysis. Let us begin by considering the continuous signal $x(t)$ to be periodic T, and to have an unlimited bandwidth. The Fourier series representation for such a signal is given by

$$x(t) = \sum_{n=-\infty}^{\infty} X_T(n) \exp(2\pi jnt/T) \tag{3.14}$$

(3. 14) where

$$X_T(n) = \frac{1}{T} \int_0^T x(t) \exp(-2\pi jnt/T)\, dt \tag{3.15}$$

(3. 15) We wish to deal with the sampled representation $x(t^o\Delta)$ and so we sample $x(t)$ every Δ sec, obtaining N samples such that $T = \Delta$. We then blindly take the DFT,

$$x^\dagger(n) = \sum_{t^o}^{N-1} x(t^o\Delta) \exp(-2\pi jnt^o/N) \tag{3.16}$$

(3. 16) We have used the dagger symbol to indicate our suspicion that something may be amiss in this representation, i.e., that $X_T^\dagger(n)$ may not be the same as $X^T(n)$. That such is the case may be seen by substituting for each sample value its Fourier series expansion as given by Eg. (3. 14) ;

$$One\, long\, equation! \tag{3.17}$$

(3.17) The exponential term here has the important property that when $m - n - 0$ or some integer multiple of N, the summation over t^o is equal to N; otherwise it is identically O. That is, for fixed m,

$$\sum_{t^o=0}^{N-1} \exp[2\pi j(m-n)t^o/N] = \begin{cases} N, & m = kN + n \\ 0, & m \neq kN + n \end{cases} \tag{3.18}$$

where k is an integer. Using this fact in Eg. (3.17), it can be seen that

$$X^{\dagger}(n) = N \sum_{k=-\infty}^{\infty} X_T(kN+n) \tag{3.19}$$

(3.19)

This means that each term in the DFT of $x(t)$ is the sum of a possibly infinite set of Fourier coefficients associated with the higher frequency components in $x(t)$. The higher frequency components are those corresponding to frequencies that are greater than N by an amount kN. If $x(t)$ has no Fourier series components for values of n equal to or greater than N/2 (corresponding to frequencies $1/2\Delta$ or greater), $X_N^{\dagger}(n) = NX_T(n) = X_N(n)$; otherwise, $X_N^{\dagger}(n) \neq NX_T(n)$. This means that the DFT for $x(t)$ yields correct results only if $x(t)$ is band-limited to frequencies below $1/2\Delta$. When $x(t)$ has a greater bandwidth, the high frequency components add to the low frequency ones, an effect that is called aliasing because the high frequency components are misrepresented or misinterpreted as low frequency ones. Once aliasing occurs, there is no way to properly sort out the $X_T(n)$ components from the $X_N^{\dagger}(n)$. This is why the cut off frequency F of the analog prefilter must be matched to the sampling rate such that $F \leq 1/2\Delta$. It is essential to the proper analysis of continuous data by sampling techniques. The numerical value of n corresponding to the highest frequency representable by the sampling procedure is $N/2$. As shown previously, it is determined by the relation $n/T = 1/2\Delta$. To see the effect of aliasing more clearly, consider Fig. 3. 1 which shows a cosine wave of frequency $F = l/2\Delta$ being sampled at the negative and positive peaks. If the frequency of the wave increases a little above F to $F + a$ (dotted line), sine waves of frequency $F + a$ and $F - a$ can be drawn through the sampling points equally well. This gives us reason to suspect that a wave of real frequency $F + a$ will, after sampling, be confused with a wave of real frequency $F - a$. With this in mind, let us examine Eq. (3.19) when n has a value of $(N/2) - i$. Then all the $X_T(kN + n)$ such that

$$kN + n = kN + (N/2) - i = (k + l/2)N - i$$

will contribute to the terms $X_N^{\dagger}(N/2) - i]$. A real frequency term at $(N/2) - i$ corresponds to complex frequency terms $X_T[(N/2) - i]$ and $X_T[(-N/2) + i]$. The aliases of $X_T[(N/2) - i]$ are at frequencies $\ldots, (-3N/2) - i, (-N/2) - i, (3N/2) - i, \ldots$ while the aliases of $X_T[(-N/2 + i]$ are at frequencies $\ldots, (-3N/2) + i, (N/2) + i, (3N/2) + i, \ldots$. If we group these aliasing terms in pairs, one term from each sequence, we find that $X_T[(-N/2) - i]$ pairs with $X_T(N/2) + i)$ to give a real frequency term at $(N/2) + i$. Similarly, there are real frequency terms

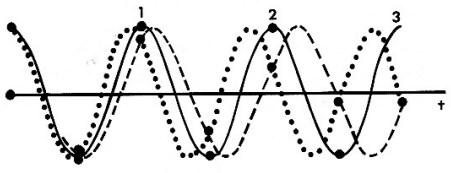

Figure 3.1: Fig. 3.1. A cosine wave of frequency F (solid line) sampled at its Nyquist rate. A higher frequency (dotted) wave, frequency F + a, is shown sampled at the same rate. At the sample times it is indistinguishable from a lower frequency (dashed) wave, frequency F - a.

at $(3N/2) + i$, $(3N/2) - i$, $(5N/2) + i$, $(5N/2) - i$, etc. Thus a real frequency data component at $(N/2) - i$ will have alias contributions from whichever of these higher frequency terms that are present in the data input to the ADC. In effect the original Fourier representation of x(t) has been folded in accordion fashion about frequencies that are multiples of $1/2\Delta$ and collapsed into the frequency region extending from 0 to $1/2\Delta$ which is also called the folding frequency, (Fig. 3.2)

It is of some interest that aliasing effects can also enter into sampled representations of data that are band-limited to the Nyquist frequency. We have seen previously how the discrete Fourier transform is a completely adequate representation of a continuous periodic band-limited signal as long as the signal samples are taken frequently enough to eliminate the possibility of aliasing. But in actuality, few of the data one analyzes are periodic or band- limited, although the latter condition can be approached as closely as desired by analog prefiltering prior to sampling. Periodicity is another matter. Even when periodic stimulation is employed and the response or signal component of the data is periodic, the remainder, the noise, is not. Periodicity is then lacking in the data. What the data analysis procedure does in this situation is to effectively create periodic data from the T sec data segment we have available to study. That is, we analyze the T sec segment as though it originated from a process with period T or greater. This introduces some complications which we need to consider. The "periodicized" process created from a T sec segment of data (1) is generally not band-limited even if the original data are, (2) can contain frequency components, apart from aliases, that are not present in the original data. Let us deal with these complications in order, using as an illustration a signal that is both band-limited and periodic, a cosine wave whose period is 3T/8, T being the period of its observation. The periodicized version of this signal is shown in Fig. 3.3. It is clear that there are discontinuities in the periodicized signal which guarantee that it will not be band-limited. In fact, it may be stated that

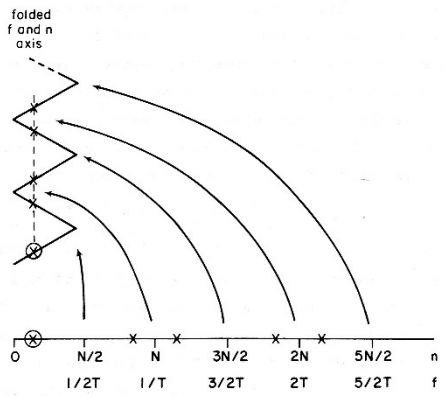

Figure 3.2: Fig. 3.2. The accordion like folding of the frequency (or n) axis due to sampling of a continuous signal. Frequency components of the original signal marked with x's on the faxis are interpreted in the sampled version as belonging to the lowest frequency, an encircled x.

unless the original signal has rather special properties, i.e., that its amplitude and time derivatives at t = 0 are the same as those at t = T, there will be discontinuities in the periodic waveform and its derivatives that guarantee that the periodicized signal will not be band-limited. We know that if we sample this process, every Δ sec such that T = NΔ, we are sure to encounter aliasing, its severity depending upon the sampling rate. If we apply the DFT to the samples and treat the resulting Fourier coefficients as though there were no aliasing involved, we effectively consider the data as having arisen from a periodic band-limited process, i.e., one that has no discontinuities of any kind at the ends of the interval. This recreated signal is also shown in Fig. 3. 3 for $N = 16, \Delta = T/16$. This means that the sampling has distorted the original data, primarily at the ends of the interval. The high frequency components associated with the discontinuities at 0 and T have been aliased into the spectral representation. The numeric results obtained from the DFT show the results of this aliasing. Both covariance and spectral analysis of the data can be affected. Fortunately, the larger N is, the smaller the end effects tend to become. They also diminish as the severity of the discontinuities diminishes.

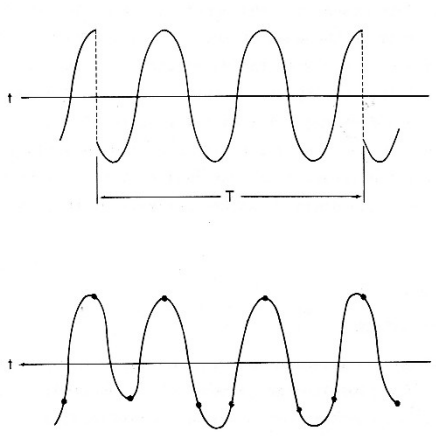

Figure 3.3: Fig. 3. 3. Top, a periodicized segment of a cosine wave. T is the observation time and 3T/8 the period of the wave. Note the discontinuities at 0 and T. Bottom, a continuous and periodic band-limited wave drawn through the sample points $\Delta = T/16$ sec apart.

3.4 Leakage

3.4.1 A. Fourier series

Besides the aliasing that is introduced into the DFT representation of a time- limited segment of a nonperiodic signal, we must deal with another form of signal misrepresentation, referred to as spectral leakage. It occurs with all aperiodic data and even with periodic band-limited data whose period is not integrally related to the time of observation. In the Fourier analysis procedures, the frequency composition of the data is computed to be a set of frequency constituents harmonically related to l/T, the fundamental of the time of observation. The frequency components that are closest to the original frequencies in the data contribute most to the analysis, but more remote frequencies may also be interpreted as being present When in fact they are not. To see a specific example of this, consider the signal to be the cosine wave whose period is $3T/8$ (Fig. 3.3). We compute the Fourier series representation of this signal first because it avoids all aliasing effects. The Fourier series coefficients are given by

$$X_T(n) = \frac{1}{T}\int_0^T \cos(2\pi\frac{8}{3T}t)\exp(-2\pi j\frac{n}{T}t)\,dt$$

for $-(N-1)/2 \leq n \leq (N-1)/2$.

$$A_T(n) = \frac{2}{T}\int_0^T \cos(2\pi\frac{8}{3T}t)\cos(2\pi\frac{n}{T}t)\,dt \tag{3.20}$$

$$B_T(n) = \frac{2}{T}\int_0^T \cos(2\pi\frac{8}{3T}t)\sin(2\pi\frac{n}{T}t)\,dt$$

for $0 \leq n \leq (N-1)/2$. The values for $A_T(n)$ and $B_T(n)$ are obtained by standard integration formulas and are tabulated in Table 3. 1 for $n = 1, 2, \ldots 8$.

Table 3. 1 Fourier Series and DFT Coefficients for cos(2n8t/3T)

Inspection of the Fourier components as determined by Eq. (3. 20) reveals that the analysis has decomposed the original cosine wave into frequency components at all values of n. None of these corresponds to the frequency of the original signal which lies slightly below n = 3, but the coefficients are largest at n = 3 and next largest at n = 2. There is a gradual diminution of component amplitudes as n departs from these values. What has happened is that the power of the original signal has been dispersed or "leaked " out from the original signal frequency into the neighboring frequencies of the Fourier analysis. No spurious power is added by the analysis, for if all the $A_T(n)$ and $B_T(n)$ were squared and summed, their total contribution would equal that of the original signal in the T sec interval. The net effect, however, is a rather serious misrepresentation of the original signal whose spectrum is a single real frequency component at 8/3T. The cause of the misrepresentation is that only a finite length of the signal segment has been used for the analysis. It is possible to show that the Fourier representation of a T sec segment of data results from a convolution of the spectrum of the original, infinite duration signal with the sinc function $\sin(\pi nt/T)/(\pi nt/T)$. To see how this comes about, we refer back to the expression for $X_T(n)$ in Eq. (3. 20) where we replace the illustrative frequency $8/3T$ by the general frequency f so that $x(t) = \cos 2\pi ft$. We can calculate the $A_T(n)$ and $B_T(n)$ for this signal and find them to be

$$A_T(n) = \frac{1}{T}\left[\frac{\sin 2\pi T(f-(n/T))}{2\pi(f-(n/T))} + \frac{\sin 2\pi T(f+(n/T))}{2\pi(f+(n/T))}\right]$$

$$B_T(n) = -\frac{1}{T}\left[\frac{\cos 2\pi T(f-(n/T))}{2\pi(f-(n/T))-1} - \frac{\cos 2\pi T(f+(n/T))-1}{2\pi(f+(n/T))}\right] \tag{3.21}$$

(3. 21) The terms containing $f - n/T$ and $f + n/T$ are a manifestation of the fact that cosine and sine waves consist of positive and negative complex frequency terms. We are considering real (positive) frequency data and so both f and n are greater than O. In most cases f will be sufficiently greater than 0 to make the second

term of Eg. (3.21) negligible compared to the first. This results in the approximation

$$A_T(n) \simeq \frac{1}{T}\frac{\sin 2\pi T(f-(n/T))}{2\pi(f-(n/T))}$$

$$B_T(n) = -\frac{1}{T}\frac{\cos 2\pi T(f-(n/T))}{2\pi(f-(n/T))-1} \tag{3.22}$$

(3.22) From this we obtain the spectral power at real frequency n/T:

$$\begin{cases} |X_T(n)|^2 + |X_T(-n)|^2 & = \frac{1}{2}[|A_T(n)|^2 + |B_T(n)|^2] \\ & \simeq [\frac{\sin \pi T(f-n/T)}{\pi T(f-n/T)}]^2 \end{cases} \tag{3.23}$$

(3.23)

The total power of $x(t) = cos2\pi ft$ is 1/2 and is concentrated solely at frequency f. The Fourier analysis has in effect dispersed or leaked this power out into neighboring frequencies that are harmonically related to $1/T$. This also means that if one is interested in estimating the spectral component of the data at a particular frequency, there will be included in the estimate a contribution from nearby spectral components that have had their power leaked into the frequency where the estimate is being made. The weighting factor for these extraneous contributions is that given by the bracketed term in Eq. (3.23). It shows that the larger T becomes, the smaller is the frequency range over which leakage is a significant factor.

Leakage may also magnify the undesirable effects of 60 Hz or other single frequency artifacts in the data. These may arise from a variety of causes: ineffective electrical shielding, stray coupling of stimulus frequencies into the responses, and so on. An important attribute of a signal with a line spectrum, one expressed by delta functions in the spectrum, is that a rather substantial amount of power is confined to an infinitesimally narrow frequency band rather than being spread out over a broader range of frequencies. It is this concentration of power that can be so potent in producing leakage into the estimates of power density in the neighboring regions of the spectrum. The leakage occurs, as Eq. (3. 23) indicates, if the line component is not exactly located at a harmonic of the fundamental analysis interval. To see this, suppose a spurious line component is located midway between adjacent harmonic frequencies of the analysis interval and that the rms strength of the line is σ_a. The leakage of this component into the neighboring frequency terms is well approximated by Eq. (3. 23) as long as the line is reasonably far from 0 frequency. It can be seen that the larger N is, the narrower will be the frequency range over which significant amounts of leakage occur. Because of the side lobes of the sine function, leakage effects can occur between rather widely spaced frequencies when a is large It is also true that the closer the frequency of a line component is

to a harmonic of the analysis interval, the smaller is the leakage effect. The most generally useful way of minimizing leakage is by means of spectral "windowing" techniques of which more will be said later. These techniques, which are another form of linear filtering, have the effect of estimating the spectrum in a way that greatly minimizes the side lobe contributions to the spectral estimate.

3.4.2 B. Discrete Fourier transforms

Leakage is not alleviated by resort to the DFT. Rather, the situation persists and is also overlayed with aliasing effects so that the resulting data representation contains both inextricably combined. To see this we refer again to the signal $x(t) = cos2n8t/3T$ and represent it by its DFT as given by Eqs. (3. 7) and (3.9), rewritten here for N = 16:

$$X_N(n) = \sum_{t^o=0}^{15} \cos(2\pi t^o/6) \exp(-2\pi j n t^o/16)$$

$$A_N(n) = \sum_{t^o=0}^{15} \cos(2\pi t^o/6) \cos(-2\pi j n t^o/16)$$

$$X_N(n) = \sum_{t^o=0}^{15} \cos(2\pi t^o/6) \sin(-2\pi j n t^o/16) \tag{3.24}$$

(3. 24a) (3. 24b) (3.24c)

In Table 3.1 we show the DFT coefficients for n ranging from 1 to 8 when there are two different sample intervals, the first being T/16 with N = 16, and the second T/256 with N = 256. The discrepancy between the tabulated values for either situation and those obtained from the continuous Fourier series expansion arises from the aliasing introduced by sampling. As the sampling interval becomes shorter, the discrepancy diminishes and what remains is the pure leakage effect. Again, what causes it is the finite length of the signal segment, N samples in duration. If we let $x(t) = \cos 2\pi f t$ and perform a calculation similar to that just done for the Fourier series, we find that power has leaked from frequency f into frequency n/N. The amount that has leaked is given by

$$\begin{cases} |X_N(n)|^2 + |X_N(-n)|^2 & = \frac{1}{2}[|A_N(n)|^2 + |B_N(n)|^2] \\ & \simeq [\frac{\sin \pi T(f\Delta - n/N)}{N \sin \pi (f\Delta - n/N)}]^2 \\ & \text{Is that right? Sine in the denominator?} \end{cases} \tag{3.25}$$

The approximation arises as before because of the fact that we have ignored the usually small terms involving $f + (n/N)$. From Eq. (3. 25) we see that the

leakage from frequency f into the nth component of the DFT has very nearly the same behavior as it had for the Fourier series representation. Thus leakage in the two cases is comparable although the leakage in the DFT tends to be the larger of the two because the denominator of Eq. (3. 25) is smaller than that of Eq. (3. 22).

Another aspect of leakage is associated with the presence of a constant dc component in the data. If only the spectrum of the data is of interest, leakage is not a factor because the steady component shows up only in the n = 0 term of the Fourier representation. But when one uses the spectrum as an intermediary step for obtaining an estimate of the acvf (or ccvf) of the data, then leakage does become a factor. Such a procedure is quite common when one employs the fast Fourier transform to first obtain the spectral estimate and then the acvf from it. The reason that leakage becomes a factor is that in this procedure it is necessary to pad out the original sequence of N data points with a sequence of zero amplitude samples, L of them if one wishes to estimate the acvf for lags up to $L\Delta$ This means that the DFT that one,works with is

$$X_{N'}(n) = \sum_{t^o=0}^{N-1} x(t^o\Delta)\exp(-2\pi jnt^o/N') \tag{3.26}$$

(3. 26) The upper limit is N - 1 rather than N' - 1, (N ' = L + N), because the last L values of $x(t\Delta)$ are taken to be 0. When $x(t)$ has an average value a, the contribution of this to $x_N(n)$ is

$$[X_{N'}(n)]_{dc} = a\sum_{t^o=0}^{N-1}\exp(-2\pi jnt^o/N') = a\frac{1-\exp(-2\pi jnN/N')}{1-\exp(-2\pi jn/N')} \tag{3.27}$$

(3. 27) The contribution to the raw spectral estimate $[C_{XX}(n)]_{dc} = |[X_{N'}(n)]_{dc}|^2$ follows directly. It is

$$|[X_{N'}(n)]_{dc}|^2 = a^2[\frac{sin(\pi nN/N')}{\sin(\pi n/N')}]^2 \tag{3.28}$$

(3. 28) For n = 0, the result is $(aN)^2$ as is to be expected. If a data record of length N = 1000 were padded with 10 zeros to permit estimation of the acvf out to lOΔ, the dc leakage at n = 1 would be $100a^2$. If the record were padded with 100 zeros, the dc leakage at n = 1 would be $1.053x10^4a^2$. The effect obviously depends upon the strength of the dc term. In the second case, if a is 5 times the amplitude of the real component at n = 1, one could expect an error in the spectral estimate amounting to about 24%, a rather serious matter. To eliminate leakage, a good procedure is to first remove the average value from the data before padding it with zeros.

3.5 Trend

Another effect that we need to be aware of is one that is brought about by the presence of very low frequency components in the data, frequencies that are less than that of the fundamental frequency of the analysis interval. Such components are referred to as producing trends in the data. These are progressive changes in the short term mean of the data, a mean that is calculated over a relatively small segment of the data. Trends may also be found in other properties of the data such as the variance and covariance functions, but here we are concerned only with trends in the mean and, more specifically, linear trends, i.e., those trends that can be described by data having the form $x(t) = bt + v(t)$, bt being the trend component and v(t) the component one normally considers in a trend-free situation. It is also possible to take into account trends which are not linear (Otnes and Enochson, 1974) but here we are only interested in seeing how linear trends affect a spectrum analysis. When a linear trend is present in an N sample sequence of data, it will contribute to the DFT according to

$$[X_N(n)]_{trend} = \sum_{t^o=0}^{N-1} bt^o \exp(-2\pi j n t^o/N) \tag{3.29}$$

(3. 29) The expression can be summed without difficulty. When n is small compared to N, we find that

$$[C_{XX}(n)]_{trend} = |[X_N(n)]_{trend}|^2 \simeq (bN/2\pi n)^2 \tag{3.30}$$

(3. 30)

In effect the trend leaks into the nearby low frequency components in a manner that is inversely proportional to n^2. Note that bN is the total trend in the data from the beginning to the end of the sequence. To eliminate contamination of the spectral estimates by trends, the trends should be estimated and removed before a spectrum analysis. Procedures for doing this are given in Otnes and Enochson (1974) and Blackman and Tukey (1958).

3.6 The power spectrum, general considerations

When investigating the properties of samples of random variables, it is useful to characterize them by population statistics. In the case of a simple, univariate random variable, the mean is a measure of its location (from zero), and the variance is a measure of its dispersion about the mean. These two statistics are also of use when investigating random signals. The mean specifies a baseline about which the signal

fluctuates. If the physical signal is an electrical one, then the mean corresponds to the de level of the signal. The variance provides a measure of the magnitude of the signal's fluctuation about its mean. For electrical signals, the variance corresponds to the power of the ac component of the signal. While the mean and variance are useful and readily computed statistics, they provide no information concerning the temporal character of the fluctuation of a random signal. We cannot infer from them whether the signal's fluctuations are slow or rapid or whether they possess some rhythmicity or a high degree of irregularity. However, as we noted in Chapter 1, if the signal is wide sense stationary, such information can be provided by the power spectrum of the signal. The power spectrum provides a statement of the average distribution of power of a signal with respect to frequency. If the signal varies slowly, then its power will be concentrated at low frequencies; if the signal tends to be rhythmic, then its power will be concentrated at the fundamental frequency of the rhythm, perhaps at its harmonic frequencies, if the signal lacks rhythmicity, then its power will be distributed over a broad range of frequencies.

A way of obtaining an estimate of the power spectrum of a signal at a given frequency is to pass the signal through a narrow band linear filter centered at the frequency of interest, and then to compute the variance (power) of the filter output. This operation can be performed at any frequency of interest. The variance of the output of the filter will be proportional to the amount of power in the signal at frequencies close to the filter center frequency. The variance can then be plotted as a function of the filter's center frequency and the resulting graph will be an approximate indication of the frequency distribution of the signal's power. This filtering approach was the traditional way of analyzing spectra before the advent of high speed digital computers. It is still useful conceptually although the mechanization of the filtering techniques has been changed drastically by the computer. The concept of a power spectrum applies to both T-continuous and T-discrete signals. Because we are usually interested in continuous signals, we will begin with a discussion of the power spectra of wide sense stationary continuous signals. Then we move to consider more fully the computation and interpretation of power spectra from wide sense stationary sampled data. This is the representation of continuous signals that digital computers usually operate upon.

Illustrations of how a power spectrum characterizes the temporal behavior of a signal are provided in the following examples. First, consider an EEG recording from a subject in deep sleep (Fig. 3.4a). In such a case the BEG consists primarily of slowly fluctuating, high amplitude delta wave activity. Consequently, most of the power is concentrated at low frequencies and so the spectrum will be relatively large at those frequencies, and small elsewhere (Fig. 3.4b). As a second example, consider the EEG of an awake but resting subject. In this case the EEG may consist of primarily rhythmic, quasisinusoidal alpha wave activity in the 9 to 12 Hz

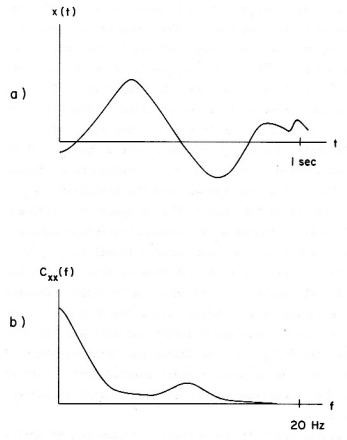

Figure 3.4: Fig. 3.4. (a) A hypothetical example of a low frequency EEG waveform recorded from an individual in deep sleep. (b) Power spectrum corresponding to the low frequency EEG process.

frequency range (Fig. 3.5a). The associated power spectrum will have a peak in the 9 to 12 Hz range and be relatively small elsewhere (Fig. 3.Sb). In the third example, consider the EEG of an alert subject. Here the EEG tends to consist of low amplitude waves with rapid, irregular fluctuations (Fig. 3.6a). No predominant rhythms or slow fluctuations are apparent. The corresponding power spectrum will tend to be broadly distributed over the frequency range of the EEG (Fig. 3. 6b), a range which extends to an upper frequency of about 30 to 50 Hz.

The three foregoing examples illustrate how the power spectrum provides a characterization of the "average" temporal behavior of a random signal. But it does not uniquely specify the signal it is derived from. One cannot reconstruct the signal given only its power spectrum because the power spectrum does not preserve the phase information in the signal. In effect, the spectrum specifies the average strength of a signal at each frequency. The average strength at a given frequency reflects both the amount of time during which there is activity at that frequency and the strength of that activity. For example, consider Fig. 3. 7 which illustrates both a persistent, relatively low amplitude rhythmic random signal (a), and a signal in which relatively high amplitude bursts of rhythmic activity occur irregularly (b). The magnitudes of the power spectra corresponding to the two signals may be the same near the frequency of the rhythm. Although the signal in Fig. 3.7b has higher amplitudes during the bursts of rhythmic activity, the average power near the frequency of the rhythm is no greater than that of the signal in Fig. 3.7a because

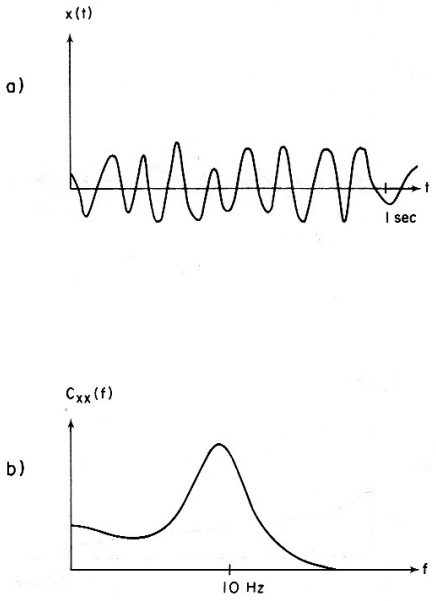

Figure 3.5: Fig. 3. 5. (a) A hypothetical example of EEG alpha activity. (b) Power spectrum corresponding to an EEG process with pronounced alpha activity.

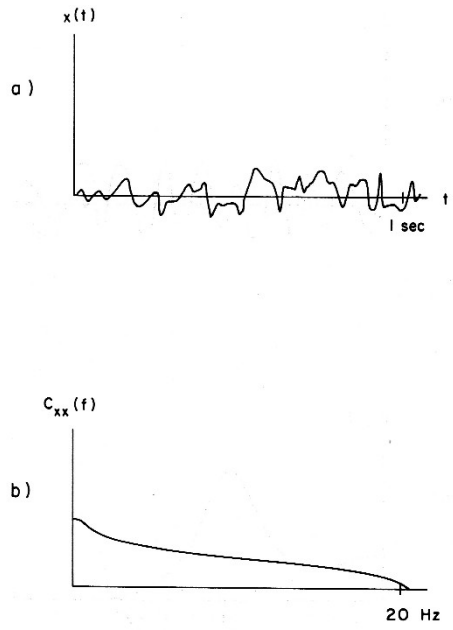

Figure 3.6: Fig. 3. 6. (a) A hypothetical example of rapid, irregularly fluctuating EEG recorded from an alert individual. (b) Power spectrum corresponding to the rapid, irregularly fluctuating EEG.

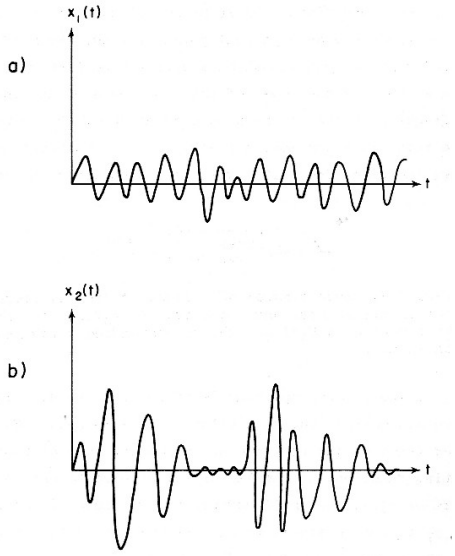

Figure 3.7: Fig. 3. 7. Hypothetical example of (a) a low amplitude random process with persistent rhythmic activity, and (b) a random process with irregularly occurring bursts of high amplitude rhythmic activity.

the duration of the rhythmic activity in (b) is less than in (a).

3.7 Power spectrum of continuous random signals

In the above discussion we presented the concept of the power spectrum from an empirical point of view. We held that the variance of the output signal of a narrow band linear filter provides a measure of the power of the components of the input signal whose frequencies are in the pass band of the filter. We now examine this statement more closely, taking a mathematical point of view. Consider Fig. 3.8. $x(t)$ is a wide sense stationary random signal whose power spectrum is of interest to us. For simplicity, we assume that the mean value of $x(t)$ is zero. $H(f)$ is the transfer function and $h(\tau)$ the corresponding impulse response (weighting function) of the linear filter used to obtain a spectral estimate of $x(t)$. We shall compute the variance of the filter's output $x_h(t)$ and relate it to $x(t)$ as well as to $h(\tau)$ and $H(f)$ and to the power spectrum of $x(t)$, $C_{XX}(f)$.

We first state the output signal in terms of the convolution relation between

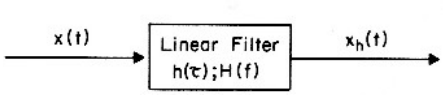

Figure 3.8: Fig. 3.8. Block diagram of a linear filtering operation. The input signal is $x(t)$ and the output is $x_h(t)$. The filter transfer function is $H(f)$ and the corresponding weighting function is $h(\tau)$

output and input, established in Chapter 2:

$$x_h(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)/, d\tau \tag{3.31}$$

(3 .3l) The variance of the output can be expressed in terms of the variance of the input signal and the filter impulse response function, as given above. Since the output, like the input, has zero mean,

$$var[x_h(t)] = E[x_h^2(t)] = E\{[\int_{-\infty}^{\infty} h(\tau)x(t-\tau)/, d\tau]^2\} \tag{3.32}$$

(3.32)

Now the square of an integral can be exPressed as the product of two identical integral s, differing only in the symbols used to denote the variable over which the integrati on is performed. Then we have

$$var[x_h(t)] = E[\int_{-\infty}^{\infty} h(\tau)x(t-\tau)/, d\tau \int_{-\infty}^{\infty} h(u)x(t-u)/, du] \tag{3.33}$$

(3.33) Since the averaging operation is with respect to the random variable $x(t)$, Eq. (3.33) can be rearranged so that the averaging operation is performed prior to integration over τ and u.

$$var[x_h(t)] = \int_{-\infty}^{\infty} h(\tau)/, d\tau \int_{-\infty}^{\infty} h(u)E[x(t-\tau)x(t-u)]/, du] \tag{3.34}$$

(3.34) $E[x(t-\tau)x(t-u)]$ is the autocovariance function (acvf) of $x(t)$. Since $x(t)$ is wide sense stationary, the acvf is a function only of the difference between τ and u. Denote the acvf by $c_{XX}(t)$ and substitute $c_{XX}(\tau - u)$ into Eq. (3.34). This gives

$$var[x_h(t)] = \int_{-\infty}^{\infty} h(\tau)\, d\tau \int_{-\infty}^{\infty} h(u)c_{XX}(\tau-u)]\, du] \tag{3.35}$$

(3.35) Equation (3.35) indicates that the variance of $x_h(t)$, the filter output, is determined solel y by the filter characteristics and the second-order statistics (acvf) of the input signal. However, Eq. (3.35) does not show clearly just how the filter's action upon the input signal determines the variance of $x_h(t)$. This can be brought out if Eq. (3.35) is expressed in terms of the frequency response of the filter and the spectrum of the signal as we shall do in the next step. But some comments upon this step are first in order. Up to this point we have used a deductive argument to arrive at Eq. (3.35). We have assumed nothing about the nature or even the existence of the power spectrum. We have only assumed that the input process is stationary and that it has the acvf $c_{XX}(t)$. We now make use of the fact, first mentioned in Chapter I, that the power spectrum and the acvf of a wide sense stationary process constitute a Fourier transform pair. The power spectrum is the direct Fourier transform of the acvf, and the acvf is the inverse Fourier transform of the power spectrum. The latter is indicated below, with the power spectrum of $x(t)$ denoted by $C_{XX}(f)$.

$$c_{XX}(t) = \int_{-\infty}^{\infty} C_{XX} \exp(2\pi j f t)\, df \tag{3.36}$$

(3 .36) Substitution of Eg. (3.36) into Eq. (3.35) yields an expression which relates the variance of the filter output to the power spectrum of the input signal:

$$var x_h(t) = \int_{-\infty}^{\infty} h(\tau) d\tau \int_{-\infty}^{\infty} h(u) du \int_{-\infty}^{\infty} C_{XX}(f) \exp(2\pi j f(\tau - u))\, df \tag{3.37}$$

(3 .37) Equation (3 .37) can be further simplified by changing the order of integration, as follows:

$$var[x_h(t)] = \int_{-\infty}^{\infty} C_{XX}(f)\, df \int_{-\infty}^{\infty} h(\tau) \exp(2\pi j f \tau)\, d\tau \int_{-\infty}^{\infty} h(u) \exp(-2\pi j f u)\, df \tag{3.38}$$

(3.38) The two right-most integrals in Eg.(3.38) are Fourier transforms of the filter impulse response, and hence may be stated in terms of the filter's transfer function:

$$\int_{-\infty}^{\infty} h(u) \exp(-2\pi j f u)\, du = H(f) \int_{-\infty}^{\infty} h(\tau) \exp(2\pi j f \tau)\, d\tau = H(-f) = H^*(f) \tag{3.39}$$

(3.39)

Note that $H^*(f)$ is the complex conjugate of $H(f)$. since the product of a complex quantity and its conjugate equals the squared magnitude of the quantity, substitution of Eqs. (3.39) into Eq.(3.38) yields

$$var[x_h(t)] = \int_{-\infty}^{\infty} C_{XX}(f) |H(f)|^2\, df \tag{3.40}$$

(3. 40) This can be seen to specify the variance of the filter's output in terms of both the power spectrum of the input signal and the squared magnitude of the filter's transfer function.

Equation (3 .40) indicates that the power spectrum of a random signal is the density of average power at a given frequency.The units are power per Herz. To see this, suppose that the filter transfer function is unity over a narrow band b of frequencies centered at frequency f_c and zero elsewhere. Then,

$$|H(f)| = \begin{cases} 1, \ f_c - \frac{b}{2} \leq f \leq f_c + \frac{b}{2} \\ 0 \quad elsewhere \end{cases} \tag{3.41}$$

(3.41) Substitution of Eq. (3.41) into Eq. (3.40) yields

$$var[x_h(t)] = \int_{f_c - b/2}^{f_c + b/2} C_{XX}(f) df \tag{3.42}$$

(3.42) Since b is small, the integral in Eq. (3.42) can be approximated by

$$var[x_h(t)] \simeq b C_{XX}(f_c) \tag{3.43}$$

(3.43) Rearranging Eq.(3.43), and taking the limit as b becomes infinitesimally small, yields

$$C_{XX}(f_c) = \lim_{b \to 0} \frac{var[x_h(t)]}{b} \tag{3.44}$$

(3.44) Note that $var[x_h(t)]$ represents the total average power of the random process in the narrow pass band of the filter : $f_c - (b/2)$ to $f_c + (b/2)$ Thus, from Eq. (3.44) it can be seen that the spectrum is a density function. The integral of $C_{XX}(f)$ over all frequencies equals the total power of the random process. This can be inferred from Eq. (3.40) by setting $|H(f)|^2 = 1$ for all f. Passing a signal through a filter with a transfer function of unity magnitude in no way alters the amount or the frequency distribution of the average power of a signal. Hence, for this case Eq. (3.40) reduces to

$$var[x(t)] = \int_{-\infty}^{\infty} C_{XX}(f) \, df \tag{3.45}$$

(3.45) It is useful here to reconsider two important properties of the power spectrum previously discussed in Chapter 1.

- (1) As Eqs. (3 .40) and (3.42) indicate, $C_{XX}(f)$ is nonnegative at all frequencies.

- (2) It is an even function of frequency.

With regard to the first property, if negative values could occur, then by suitable filtering one could obtain an output signal with negative power. However, this is impossible since the power of a signal is the signal's variance, and variance, being the average of a squared quantity, can never be negative. The second property can be inferred from the Fourier transform relationship between the power spectrum and acvf, as follows.

$$C_{XX}(f) = \int_{-\infty}^{\infty} c_{xx}(t) \exp(-2\pi j f t)\, dt \tag{3.46}$$

(3.46) Replacing the exponential in Eq. (3.46) with its Euler identity yields

$$C_{XX}(f) = \int_{-\infty}^{\infty} c_{xx}(t)(cos2\pi ft - jsin2\pi ft)dt \tag{3.47}$$

(3.47) Since $c_{xx}(t)$ is an even function of t and $sin2\pi ft$ is an odd function of t, the integral of the product of the acvf with the sinusoid will be zero. Hence

$$C_{XX}(f) = \int_{-\infty}^{\infty} c_{xx}(t)cos2\pi ft\, dt \tag{3.48}$$

(3.48) Changing f to - f in Eq. (J .48) does not alter the cosine and therefore does not alter the integral. Consequently, $C_{XX}(f)$ must be a real, even function of .

3.8 The power spectrum of T-Discrete random signals

Use of a digital computer for power spectrum computations requires that the continuous signal be sampled.It is important that aliasing errors be avoided if an accurate estimate of the power spectrum is to be obtained.When the signal is band-limited, sampling at the Nyquist rate or faster will insure that aliasing will not occur. If the signal is not band-limited or cannot be sampled at twice its upper band-limit, then it should be low- pass filtered prior to sampling, so that activity at frequencies above One-half the sampling frequency will be effectively eliminated. A power spectrum estimate that is free of aliasing errors can then be obtained for frequencies below one- half the sampling frequency. However, information concerning activity at higher frequencies will necessarily be lost. Although the power spectrum properties of discrete signals are closely related to those of the original continuous signals, there are important differences which it is most useful to discuss.

The two approaches commonly used to estimate power spectra via digital computation are:

- (1) The estimation f irst of the acvf illld from it the power spectrum by the use of the discrete Fourier transform (DFT) .
- (2) The computation of the periodogram, the "raw" spectrum estimate, by applying the DFT to a finite N sample segment of the signal .

with the advent of the fast Fourier transform algorithm (Oppenheim and Schafer, 1975), the periodogram approach is usually the more rapid one. Once the periodogram has been obtained, further steps are necessary to improve the goodness of the spectral estimate. We will discuss these after paying initial at'tention to the properties of the periodogram.

3.9 The Fourier transform for T-Discrete signals

The Fourier transform relationship between the power spectrum and the acvf for T-continuous signals has been developed and discussed in Chapter 1. 111e Fourier transform pair is restated here.

$$C_{XX}(f) = \int_{-\infty}^{\infty} c_{xx}(t) \exp 2\pi f t \, dtexp(-j2wft)dtxx \tag{3.49}$$

(3.49)

$$C_{xx}(t) = \int_{-\infty}^{\infty} C_{CC}(f) \exp 2\pi f t \, df \tag{3.50}$$

(3.50) An analogous relationship can be shown to hold for T-discrete signals. If the period between samples is Δ sec and the upper bandlimit of the signal is less than or equal to $1/2\Delta$, then Eq. (3.50) becomes

$$c_{xx}(t^o\Delta) = \int_{-1/2\Delta}^{1/2\Delta} C_{XX}(f) \exp(2\pi f t^o \Delta, df \tag{3.51}$$

(3.51) The acvf is defined only at the discrete times of $t^o\Delta$, where t^o is an integer that can range from minus to plus infinity.However, $C_{XX}(f)$ is a continuous function of frequency. Note that Eq. (3 .51) is obtained from Eq. (3.50) by direct substitution of $t\Delta$ for t and setting the limits of integration to correspond to one-half the Nyquist frequency.

The discrete analog of Eq.(3.49) is a summation over the discrete set of acvf values :

$$C_{XX}(f) = \Delta \sum_{t^o=-\infty}^{\infty} c_{xx}(t^o\Delta) \exp(-2\pi j f t^o \Delta) \tag{3.52}$$

(3.52) When Eq.(3.52) is compared with Eg. (3.49), we see that $t\Delta$ replaces t, a summation replaces the integral, and the finite time increment Δ replaces the infinitesimal dt.The correspondence between Eq. (3.52) and (3.49) has been given here by making some intuitively reasonable changes in the original T-continuous transform pair. We will now demonstrate that the relationship is a mathematically valid one. This is done by substituting for $C_{XX}(f)$ in Eq. (3.51) the right side of Eq. (3.52) .

$$C_{XX}(t^o\Delta) = \int_{-1/2\Delta}^{1/2\Delta} \Delta \sum_{\tau=-\infty}^{\infty} c_{xx}(\tau^o\Delta) \exp(-2\pi j f \tau^o \Delta) \exp(2\pi j f t^o \Delta) df \tag{3.53}$$

(3.53) Interchange of the order of integration and summation yields

$$c_{XX}(t^o\Delta) = \Delta \sum_{\tau=-\infty}^{\infty} c_{xx}(\tau^o\Delta) \int_{-1/2\Delta}^{1/2\Delta} \exp(-2\pi j f (t^o - \tau^o)\Delta) df \tag{3.54}$$

(3 .54) The integral on the right side is easily shown to be

$$\frac{1}{\Delta} \frac{\sin \pi(t^o - \tau^o)}{\pi(t^o - \tau^o)} = \int_{-1/2\Delta}^{1/2\Delta} \exp(2\pi j f (t^o - \tau^o)\Delta) df \tag{3.55}$$

(3.55) When both t^o and τ^o are integers, the above integral is zero except for $t^o = \tau^o$, for which case the integral equals $1/\Delta$. Hence, substitution of Eq. (3.55) into Eq. (3.54) results in the elimination of all terms in the summation over τ^o, except the $t^o = \tau^o$ term. The Δ and l/Δ factors cancel. What is left is an identity proving the equality of Eq. (3.54) and demonstrating the validity of the Fourier transform pair for T-discrete signals, Eqs.(3.51) and (3. 52) .

We noted above that $C_{XX}(f)$ is a continuous function. Examination of Eg. (3. 52) also indicates that $C_{xx}(f)$ is a periodic function of frequency, since all the complex exponentials in the summation are periodic with the fundamental frequency being $1/\Delta$. Thls property was to be expected in view of the discussion of allasing in Section 3. 3. Note that only the frequency components between $-1/2\Delta$ and $1/2\Delta$ are needed to describe the signal.

3.10 The periodogram

The intention of this section is to show that the power spectrum of a stationary random process can be estimated through use of the periodogram without having first to estimate the acvf. We will show that the periodogram is equivalent to a Fourier transform of the acvf. To do this we first discuss (1) the properties of an

estimated acvf which is based upon a finite segment of a T- discrete waveform and (2) the properties of an estimated power spectrum which is based upon the Fourier transform of such a specimen acvf.

An estimate of the acvf of a stationary random process can be computed from a T sec segment of the process.A set of N consecutive samples spaced Δ sec apart is used as follows:

$$\hat{c}_{xx}(\tau^o\Delta) = \frac{1}{N}\sum_{t^o=0}^{N-|\tau^o|-1} x(t\Delta)x[(t+\tau^o)\Delta] |\tau^o| \leq N-1 \tag{3.56}$$

(3.56) Note that the upper limit of the summation is a function of τ^o. This is because there are only a finite number of sample products available. For example, in the $\tau^o = 0$ case, all N points can be used to compute the cross products $x(t^o\Delta)x(t^o\Delta)$. In the $\tau^o = 1$ case, only N - 1 points can be used to compute the cross products since, when $t^o = N-1$, the cross product becomes $x[(N-1)\Delta]x(N\Delta)$. The only data samples available are for the time points at 0 through $(N-1)\Delta$. There is no $N\Delta$ time sample available unless, as noted in Section 3. 2, the data are periodicized. This will be discussed further in Section 3.18. Thus the summation over the cross products must be limited to the range of $t^o = 0$ to $t^o = N-2$ when $\tau^o = 1$. Similar reasoning is applicable to larqer magnitudes of τ^o, in which case still fewer sample cross products are available. The expected value of $\hat{c}_{xx}(\tau^o\Delta)$ is

$$\begin{cases} E[\hat{c}_X X(\tau^o\Delta)] & = \frac{1}{N}\sum_{t^o=0}^{N-|\tau^o|-1} E\{x(t^o\Delta)x[(t^o+\tau^o)\Delta]\} \\ & = \frac{1}{N}\sum_{t^o=0}^{N-|\tau^o|-1} c_{xx}(\tau^o\Delta) \\ & = \frac{c_{xx}(\tau^o\Delta)}{N}\sum_{t^o=0}^{N-|\tau^o|-1} 1 \end{cases} \tag{3.57}$$

(3.57)

Since the summation consists of $N - |\tau^o|$ terms, all equal to unity, the sum equals $N - |\tau^o|$ and so Eq. (3.57) becomes

$$E[\hat{c}_X X(\tau^o\Delta)] = (1 - \frac{|\tau^o|}{N})c_{xx}(\tau^o\Delta) \tag{3.58}$$

(3.58) Equation (3.58) indicates that Eq.(3.56) is a biased estimator of the acvf, and that as the number of sample times N becomes large with respect to $|\tau^o|$, the bias becomes small.

An estimate of the power spectrum can then be obtained by using Eq. (3.52) to compute the Fourier transform of the acvf estimate, Eq.(3.56). The summation index $|\tau^o|$ is confined to the range $-(N-1)$ to $(N-1)$ since only N time points are

available in the original sampled data segment and positive and negative values of $|\tau^o|$ are permitted up to N - 1. We then have for the estimate of the power spectrum,

$$\hat{C_x}x(f) = \frac{\Delta}{N} \sum_{\tau^o=-(N-1)}^{N-1} \sum_{t^o=0}^{N-|\tau^o|-1} x(t\Delta)x[(t+\tau^o)\Delta] \exp(-2\pi j f \tau^o \Delta) \quad (3.59)$$

(3.59) This equation forms a basis for estimating the power spectrum although, as will be shown, some modifications are needed so as to obtain statistically acceptable results .

From a practical point of view, evaluation of Eq. (3 .59) can entail relatively large amounts of computer time when N is large. For this reason it may be advantageous to estimate the power spectrum directly by means of the periodogram of the waveform specimen, as expressed by the following equation:

$$P_{xx}(n) = \frac{\Delta}{N} | \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2\pi j n t^o/N)|^2 \quad (3.60)$$

(3.60) $P_{XX}(f)$ is the symbol for the periodogram. It arises from the Fourier transform of the unsampled T sec data segment of x (t),

$$P_{xx}(f) = \frac{1}{T}|X(f)|^2 \quad (3.61)$$

(3.6l) When x(t) is band-limited, we can resort to the sampled representation and the Fourier transform,

$$P_{xx}(f) = \frac{\Delta}{N} | \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2\pi j f t^o \Delta)|^2 \quad (3.62)$$

(3. 62) Usually only the harmonic frequencies $f_n = n/T = n/N\Delta$ are of interest to us (by periodicizing the original data), and we can obtain the periodogram from the DFT;

$$\begin{cases} P_{xx}(f_n) & = \frac{\Delta}{N}|X_N(n)|^2 a \\ & = \frac{\Delta}{N}| \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2\pi j n t^o/N)|^2 \end{cases} \quad (3.63)$$

(3.63) On occasion we shall write $P_{XX}(f_n)$ as $P_{XX}(n)$ so that the two notations are equivalent.

We will now show that $P_{XX}(f)$ is equal to the $\hat{C}_{XX}(f)$ defined in Eg. (3.59). As a first step, we note that the square of the magnitude of a complex quantity is

equal to the product of that quantity and its complex conjugate. Hence,

$$P_{xx}(f_n) = \frac{\Delta}{N} \sum_{t^o=0}^{N-1} x(t^o\Delta) \exp(-2\pi j f_n t^o \Delta) \sum_{u^o=0}^{N-1} x(u^o\Delta) \exp(2\pi j f_n u^o \Delta)$$
$$= \frac{\Delta}{N} \sum_{t^o=0}^{N-1} \sum_{u^o=0}^{N-1} x(t^o\Delta) x(u^o\Delta) \exp[-2\pi j f_n (t^o - u^o)\Delta] \tag{3.64}$$

(3.64) We now make a change of variables, substituting τ^o for $t^o - u^o$. Since both t^o and u^o range from 0 to N - 1, the range will be - (N - 1) to (N - 1). Hence, Eq. (3. 64) becomes

$$P_{xx}(f_n) = \frac{\Delta}{N} \sum_{\tau^o=-(N-1)}^{N-1} \sum_{u^o=0}^{N-|\tau^o|-1} x(u^o\Delta) x[(u^o + \tau^o)\Delta] \exp(-2\pi j f_n \tau^o \Delta) \tag{3.65}$$

(3.65) Note that the upper limit of the summation over u^o has been reduced by $|\tau^o|$. The reasons are the same as for the summation in Eq.(3.56). Comparison of Eq. (3.59) with Eq.(3.65) indicates that the periodogram $P_{XX}(f_n)$ is identical with the spectral estimate $\hat{C}_{XX}(f)$ obtained by means of the Fourier transform of the sample acvf. The reason for preferring the periodogram as the vehicle for spectral estimation is that it can be computed more rapidly, provided that a fast Fourier transform algorithm is used.

3.11 Statistical Errors of the Periodogram–Bias

We previously indicated that the specimen or sample avcf, which is used explicitly in Eq. (3.59) and implicitly in Eq. (3.60), provides a biased estimate of the acvf. Consequently, the periodogram will provide a biased estimate of the power spectrum. The expected value of the periodogram can be obtained by substituting Eq. (3.58), the expected value of the sample acvf, into Eq. (3.65).

$$E[P_{xx}(f_n)] = \Delta \cdots \tag{3.66}$$

(3.66) Comparison of this equation with Eq. (3.52), which defines $C_{xx}(f)$ as the Fourier trans form of $c_{xx}(\tau^o\Delta)$, yields

$$E[P_{xx}(f_n)] = \Delta \cdots \tag{3.67}$$

(3.67)

The three right- most terms in Eq. (3.67) constitute the bias. Assuming that K(t) is a zero mean random process, the bias will tend toward zero as N becomes large.

To examine the nature of the bias in the frequency domain, we can rewrite Eq. 3.66 in a somewhat more general form, as follows :

$$E[P_{xx}(f_n)] = \Delta \sum_{\tau^o=-\infty}^{\infty} w_B(\tau^o \Delta) \cdots \tag{3.68}$$

(3.68)

where

$$w_B(\tau^o \Delta) = \begin{cases} 1 - \frac{|\tau^o}{|} N, \, |\tau^o| < N \\ 0, \, |\tau^o| \geq N \end{cases} \tag{3.69}$$

(3.69)

The function $w_B(\tau^o \Delta)$ can be thought of as a "lag window" function which multiplies or weights the set of acvf terms, and, since it is different from unity, "causes" the periodogram to be a biased estimate of the power spectrum. Since we showed in Chapter 1 that multiplication in the time domain is the equivalent of convolution in the frequency domain, Eq. (3.68) can be stated in the frequency domain as

$$E[P_{xx}(f_n)] = \Delta \int_{-1/2\Delta}^{1/2\Delta} C_{XX}(f) W_B(f_n - f) \, df \tag{3.70}$$

(3.70) where $W_B(f)$ is the Fourier transform of $w(t)$.It can be shown that

$$\begin{cases} w_B(f) & = \sum_{\tau^o=-(N-1)}^{N-1} (1 - \frac{|\tau^o|}{N}) \exp(-2\pi j f \tau^o \Delta) \\ & = \frac{1}{N} (\frac{\sin(\pi N \Delta f)}{\sin(\pi \Delta f)})^2 \end{cases} \tag{3.71}$$

(3. 71) $W_B(f)$ can be thought of as a " frequency window" function. Substituting Eq. (3.71) into Eq. (3.70) gives

$$E[P_{xx}(f_c)] = \Delta \int_{-1/2\Delta}^{1/2\Delta} \frac{1}{N} (\frac{\sin \pi N \Delta (f_c - f)}{\sin \pi \Delta (f_c - f)})^2 C_{xx}(f) \, df \tag{3.72}$$

(3.72)

A plot of $W_B(f)$ is provided in Fig. 3.9a. Note that within the frequency band of interest, $-1/2\Delta$ to $1/2\Delta$, $W_B(f_n - f)$ is near zero except at $f = f_n$. Hence, only that portion of $C_{xx}(f)$ which is near frequency f will contribute much to the periodogram estimate of the power spectrum. This is illustrated in Fig. 3.9b .

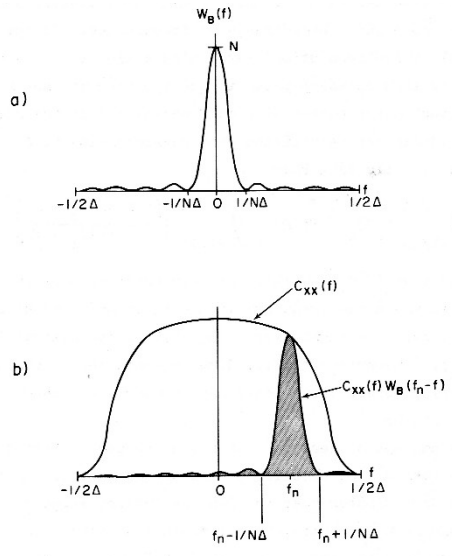

Figure 3.9: Fig. 3.9. (a) Plot of the window function, $W_B(f) = (\sin \pi N \Delta f)^2/N(\sin \pi \Delta f)^2$, for N = 10.(b) An illustration of how the product of $C_{XX}(f)$ and $W_B(f_n - f)$ determines the expected valyue of the estimate of the power spectrum

Note that E(P xlfc)l is equal to an area determined by the product of C (f) and W. (E - f).As N becomes large, the area becomes more closely confined to frequencies that are near to f_n. This means that by increasing N we can obtain a high resolution, small bias estimate whose expected value is close to $C_{xx}(f_n)$. If N is small, the expected value of the estimate will contain spectral components covering a broad range of frequencies. In this circumstance only a low resolution, large bias estimate can be obtained, and nuances such as sharp peaks in the spectrum may not be detected.

A rough guide to the size of N necessary for the bias to become negligible can be obtained from consideration of Eq. (3.72), Fig. 3.9b, and the fact that

$$\Delta \int_{-1/2\Delta}^{1/2\Delta} W_B(f_n - f)\, df = \Delta \int_{-1/2\Delta}^{1/2\Delta} \frac{1}{N} \left(\frac{\sin \pi N \Delta (f_c - f)}{\sin \pi \Delta (f_c - f)}\right)^2 df = 1.0 \quad (3.73)$$

(3.71)

This means that if $C_{xx}(f)$ is relatively constant over the band of frequencies where the frequency window function is markedly greater than zero, then the right side of Eq. 3.72 is approximately equal to $C_{xx}(f_n)$. Inspection of Fig. 3.9b suggests that N should be such that $C_{xx}(f)$ does not vary significantly over a frequency bandwidth of about $4/N\Delta$ Hz.

The concept of leakage that was discussed in Section 3.4 is simply another way of describing bias or resolution. Examination of Fig. 3.9b indicates that the low resolution, large bias situation is one in which activity at frequencies other than the one of interest contributes to (i.e., leaks into) the estimate of $C_{xx}(f_n)$.

3.12 Statistical Errors of the Periodogram–Variance

Since $P_{xx}(f_n)$ is a function of the set of N random variables, the $x(t^o\Delta)$, $P_{xx}(f_n)$ is also a random variable. We already know that as N becomes large, the mean of $P_{xx}(f)$ approaches $C_{xx}(f)$, the power spectrum of x(t). We now arrive at a troublesome property of the periodogram, namely, that its variance does not become small as N increases. Instead, the estimation errors contained in $P_{xx}(f)$ will be of the same order of magnitude as the $P_{xx}(f)$ itself, regardless of N. Consequently, the raw periodogram is not a consistent estimate of the power spectrum $C_{XX}(f)$. For $P_{XX}(f)$ to be a consistent estimate in the statistical sense, its mean must approach the true spectrum and its variance must become small as N becomes large. The periodogram meets the former but not the latter criterion. However, by application of suitable averaging procedures the latter criterion can also be satisfied. We will now discuss the basis of such averaging procedures.

First, to gain insight into the nature of the variance of the periodogram, let us consider the case of a zero mean, white Gaussian process. In this case, the samples $x(t^o\Delta)$ are independent of one anot her and the power of the process is uniformly distributed over the frequency band from $-1/2\Delta$ to $1/2\Delta$. The acvf, $C_{XX}(\tau^o\Delta)$, is zero for all τ^o except $\tau^o = 0$, at which point it has the value σ_x^2. The spectrum $C_{XX}(f)$ of such a process is easily seen to be equal to $\sigma_x^2\Delta$ for all f.

We find the mean and variance of the perlodogram at zero frequency by setting $f = 0$ in Eq. (3.60). This yields

$$P_{xx}(0) = \frac{\Delta}{N}[\sum_{t^o=0}^{N-1} x(t^o\Delta)]^2 \tag{3.74}$$

(3.73) In Section 1.13 it was pointed out that the distribution of the sum of N identically distributed normal (μ, σ) random variables is normal ($N\mu$, $\sqrt{N}\sigma$). Hence, the distribution of $\sigma_{n=0}^{N-1} x(t^o\Delta)$ is Gaussian with a mean of zero and variance equal to $N\sigma_{x^2}$. It was further shown in Section 1.13 that the square of a zero mean, unit-standard-deviation Gaussian random variable has a chi- squared distribution with one degree of freedom. Thus inspect ion of Eq. (3.73) indicates that it can be expressed as the product of a constant times a chi-squared random variate, as follows :

$$P_{XX}(0) = \Delta\sigma_x^2[\frac{1}{\sqrt{N}\sigma_x}\sum_{t^o=0}^{N-1} x(t^o\Delta)]^2 \tag{3.75}$$

(3.74)

The square of the quantity within the brackets has a chi-squared distribution with one degree of freedom. Two other results from Section 1.13 are useful here. The first is that the mean of a chi-squared variable with m degrees of freedom equals m and its variance equals 2m. The second is that the product of a constant a times a chi-squared random variate with m degrees of freedom has a mean equal to am and a variance equal to $2\alpha_n^2$. Applying these to Eq. (3.74), we have

$$E_{xx}[P(0)] = \Delta\sigma_x^2 \tag{3.76}$$

(3.75)

$$var_{xx}[P(0)] = 2\Delta^2\sigma_x^4 \tag{3.77}$$

(3.76) Hence, the standard deviation of $P_{XX}(0)$ is $\sqrt{2}\Delta\sigma_x^2$. Since $C_{XX}(0)$ equals $\Delta\sigma_x^2$, the standard deviation of $P_{XX}(0)$ equals $\sqrt{2}C_{XX}(0)$. This shows that although the expected value of $P_{XX}(0)$ equals $C_{XX}(0)$, the variance of $P_{XX}(0)$ is independent of N, the number of time samples. The coefficient of variation of $P_{XX}(0)$ is $\sqrt{2}$.

The above result applies only to the estimate at zero frequency. A basically similar but computationally more tedious development can be made for the value of a periodograrn of a white, Gaussian process at any arbitrary frequency. The mean and variance of the periodogram are (Oppenheim and Schafer, 1975)

$$E[P_{XX}(f)] = \Delta\sigma_x^2 \tag{3.78}$$

(3.77)

$$var[P_{xx}(f)] = \Delta^2\sigma_x^4[1 + (\frac{\sin 2\pi f\Delta N}{N\sin 2\pi f\Delta})^2] \tag{3.79}$$

(3.78)

Equation (3.78) reduces to Eq. (3.76) when f equals zero or $1/2\Delta$. Since $[\sin 2\pi f\Delta N/N\sin(2\pi f\Delta)]^2$ ranges between zero and one, we find that $var[P_{xx}(f)]$ does not become small, viz.,$\Delta^2\sigma_x^4 \leq var[P_{XX}(f)] \leq 2\Delta^2\sigma_x^4$. In practice it is only necessary to compute the periodogram at the discrete set of frequencies $f = n/N\Delta$, where n is an integer. This makes it possible to use a fast Fourier transform algorithm. At these frequencies, $var[P_{XX}(n/N\Delta)]$ equals $\Delta^2\sigma_x^4$. Hence for a zero mean, white Gaussian process the expected value of the periodogram equals $C_{xx}(f)$ while its standard deviation is also approximately equal to $C_{xx}(f)$. Increasing N will not reduce the standard deviation.

For a nonwhite random process the results are quite similar. The periodogram again provides a biased estimate of $C_{xx}(f)$, as indicated by Eqs. (3.67), (3.68) and (3.72). An approximate expression for the variance of the periodogram is (Oppenheim and Schafer, 1975)

$$var[P_{XX}(f)] \simeq C_{XX}^2(f)[1 + (\frac{sin2\pi f\Delta N}{N\sin 2\pi f\Delta})^2] \tag{3.80}$$

Thus, as in the case of a white Gaussian process, the standard deviation of the periodogram is equal to $C_{xx}(f)$ at frequencies $n/N\Delta$ and is slightly larger at other frequencies. While increasing N will decrease the bias, as indicated by Eq. (3.72), it will not effectively decrease the standard deviation of the periodogram estimate .

3.13 Averaging the Periodogram-the Bartlett Estimator

It is apparent from the preceding discussion that since the periodoqram is not a consistent estimator of the power spectrum, a procedure is required that will attenuate the random fluctuations associated with the periodogram and produce a useful spectrum estimate. One such procedure is to divide the signal specimen into a series of subsegments, compute a periodogram for each of them and then average the

periodoqrams. This approach, first suggested by Bartlett (Oppenheim and Schafer, 1975), also gives one the opportunity of testing for stationarity. It is implemented as follows. Let the signal segment be divided into M subsegments, each Nh sec long. Denote the signal in the m^{th} subsegment by

$$x^{(m)}(t^o\Delta) = x)[t^o + N(m-1)]\Delta), 0 \leq t^o \leq N-1,\ 1 \leq m \leq M \qquad (3.81)$$

(3.80) The corresponding periodogram for the m^{th} subsegment is

$$PP_{XX}^{(m)}(f) = \frac{\Delta}{N}|\sum_{t^o=0}^{N-1} x^{(m)}(t^o\Delta)\exp(-2\pi j f t^o\Delta)|^2 \qquad (3.82)$$

(3.8l) Thus, using Bartlett's method, the estimate of the power spectrum of x(t) is

$$B_{XX}(f) = \frac{1}{M}\sum_{m-1}^{M} P_{XX}^{(m)}(f) \qquad (3.83)$$

(3.82) The expected value of the Bartlett estimator at frequency f_n is

$$E[B_{XX}(f_n)] = \frac{1}{M}\sum_{m=1}^{M} E[P_{XX}^{(m)}(f_n)] \qquad (3.84)$$

(3.83) The expected value of the periodogram, $P_{XX}^{(m)}(f_n)$, is the same for all m and is given by Eg. (3 .72). Hence, the expected value of the Bartlett estimator is the same as the expected value of the individual periodograms and is given by

$$E[B_{XX}(f_n)] = \Delta\int_{-1/2\Delta}^{1/2\Delta}\frac{1}{N}[\frac{\sin\pi N\Delta(f_n) - f)}{\sin\pi\Delta(f_n) - f)}C_{XX}(f)\,df \qquad (3.85)$$

(3.84) The bias leakage properties of the Bartlett estimator are also the same as that of the individual periodograms so that the remarks following Eg. (3.72) concerning bias and leakage of raw periodograms apply here as well. What is most important is that the variance of the Bartlett estimator is less than that of the periodogram, as we shall show in the next section. The argument is based upon the assumption that there is a total of $N_m M$ data points available, N_m being the number of data points in each of the M subsegments.

3.14 Variance of the Bartlett Estimator

If N_m, the number of time points in a subsegment, is sufficiently large so that $cxx(\tau^O\Delta)$ is small for $\tau^o > N_m$, then the various subsegment periodograms, $P_{xx}^{(m)}(f)$, will tend to be statistically independent of one another. This means that the variance of the average of the M periodograms will be approximately equal to the variance of the individual periodograms divided by M. (See Sections 1. 13 and 4.1.) Using this and Eg. (3.79), it follows that the variance of the Bartlett estimator is approximately

$$var[B_{XX}(f)] = var[P_{XX}^{(m)}(f)]/M = \frac{C_{XX}^2(f)}{M}[1 + (\frac{\sin 2\pi f\Delta N_m}{N_m \sin 2\pi f\Delta})^2] \quad (3.86)$$

(3.85)

Thus, for $f_n = n/N_m\Delta$ and unequal to zero or $1/2\Delta$,

$$var[B_{XX}(f_n)] \simeq C_{XX}^2(f_n)/M \quad (3.87)$$

(3.86a) and when f = 0 or $1/2\Delta$,

$$var[B_{XX}(0)] = 2C_{XX}^2(0)/M \quad (3.88)$$

$$var[B_{XX}(1/2\Delta)] = 2C_{XX}^2(1/2\Delta)/M \quad (3.89)$$

(3.86b) (3.86c) This means that Bartlett's method is a consistent estimator of the power spectrum since, as the total number of data points $N = N_mM$ increases, both the bias and variance of the estimate become small. The bias, as given by Eq. (3.84), is determined solely by the length of the subsegments N_m and diminishes as N_m increases. The variance is determined by the number of subsegments M to which it is inversely proportional.

Since only a fixed number of time samples N_mM is available for estimation of the power spectrum, however it is done, there is a trade-off between the size of the variance and the resolution of the Bartlett estimator. Variance is reduced by dividing the data segment into as many subsegments as possible, thereby increasing M.But by so doing, one shortens the length of the subsegments N_m, and hence increases the bias and decreases the resolution. Thus, the size of variance and bias are inversely related to one another: as one increases, the other decreases. Variance itself is related closely to spectral resolution, the ability to detect fine structure in the spectrum. Decreasing the variance of an estimate is brought about by decreasing the length Nm of a data subsegment. This means that the periodograms have fewer frequency components in them (smaller N_m) so that the frequency resolution decreases. Reduced resolution is therefore concomitant with reduced variance. Later

we shall show this is another way by speaking of frequency resolution in terms of bandwidth.

3.15 Fast Fourier Transform and Power Spectrum Estimation

We mentioned in Section 3.8 that the main reason for using the periodogram approach to power spectrum estimation is that it can be carried out more rapidly than by computing the acvf and then taking its Fourier transform. The savings in time come about by use of the fast Fourier transform algorithm (Bergland, 1969; Oppenheim and Schafer, 1975) to compute the Fourier transforms of the original data, as specified by Eqs. (3.60) and (3.81). In order to take advantage of the fast Fourier transform or FFT, we must confine the frequencies for which the spectral estimate is computed to the discrete set of $f_n = n/N\Delta, n = 0, \ldots, N-1$, where $N\Delta$ is the duration of the segment. This is no restriction since the periodogram of a band-limited process is completely represented by its sample values at frequencies $n/N\Delta$. We must emphasize the fact that the value of the spectral estimate at each frequency does not depend upon whether the FFT or some other algorithm is used. Neither are the bias and variance of the estimate affected by the choice of the algorithm. The only difference may be in computational round-off error, which may be smaller with the FFT, since the FFT entails fewer steps.

3.16 Smoothing of Spectral Estimates by Windowing

We have shown above that although the periodogram itself is not a consistent estimator of the power spectrum, a way of obtaining one is to average across a set of sequentially obtained periodograms. Here we shall develop a different approach to a smoothing of the periodogram Which also yields a consistent spectral estimate. Our argument will apply mainly to estimates obtained at the discrete set of frequencies $f_n = n/N\Delta$.

Rather than dividing the data into numerous time sequential subsegments and averaging across time, the periodogram can be smoothed by averaging over narrow bands of frequency. One important property of periodogram estimates that we make use of here is that $P_{XX}(f_n)$ for a white Gaussian process is the sum of the square of two identical and independent Gaussian random variables (Jenkins and Watts, 1968), except when $f = 0, 1/2\Delta$. This property is also approximately valid when the Gaussian restriction is eliminated and any peak in the spectrum is broad compared to $l/N\Delta$. This means that in most situations of interest, $P_{XX}(f_n)$ is

proportional to a chi- squared random variable with two degrees of freedom. Since

$$E[P_{XX}(f_n)] = C_{XX}(f_n) \tag{3.90}$$

and

$$var[P_{XX}(f_n)] = C_{XX}^2(f_n) \tag{3.91}$$

$2P_{XX}(f_n)/C_{XX}(f_n)$ is a χ_2^2 random variable. A second important property of periodogram estimates is that for a white Gaussian process $cov[P_{XX}(f_n), P_{XX}(f_m)] = 0$ when $n \neq m$. This property is also approximately valid for nonwhite and some non-Gaussian processes. Thus one can treat values of the periodogram at integer multiples of $l/N\Delta$ as uncorrelated random variables.For more details, see Jenkins and Watts (1968).

Let us now consider a spectral estimate made up of a weighted sum of periodogram values;

$$\hat{C}_{XX}(f_n) = \sum_{k=n-K}^{n+K} P_{XX}(f_k)W(f_n - f_k) \tag{3.92}$$

(3.87) The $W(f_k)$ are the weights of a spectral smoothing filter which weights and sums the periodogram estimates from f_{n-K} to f_{n+K}. $\hat{C}_{XX}(f_n)$ is a new random variable, and when the process is Gaussian, its mean and variance are given by

$$E[\hat{C}_X X(f_n)] = \sum_{k=n-K}^{n+K} E[P_{XX}(f_k)]W(f_n - f_k) \tag{3.93}$$

(3.88a)

$$var[\hat{C}_X X(f_n)] = \sum_{k=n-K}^{n+K} var[P_{XX}(f_k)]W^2(f_n - f_k) \tag{3.94}$$

(3.88b) Since frequency averaging is usually applied to periodograms obtained from long data segments, the results of Section 3.12 indicate that Eqs. (3.88a and b) can be approximated by

$$E[\hat{C}_X X(f_n)] = \sum_{k=n-K}^{n+K} C_{XX}(f_k)W(f_n - f_k) \tag{3.95}$$

(3.89a)

$$var[\hat{C}_X X(f_n)] = \sum_{k=n-K}^{n+K} C_{XX}^2(f_k)W^2(f_n - f_k) \tag{3.96}$$

(3.89b)

These equations can be further simplified when the process is a white one (even if only in the range of frequencies covered by the summation), in which case its mean and variance are given by

$$E[\hat{C}_X X(f_n)] = C_{XX}(f_n) \sum_{k=-K}^{K} W(f_k) \tag{3.97}$$

(3.90a)

$$var[\hat{C}_X X(f_n)] = C_{XX}^2(f_n) \sum_{k=-K}^{K} W^2(f_k) \tag{3.98}$$

(3.90b) It is convenient to use only positive weights and to set $\sum_k W(f_k) = 1$. This results in no loss of generality. Since each weight must be no larger than unity, the variance of $\hat{C}_{XX}(f_n)$ must be less than the variance of $P_{XX}(f_n)$. A rectangular filter, one which weighs equally all the periodoqram values from f_{n-K} to f_{n+K} has weights $W(f_k) = 1/(2K+1)$. For a white noise process, the variance of Cxx (fn) with such a filter is 1/(2K + 1) that of $P_{XX}(f_n)$.

Because $\hat{C}_{XX}(f_n)$ is the weighted sum of a set of $P_{XX}(f_k)$ and each $P_{XX}(f_k)$ is closely proportional to a χ_2^2 random variable, $\hat{C}_{XX}(f_n)$ is itself closely proportional to a $\chi_{d.f.}^2$ random variable and can be dealt with in this way. This was discussed earlier in Section 1.13.The degrees of freedom d.f., and the constant proportionality α for the random variable are given by

$$d.f. = \frac{2(E[\hat{C}_{XX}(f_n)])^2}{var[\hat{C}_{XX}(f_n)]} \simeq \frac{2}{\sum_{k=-K}^{K} W^2(f_k)} \tag{3.99}$$

(3.91a)

$$\alpha = \frac{E[\hat{C}_{XX}(f_n)]}{d.f.} \simeq \frac{C_{XX}(f_n)}{d.f.} \tag{3.100}$$

(3.9lb) This means that we can consider $d.f.[\hat{C}_{XX}(f_n)/C_{XX}(f_n)]$ to be a $\chi_{d.f.}^2$ random variable. Applying this result to a rectangular smoothing filter, one which weights equally the periodogram values from f_{n-K} to $f_{n+K'}$ we find $d.f. = 2(2K+1)$. This was to be expected since $2K+1$ periodogram values, each with 2 degrees of freedom, were used to construct the estimate.Computations of this sort can be carried out for any smoothing window of interest. For example, a Bartlett estimator which is obtained by sectioning an N sample record into M segments, each of length N/M, can be shown to have 3M degrees of freedom.

Equation (3.91a) shows that the degrees of freedom and the variance of the estimator are inversely related.The equation also bears a close relationship to the

number of frequency components being summed over: the greater the number, the greater the degrees of freedom and the smaller the variance. We may assign to the smoothing filter a generalized bandwidth parameter. This is the bandwidth (or the number of frequency components averaged over) that a uniformly weighted filter would have in order to yield an estimator with the same variance as the actual smoothing filter . This assumes the data have a flat spectrum over the range of the smoothing filter. The bandwidth and variance are inversely related so that their product is a constant. This can be readily seen for a white noise process being smoothed by a uniformly weighted filter. The bandwidth. 2X + 1 and the variance $var[P_{xx}(f_n)/(2K+1)$. This means that there is always a trade-off between variance and bandwidth. Small variance is obtained at the cost o f large bandwidth (or low resolution) and vice versa.

The trade-offs between variance and resolution are much the same whether a Bartlett estimator, Eq. (3.82), or a more general windowing approach, Eg. (3.87), is used. However, there are differences in details.Inspection of Eq. (3.84) indicates that the Bartlett estimator is the equivalent of using a frequency smoothing filter of the form $(sin\pi N\Delta f/Nsin\pi\Delta f)$, referred to as the Bartlett window. While the Bartlett window has been widely used and provides a reasonable balance between variance and resolution, in some instances other window shapes may be more desirable.An advantage of the averaging over the frequency approach is that a wide variety of window functions can be devised according to the particular spectral smoothing problem at hand. Details such as the precise width of the window function can be controlled by direct specification of the $W(f_k)$ terms.

Although there is some latitude in selecting a spectral window function $W(f_k)$ for a given application, there are practical constraints that should be evaluated. Thus, while a window that extends over a broad frequency range will yield a low variance estimate, it is associated with leakage from frequencies that are far from the one at which the spectrum is being estimated. If these distant spectral components are large, the window width should be narrowed to reduce the leakage. Another consideration has to do with the values of the $W(f_k)$. There are relatively common window functions that have negative values for some of the $W(f_k)$. Such windows must be used with caution since they can lead to negative spectrum estimates .

From Eqs. (3.90b) and (3.91a) it can be seen that the magnitude of $\sum_{k=-K}^{K} W^2(f_k)$ is crucial in determining the variance of the spectrum estimate. The smaller the sum of the squares, the smaller the variance. Given the constraint that $\sum_{k=-K}^{K} W(f_k) = 1$, it can be shown that the sum of the $W^2(f_k)$ terms will be smallest when all $W(f_k) = 1/(2M+1)$ in which case the sum of the squares equals $1/(2M+l)$.

Spectral windowing can also be implemented in the time domain by dealing with the acvf. Since convolution in the frequency domain is equivalent to multipli-

cation in the time domain, the time domain equivalent of Eq.(3. 87) is

$$C_{XX}(f) = \Delta \sum_{t^o=-(N-1)}^{N-1} w(t^o\Delta)\hat{c}_{XX}(t^o\Delta)\exp(-2\pi j f t^o\Delta) \qquad (3.101)$$

(3.92) where $\hat{c}_{xx}(t^o\Delta)$ is given by Eq. (3.56), and $w(t^o\Delta)$, commonly referred to as a "lag window," is in effect the Fourier transform of w(i). Prior to the late 1960's, when the FFT became widely known, windowing was usually implemented in the time domain, via Eq. (3.92). It is the Fourier transform of these lag windows that sometimes yields negative $W(f_k)$. A detailed discussion of the properties of spectral and lag window functions and their implementation can be found in Jenkins and Watts (1968), Otnes and Enochson (1972), and Welch (1967).

3.17 The Cross Spectrum

In Chapter 1 we discussed the concept of the cross covariance function (ccvf).The Fourier transform of the ccvf is referred to as the cross spectrum. The cross spectrum provides a statement of how common activity between two processes is distributed across frequency. The cross spectrum is the Fourier transform bf the ccvf, as indicated by Eq. (l.69). As an example, consider two processes each of which consists of a quasiperiodic signal embedded in wide band noise processes. Suppose the quasiperiodic signals are due to a common phenomenon so that they are closely related. The wide band noise processes, on the other hand, are due to random fluctuations that are unique to each process and so are unrelated. The cross spectrum of the two processes would be relatively large in the frequency band of the shared, quasiperiodic signal and small at other frequencies, since the wide band noise processes are independent and not shared activity.

To some extent the cross spectrum can provide insight into the relationships between a pair of random processes. Further insight can be obtained from the coherence function, which is derived from the power spectra and cross spectrum of the pair of random processes. The coherence function will be discussed in Section 3.19.

The procedures and problems in estimating cross spectra are similar to those described in the preceding discussion of the power spectra. It can be computed by Fourier transform of the sample ccvf. However, with the availability of the FFT algorithm, a periodogram approach in some instances may be preferable. The bias-resolution and variance properties of the cross spectrum are the same for both approaches and are similar to those of the power spectrum.

For example, consider the ccvf and cross spectrum for two wide sense stationary random signals, $x(t)$ and $y(t)$.The sample ccvf may be computed i n the same

manner as an acvf [see Eq. (3.56)], as follows,

$$\hat{c}_{xy}(\tau^o\Delta) = \frac{1}{N}\sum_{t^o=0}^{N-|\tau^o|-1} x(t^o\Delta)y[(t^o+\tau^o)\Delta], \qquad |\tau^o| \leq N-1 \tag{3.102}$$

(3.93)

The sample cross spectrum can be obtained in the same manner as the sample power spectrum [see Eq. (3.59)], as follows,

$$\hat{C}_{xy}(f) = \Delta \sum_{\tau^o=-(N-1)}^{N-1} \hat{c}_{xy}(\tau^o\Delta)\exp(-2\pi j f\tau^o\Delta) \tag{3.103}$$

(3.94)

The expected value of the above cross- spectrum estimate can be found by the same steps used to arrive at Eq .(3.72), the expected value of the periodogram estimate. The result is

$$E[\hat{C}_{xy}(f_n)] = \Delta\int_{-1/2\Delta}^{1/2\Delta} \frac{1}{N}\left(\frac{\sin\pi NT(f_n-f)}{\sin\pi T(f_n-f)}\right)^2 C_{xy}(f)df \tag{3.104}$$

(3.95)

Eq.(3.95) is directly comparable to Eq.(3.72), the expression for the expected value of the periodogram estimate of the power spectrum. As in the case of the periodogram, increasing the length of the epoch segment N will decrease the bias of the cross spectral estimate but its variance will not be effectively decreased. Consequently, averaging and/or windowing techniques, as described earlier for estimation of the power spectrum, must also be employed when estimating the cross power spectrum. Further details about cross- spectral estimates may be found in chapters 8 and 9 of Jenkins and Watts (1968).

3.18 Covariance Functions

The auto- and cross covariance functions were introduced in Chapter 1 and shown to be a way of representing the temporal relationships within an individual dynamic process and also between different dynamic processes. The Fourier relationship between the cvfs a nd power spectra was also established for continuous stationary processes and for T sec realizations of them. To do this for the power spectra we re sorted to the artifice of considering a T sec segment of data to be one period of a periodic process. This provided us with an estimator for the cvf and the spectrum of the continuous aperiodic process. The properties of the spectral estimators

have been discussed in the preceding section.Now we move to a more detailed consideration of the covariance function, pointing out some essential features of its estimation and how this estimation is related to power spectrum estimation. - We begin with the autocovariance function.

3.18.1 A. Some Statistical Properties of the ACVF Estimator

The representation of a T sec segment of data as one period of a periodicized specimen function $x(t)$ means that the estimated acvf is given by

$$\tilde{c}_{xx,N}(\tau^o) = \frac{1}{N}\sum_{t=0}^{N-1} \tilde{x}(t^o)\tilde{x}^*(t^o+\tau^o) \tag{3.105}$$

(3.96) and is itself periodic, N. We use the tilde to denote that the acvf has arisen from periodicized data $\tilde{x}(t)$. The subscript N indicates the periodicity. (Throughout this discussion we will assume the original specimen function to be band limited, $F = 1/2$, and sampled at the Nyquist rate so that $\Delta = 1$.) Whenever $t^o + \tau^O$ exceeds N - 1, $t^o + \tau^O$ is to be considered as having its value taken "modulo N." That means, in this instance, that if $t^o + \tau^O = 117$ and N = 100, the value taken for $t^o + \tau^O$ is 17 and $\tilde{x}(117) = x(17)$. This follows from the periodicity of $\tilde{x}(t^o)$. The estimated acvf that results from the use of Eq. (3.96) is sometimes referred to as a circular covariance function because of this method of computation–the data are in effect considered to be wrapped around a cylinder whose circumference is $T = N\Delta$. The circular covariance function estimator has a serious deficiency that limits its usefulness. The nature of this deficiency can be seen by representing it as two summations:

$$\tilde{c}_{xx}(\tau) = \frac{1}{N}[\sum_{t^o=0}^{N-1-|\tau^o|} \tilde{x}(t^o)\tilde{(x)}^*(t^o+\tau^o) + \sum_{t^o=N-|\tau|}^{N-1} \tilde{x}(t^o)\tilde{x}^*(t^o+\tau^o-N)] \tag{3.106}$$

(3.97) The absolute value sign serves to make the equation applicable to both positive and negative delays, though from the symmetry of the acvf about $\tau = 0$, only positive values need be considered. Using this fact, it can be seen that the above equation simplifies to

$$\tilde{c}_{xx,N}(\tau^o) = (\frac{N-|\tau^o|}{N})\tilde{c}_{XX}(\tau^o) + \frac{|\tau^o|}{N}\hat{C}(N-|\tau^o|) \tag{3.107}$$

(3.98)

$\tilde{c}_{XX}(\tau^o)$, of course, is just the average of products of the form $\tilde{x}(t^o)\tilde{(x)}^*(t^o + \tau^o)$. This means that the circular acvf estimator is a combination of two estimators

of the acvf, one for τ^o to and the other for $N - |\tau^o|$. These two are inseparable from one another in this method of estimation. Interpretation of the estimated acvf can therefore be a problem. Of course, this is of no consequence when the data really do arise from a process with period T. However, this is not usually the case. Consequently, it is desirable to look for acvf estimation procedures that are free of this problem. We need not seek far for one. All we need do is adopt another periodicity artifice, one that begins by padding out the original sequence of N samples with a sequence of samples of 0 amplitude, let us say L of them. Then the data may be considered to arise from a specimen of a periodic process whose period is N' =N + L. We consider this for the simplest situation, when L = N and N' = 2N.

In our new sequence, $\tilde{x}(t^o)$ of length 2N, data samples $\tilde{x}(N)$ through $\tilde{x}(2N-1)$ are O. Because of this, at each time lag τ^o there can be only $N - |\tau^o|$ nonzero products in the acvf estimate formed from the sequence. The acvf is then estimated as the average of these products with, however, the averaging factor being taken as l/N, N being the number of nonzero products when $\tau^o = 0$ rather than $l/(N-|\tau^o|)$. The reason for using the former is that the variance of the resulting estimator turns out to be smaller at larger values of τ^o than when using the factor $l/(N - |\tau^o|)$ (See Jenkins and Watts, 1968.) This gives for the estimator

$$\tilde{c}_{xx,2N}(\tau^o) = \frac{1}{2N} \sum_{t^o=0}^{2N-1} \tilde{x}(t^o)\tilde{(x)}^*(t^o+\tau^o) = \sum_{t^o=0}^{N-|\tau^o|-1} x(t^o)x^*(t^o+\tau^o) = \hat{c}_{XX}(\tau^o) \tag{3.108}$$

(3.99)

The tilde over the data samples is unnecessary. We have also returned to the circumflex notation for the acvf estimate because circularity has been eliminated in the computation even though we have arrived at $\hat{c}_{XX}(\tau^o)$ by an argument involving a periodicity of 2N. This estimate does not have the difficulty exhibited by the circul ar acvf estimate with period N as given in Eq. (3.96). It is therefore to be preferred to $\tilde{c}_{XX}(\tau^o)$ in most instances.

The statistical properties of $\hat{c}_{XX}(\tau^o)$ are of interest. When the data $x(t)$ arise from a specimen function of random process x, we have

$$E[\hat{c}_{xx}(\tau^o)] = \frac{1}{N} \sum_{t^o=0}^{N-|\tau^o|-1} E[x(t^o)x^*(t^o + \tau^o)] \tag{3.109}$$

(3 .100) This means that $\hat{c}_{xx}(\tau^o)$ is a biased estimate of $c_{xx}(\tau^o)$ because, as shown earlier, $E[\hat{c}_{xx}(\tau^o)] - c_{xx}(\tau^o) = -|\tau^o|c_{xx}(\tau^o)/N$. Use of the averaging factor $1/(N-|\tau^o|)$ would eliminate this problem, but only, as noted above, at the expense

of increasing the vari- ance of the estimate as τ^o becomes large. This is generally thought to be undesirable.

The variance of $\hat{c}_{XX}(\tau^o)$ may be calculated from its definition in Eq. (3.99). The result depends upon the statistical properties of the process.In the Gaussian case, the one of most general interest, it can be shown (Jenkins and Watts, 1968) that

$$var[\hat{c}_{xx}(\tau^o)] \simeq \frac{1}{N} \sum_{k^o=-\infty}^{\infty} [c_{xx}^2(k^o) + c_{xx}(k^o+\tau^o)c_{xx}(k^o-\tau^o)] \qquad (3.110)$$

(3.101)

This means that the variance of the acvf estimate of a Gaussian process depends upon the acvf itself, something we generally do not know beforehand. For the particular situation in which the process is white noise with variance σ_x^2, $c_{xx}(\tau^o) = \sigma_x^2\delta(\tau^o)$ and $var[\hat{c}_{xx}(\tau^o)] = \sigma_x^4/N$ for all τ^o except $\tau^o = 0$ in which case the variance is $2\sigma_x^4/N$. Note that when x is an aperiodic process with no dc component, $c_{xx}^2(\tau^o)$ becomes small as τ^o becomes large. This means that the summation on the right-hand side of Eq. (3.101) will be finite so that when we divide it by N to obtain the variance of $C_{xx}(\tau^o)$, the result becomes small as N increases, indicating the estimator to be a consistent one. This also can be shown to hold when the process is non-Gaussian. Further scrutiny of Eq. (3.101) seems to indicate that difficulties are encountered when $x(t)$ has a periodic component in it, which can occur when there is residual interference from 60 Hz power lines. In this case, $c_{xx}^2(\tau^o)$ does not become small as τ^o increases and the summation becomes infinite. Does this mean that the variance of the estimate is infinite regardless of N? The answer is no. The difficulty arises in the formulation leading to Eq. (3.101). When proper account is taken of the pure frequency component in x(t), the variance of the estimate turns out to be the same as before.

The statistical relationship between estimates of the acvf made at neighboring time points is also of some interest. This refers to the fluctuations of the estimate about the estimated mean of the acvf. What we are in effect discussing is the covariance of the estimation errors. The problem is a thorny one, but some results exist for the Gaussian stationary process. In particular, the covariance between acvf estimates at T1and T2is given by (Jenkins and Watts, 1968)

$$cov[\hat{c}_{xx}(\tau_1^o), \hat{c}_{xx}(\tau_2^o)] \simeq \frac{1}{N} \sum_{r^o=-\infty}^{\infty} [c_{xx}(r^o)c_{xx}(r^o+\tau_1^o-\tau_2^o)+c_{xx}(r^o+\tau_1^o)c_{xx}(r^o-\tau_2^o) \qquad (3.111)$$

(3.102) This equation, from which the previous one was derived, points out some useful features of the acvf estimate. First, the estimates are uncorrelated only when

the x process is a white noise with $c_{xx}(\tau^o) = \sigma_x^2(\tau^0)$ Second, for any process which has an acvf with nonzero values extending over K successive intervals, there will be a nonzero covariance between acvf estimates that are closer than 2K apart, that is, for which $|\tau_1^o - \tau_2^o| < 2K$. Narrow band processes have covariance functions of this type. The covariance between estimates becomes smaller as $|\tau_1^o - \tau_2^o$ approaches 2K. But the major fact is that when the process is a narrow band one, a larger N is required to obtain an acvf estimate in which the covariance between estimates is to be kept beneath a given maximum. This can be of importance in dealing with acvf estimates of the EEG. An EEG with a marked alpha component will, for a fixed N, have a greater amount of covariance between acvf estimates than will an estimate of the covariance function obtained when the alpha component is small or lacking. Another aspect of the covariance function of narrow band processes is that there is little, if anything, to be gained by smoothing the acvf estimates because this does not reduce the covariance between neighboring estimates.

3.18.2 B. Estimation of the ACVF

The functional form of the estimator in Eq. (3.99) suggests the obvious "brute force" way of calculating the estimates: averaging for each value of τ^o the $N - \tau^o$ products obtained from the N samples sequence. Computationally, the procedure is a lengthy one since complete evaluation of $\hat{c}_{xx}(\tau^o)$ requires that there be $N(N+1)/2$ multiplications and N(N - 1)/2 additions, a total of N^2 arithmetic operations. When N is large, the time required to complete this task becomes excessive. While some short cuts have been found for these time domain procedures, the net time savings has not been impressive. What has brought about a significant reduction in computation time has been the fast Fourier transform algorithm. Its use makes it possible to obtain estimates of the acvf by first estimating the periodogram of the data and then taking the inverse discrete Fourier transform. Since there are about $Nlog_2N$ operations involved in estimating the periodogram and about another $2Nlog_2(2N)$ in taking the inverse DFT, the great computational savings are apparent. For example, when N = 1000, the method of Eq. (3.99) requires about 10^6 operations, while the DFT method requires about $4x10^4$ operations. The reduction in the number of operations is by a factor of over 25, a factor that increases as N increases. Because the OFT is such an efficient approach to acvf estimation when N is large, we shall describe it further.

We have already noted that the acvf estimate of Eq. (3.99) can be considered to arise from a periodicized process whose initial N samples are the $x(t^o\Delta)$ and whose final N samples are all zeros. To guard against spectral leakage effects of the dc component, we subtract out the average value of the N samples before padding the sequence with zeros. We may also de-trend the data if that seems warranted.

The resulting sequence of 2N points then possesses the avcf we are interested in. An alternative way of arriving at this acvf is to first obtain the periodograrn of the padded sequence. The periodograrn of an unpadded sequence of N data points has been given in Section 3.10, Eq. (3.60). When the sequence is padded to length N' by adding L consecutive zeros such that N' = N + L, the periodogram of the padded sequence is

$$P_{xx,N'}(n) = \frac{1}{N'}|X_{N'}(n)|^2 = \frac{1}{N'}|\sum_{t^o=0}^{N-1} x(t^o)\exp(-2\pi jnt^o/N')|^2 \qquad (3.112)$$

(3.103) The upper limit in the summation is N - 1 rather than N' - 1 because the last L values of the sequence are zero. When N' = 2N, we have

$$P_{xx,2N}(n) = \frac{1}{2N}|\sum_{t^o=0}^{N-1} x(t^o)\exp(-2\pi jnt^o/2N)|^2 \qquad (3.113)$$

(3.104) Notice that because the fundamental interval is 2N rather than N in length, there are twice as many frequencies present in the 2N periodogram. These additional frequency components are required to express the fact that the second hal f of the sampl e sequence is constrained to be zero. They afford no additional information about the ori ginal data but only serve as a computational vehicle to arrive at the acvf estimate. Note also that the presence of 2N rather than N in the denominator does not increase the number of operations involved in the computation.

Having once obtained $P_{xx,2N}(n)$, its inverse DFT can be taken and it yields the estimated acvf:

$$\hat{c}_{xx}(\tau^o) = \frac{1}{N}\sum_{n=-(N-1)}^{N-1} P_{xx,2N}(n)\exp(2\pi jn\tau^o/2N) \qquad (3.114)$$

(3.105) Use of the factor l/N rather than l/2N in the above equation might, at first glance, appear to be an error. It can be verified to be correct by taking the DFT of the padded sequence $\tilde{x}(t^o)$ and substituting this into Eq. (3.99, top). After carrying out the summations and using Eq. (3.103), we arrive at Eq. (3.105). Furthermore, because $c_{xx}(\tau^o)$ is an even function, the computation need only be carried out for positive values of τ^o. Another way of writing Eq.(3.105) takes advantage of the fact that $P_{xx,2N}(n)$ is real. Using this, we have

$$\hat{c}_{xx}(\tau^o) = \frac{P_{xx,2N}(0)}{N} + \frac{2}{N}\sum_{n=1}^{N-1} P_{xx,2N}(n)\cos(2\pi n\tau^o/2N) \qquad (3.115)$$

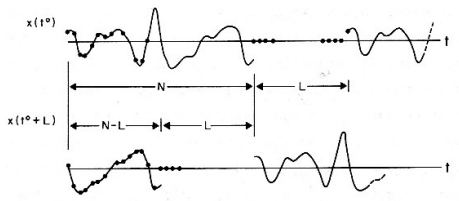

Figure 3.10: Fig. 3.10. Computation of the estimated avcf $\hat{C}_{xx}(t^o)$ at lag L from a periodicized sequence of N data points padded with L zeros.Only N - L products can differ from zero .

(3.l06)

The derivation of the estimated acvf from the periodogram has just been shown to be valid for all values of τ^o up to N. In practice, there is usually little need to carry this out to such large lag values.Usually, lags that are less than 10% of N are only of interest. Because of this there are further savings to be obtained in the use of the DFT. Let us assume that the acvf is of interest up to a lag of $L < N$. Then when we pad the original sequences of N data samples, we need to add L zeros to get an overall sequence of length $N' = N + L$. This guarantees that any estimation of $c_{xx}(\tau^o)$ at values of $\tau^o < L$ will be free from wraparound or overlap effects with the next period of the periodicized data. The effect of padding the data with L zeros is shown in Fig. 3.10 when the lag is L. It can be seen that there are N - L products which are nonzero and L which are forced to be zero.and that none of the nonzero products arises from the overlap of one period with the next.

The N' periodogram of the padded data is given (after average values and possible trends have been removed) by Eq. (3 .104) rewritten here

$$P_{xx,N'}(n) = \frac{1}{N'} \left| \sum_{t^o=0}^{N-1} x(t^o) \exp(-2\pi j n t^o / N') \right|^2, -(N'-1) \leq n \leq (N'-1) \tag{3.116}$$

(3.107) where N' - N + L. As before, we have a larger range of n to deal with, but the additional frequency terms in the periodogram only serve to take the padding with zeros into account. The inverse DFT then yields our estimate of the acvf:

$$\hat{c}_{xx}(\tau^o) = \frac{1}{N'} \sum_{n=-(N'-1)}^{N'-1} P_{xx,N'}(n) \exp(-2\pi j n \tau^o / N'), 0 \leq |\tau^o| \leq L \tag{3.117}$$

(3 .108)

Because L is usually small compared to N, the inverse transform involves not many more operations than does the computation of the periodogram.

3.18.3 C. Cross Covariance Function Estimation

The computation of the ccvf for two N length data sequences $x(t\Delta$ and $y(t\Delta)$ follows the same principles that hold for the acvf estimate. Again, we assume $\Delta = 1$. The use of the direct and inverse DFT facilitates these computations when N is large. If we are interested in estimating the ccvf for lags up to L, then padding the x and y sequences with L zeros each eliminates the possibility of an overlap in the computation. The procedure to be used, therefore, after padding the sequences, is to obtain their respective DFTs, $X_{N'}(n)$ and $Y_{N'}$. From them we obtain the raw crossspectrum estimate $P_{xy,N'}(n) = 1/N\, X_{N'}(n)Y_{N'}(n)$ and then the estimated ccvf:

$$\hat{c}_{xy}(\tau^o) = \frac{1}{N'} \sum_{n=-(N'-1)}^{N'-1} P_{xy,N'}(n) \exp(2\pi j n\tau^o/N'), -L \leq \tau^o \leq L \quad (3.118)$$

(3.109)

It will be remembered that $C_{xy}(\tau^o)$ is not an even function of τ^o and so its ccvf is to be estimated at both positive and negative values of τ^o. This means a doubling of the length of the last step of the computation, but when N is large, the FFT still produces a substantially shorter computation than the -brute force method.

The statistical properties of the ccvf are close enough to those of the acvf so that a full development of them would in the main be repetitious. Consequently, we bring out only the highlights of the development and move quickly to the results. The most common form of the ccvf estimator is the biased version

$$\hat{c}_{xy}(\tau^o) = \frac{1}{N} \sum_{t^o=0}^{N-|\tau^o|-1} x(t^o)y^*(t^o + \tau^o) \quad (3.119)$$

(3.110) $\hat{c}_{xy}(\tau^o)$ can be considered to be one period of a 2N periodic function, and, as already shown, this is especially important when it is obtained by Fourier methods. The biased version of the estimator is preferred for the same reason as is the biased version of the acvf, that it tends to yield a smaller variance in the estimate when TO becomes large. The variance of the ccvf estimator is derivable from its definition. When both processes are Gaussian, it is given by (Jenkins and

Watts, 1968),

$$var[\hat{c}_{xy}(\tau^o)] \simeq \frac{1}{N} \sum_{r^o=-\infty}^{\infty} [c_{xx}(r^o)c_{yy}(r^o) + c_{xy}(r^o + \tau^o)c_{yx}(rO - \tau^o)] \quad (3.120)$$

(3.111) This shows that the variance is calculable only when we know what the ccvf and both acvfs are. If both processes are white and uncorrelated, the second term drops out and we have

$$var[\tilde{c}_{xy}(\tau^o)] = \frac{\sigma_x^2 \sigma_y^2}{N} \quad (3.121)$$

(3.112) The principal fact about the ccvf estimator is that it is a consistent one. Also in common with the acvf estimator, the covariance between estimates at two different lag times depends upon the difference between the lags and the covariance properties of the processes. The covariance of the estimator is a generalization of Eq. (3.111) which we show here for the special case when x and y are uncorrelated:

$$cov[\hat{c}_{xy}(\tau_1^o), \hat{c}_{xy}(\tau_2^o)] \simeq \frac{1}{N} \sum_{r^o=-\infty}^{\infty} c_{xx}(r^o)c_{yy}(r^o + tau_2^o - \tau_1^o) \quad (3.122)$$

(3.113)

Equation (3.113) can be seen to be a discrete convolution of the two acvfs, the separation variable being $\tau_2^o - \tau_1^o$. Among other things, this means that when X and Y are uncorrelated narrow band (nearly sinusoidal or pacemakerlike) processes centered at about the same frequency, the covariance between estimates can rise and fall cyclically over an extensive range of time separations. This in turn can lead to spurious indications of covariance between processes unless special measures are taken, beyond merely increasing N, to reduce the magnitude of the estimated covariance between estimates. One such measure is prefiltering the X and Y data to individually "whiten" them before the covariance testing is carried out. The details of such a procedure are beyond the scope of this presentation and may be found in Jenkins and Watts (1968). However, the net import is that the use of the ccvf estimator as a means for measuring dependency between processes is beset with difficulties. These should be carefully assessed before experimentation designed to exploit ccvf estimation is entered into. There is a distinct danger of arriving at erroneous conclusions, especially in the case of pacemaker like processes.

3.19 Coherence Functions

The difficulties associated with ccvf estimation have brought about the development of an alternative method for evaluating the relationship between continuous processes, the coherence function. The coherence function is a measure based upon the auto- and crossspectral properties of the processes, not upon their cvfs. It closely resembles the square of a correlation coefficient between the spectral components of the processes at a particular frequency f. Thus the coherence function, or squared coherence, is defined as

$$\kappa_{xy}^2(f) = \frac{C_{xy}(f)|^2}{C_{xx}(f)C_{yy}(f)} \tag{3.123}$$

(3.114) Because the $|c_{xy}(f)|^2$ ranges in absolute value from 0 to $c_{XX}(f)$ $c_{yy}(f)$, $\kappa_{xy}^2(f)$ can be seen to be a normalization of the xx yy xy square of the cross spectrum by the product of the autospectra. The normalization is important because it compensates for large values in the cross spectrum that may have been brought about not by an increase in the coupling between the processes at frequency f but by an inherently large concentration of power at that frequency in either the X or Y process. If the X and Y processes are identical, then $Cxy(f) = Cxx(f) = C_{yy}(f)$ and $\kappa_{xy}^2(f) = 1$ at all frequencies. At the opposite extreme, if X and Y are independent processes, $C_{xy}(f) = 0$ and $\kappa_{xy}^2(f) = 0$ at all frequencies. Between these two extremes there lies a wealth of possible relationships between the processes that can often be measured usefully by the coherence function. It may be, for example, that X and Yare closely related but only over a limited range of frequencies. This would be the case if X and Y each represented a noisy "locked in" response to a sinusoidal signal of frequency foe In this case the coherency would be nearly unity at f o and zero elsewhere. Similar situations may exist when the processes are not driven ones. They may be highly coherent over certain ranges of frequency and incoherent elsewhere. Note should be taken here of the fact that the coherence function suppresses any phase information concerning the two processes– t considers their relationship only in terms of power at a given frequency. Later in the chapter we discuss the use of phase measures to detect process interrelationships. It is also worth noting that when one of the processes is a well-defined stimulus, coherency measures are inferior to average response or cross-correlation techniques. Coherency measures find their major application when the processes are substantially random ones.

The coherence function exemplifies a change in emphasis from temporal to frequency measures. It can bring a certain amount of clarification to interprocess relationships. In this regard the estimator of the coherence function has properties that seem to be superior to those of the ccvf estimator. It is these properties which

we consider now. The estimator $\kappa_{xy}^2(f)$ for the coherence function needs to be defined carefully. A meaningful estimate cannot be obtained directly from the raw auto- and cross-spectra of the processes. To see this, it is only necessary to examine what would happen if this were the case, viz.,

$$\frac{|P_{xy}(f_n)|^2}{|P_{xx}(f_n)||P_{yy}(f_n)|} = \frac{|X_N(f_n)Y_N^*(f_n)|^2}{|X_N(f_n)|^2|Y_N(f_n)|^2} = 1, \quad \text{forallf} \tag{3.124}$$

(3.115)

Clearly, this is a useless quantity. To be useful, a coherence function estimator must be formed from smoothed spectral estimates of the processes. The smoothing operations, however, necessitate consideration of the same issues that were dealt with in the estimation of auto- and cross-spectra, resolution, a d bias. Their effect on the coherence function estimator is more difficult to determine, simply because of the way the coherence function has been defined. Though formal solutions for the bias and covariance of coherence function estimates have not been obtained for all the situations of interest involving (a) different kinds of processes, (b) different spectra, and (c) different smoothed spectral estimators, it has been possible by the use of simulation techniques to develop useful relationships for the bias and variance in many situations of interest. A property of major interest is that the coherence function estimator obtained from smoothed spectral estimates appears to be a robust one. That is, it is insensitive to whether the processes are Gaussian or not. This means that one can employ coherence function estimation without having to be particularly concerned about whether the results of the analysis are sensitive to the amplitude distributions of the particular processes involved.

As a rule, it is the small values of coherence that are especially important to deal with. They are the ones that are normally encountered in dealing with the EEG, for example. Electrode sites that are not close usually produce data in which clear correlations are not obvious. And if they were, there would be little reason to perform a coherence function analysis. To see how large the coherence function might be in a not too unreal situation, let us consider a simple model in which the data sources X and Y consist of a common signal process S embedded in independent noise processes N_l and N_2. The temporal representation of this situation is

$$x(t) = n_1(t) + s(t)y(t) = n_2(t) + s(t) \tag{3.125}$$

(3.116)

The power spectrum representation of this situation is

$$C_{xx}(f) = C_{n_1 n_1}(f) + C_{SS}(f)C_{yy}(f) = C_{n_2 n_2}(f) + C_{SS}(f)C_{xy}(f) = C_{SS}(f) \tag{3.126}$$

(3.117) The last relationship follows from the Fourier transform of the ccvf between X and Y. We must have $C_{xy}(\tau) = C_{ss}(\tau)$ because the only correlation between X and Y is that caused by the presence of S in both. The coherence function is then

$$\kappa^2_{xy}(f) = \frac{C^2_{ss}(f)}{[C_{n_1 n_1}(f) + C_{ss}(f)][C_{n_2 n_2}(f) + C_{ss}(f)]} \tag{3.127}$$

(3.118) If we assume n_l and n_2 to have identical spectra, this can be simplified to

$$\kappa^2_{xy}(f) = \frac{1}{[1 + C_{nn}(f)/C_{ss}(f)]^2} \tag{3.128}$$

(3.119)

Let us now consider the signal process to have strength equal to the noise processes at frequency f. Then $\kappa^2_{xy}(f) = 1/4$, a rather small coherence. A signal-to-noise ratio of the order of unity tends to be large in comparison to that encountered in a number of interesting neurological situations, and so our major concern insofar as coherence function estimation is concerned must be with the behavior of $\kappa^2_{xy}(f)$ when coherency is low.

The behavior of the coherence function estimator is best known when it is derived from smoothed spectral estimates having 20 or more degrees of freedom. This means, for example,' smoothing over 10 neighboring frequencies with a rectangular spectral window or using 10 data sequences when Bartlett smoothing is employed. Under these circumstances it has been found (Enochson and Goodman, 1965) that when the squared coherence is between 0.3 and 0.98, its estimator $\hat{z}$, expressed in terms of the Fisher z variable, has nearly Gaussian distribution. $\hat{z}$ is given by

$$\hat{z} = \tanh^{-1} \hat{\kappa}_{xy} = \frac{1}{2} \log \frac{1 + \hat{\kappa}_{xy}}{1 - \hat{\kappa}_{xy}} \tag{3.129}$$

(3.120) The mean and variance of $\hat{z}$ are given by

$$\mu_{\hat{z}} = \tanh^{-1} \kappa_{xy} + \frac{1}{d.f. - 2} \sigma^2_{\hat{z}} = \frac{1}{d.f. - 2} \tag{3.130}$$

(3.121)

d.f. is the degrees of freedom associated with the spectral smoothing window and has been discussed previously. A rectangular window covering 10 neighboring frequencies has 20 degrees of freedom. The second term in the mean is a bias

which becomes small as the degree of smoothing increases. The variance of the estimate also becomes small as the width of the spectral window increases, but obviously one does not wish to widen the window too much and thereby lose spectral resolution. One may surmise, however, that the covariance of coherence function estimates at nearby frequencies increases with the degree of smoothing. When the squared coherence is less than 0.3, one can continue to deal with the z transformed version κ_{xy}, but the bias and the variance of the estimator need to be modified. Benignus (1969) has shown by using simulation techniques that a better estimate for Kxy, small or large, is

$$\tilde{\kappa}_{xy}^2 = \tilde{\kappa}^2 - \frac{2}{d.f.}(1 - \kappa_{xy}^2) \tag{3.131}$$

(3.122) The same techniques also show that a better estimate of the variance of $\hat{z}$ is given by

$$\tilde{\sigma}_{\hat{z}}^2 = \sigma_{\hat{z}}^2[1 - 0.004^{(1.6\tilde{\kappa}_{xy}^2 + 0.22)}] \tag{3.132}$$

(3.123)

Further refinements to the estimator have been made Silva et ale (1974). Confidence limits for κ_{xy}^2 may be constructed using these results. They are shown in Fig. (3.11). N is the number of segments used in Bartlett smoothing, and therefore is twice the number of degrees of freedom of the spectral estimate.

The discontinuities in the upper bounds result from the method of computation and are of no special significance. The curves are instructive. Suppose we perform Bartlett smoothing with 16 segments of data. Only when $\kappa_{xy}^2 > 0.23$ can we then say with about 95% confidence that the two processes have some coherence at the frequency tested. The expected value of the squared coherence is 0.23 but the confidence limits are 0 and 0.46. The figure clearly shows that rather large estimation errors will be the rule rather than the exception when the squared coherence is low. In view of these considerations, it is not surprising that nearly all who discuss the use of the coherence function recommend extreme caution in its use. Even large coherence function estimates may not justify the interpretation that there is dependency between the processes.

Several interesting applications of the coherence function to the study of the EEG have been made. We mention only two. Lopes da Silva et al. (1973) used the coherence function to study the relationship between cortical alpha rhythms and thalamic generators. They found instances of significant coherence between the two regions as well as cortico-cortical coherences which were high over large regions of the cortex. Another interesting application of the coherence function has been given by Gersch and Goddard (1970). They used it to test for the location of an epileptic focus in terms of its nearness to one of a number of electrode sites

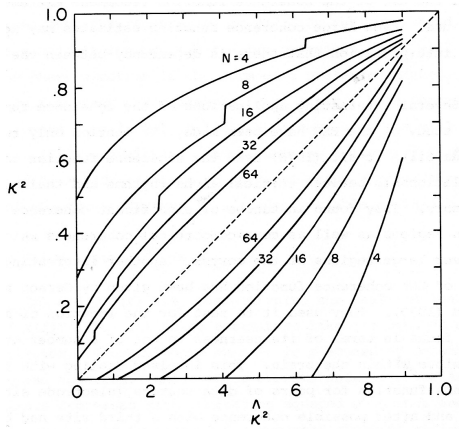

Figure 3.11: Fig. 3.11. The 95% confidence intervals of the coherence function, plotted for the number of data segments used in smoothing. The discontinuities in the upper bounds reflect the change to a one-tailed interval when the lower confidence limit descends to O. [Benignus, V. A., IEEE Trans. Audio Electroacoust. AU-17, 145 (1969).J

within the brain. This involved dealing with the coherence function for pairs of data sources (electrode sites) before and after possible coherence with a third site had been taken into account. By showing that the activity of two sites was coherent in the important frequency range of 4-12 Hz when the effects of a third site were present and then became incoherent when the effects of that site were computationally removed, they were able to infer that the third site was near the epileptic focus.

3.20 Phase Estimation

Another method for determining the existence of correlation between two processes is to use the information in the phase of the two processes rather than their power. The phase spectrum is derived from the cross spectrum by the relationship

$$F_{xy}(f) = \arctan[-Q_{xy}(f)/L_{xy}(f)] \tag{3.133}$$

(3.124)

The denominator is the real part of $C_{xy}(f)$ and the numerator the imaginary part. If the processes X and Y are uncorrelated, then no particular phase relationship is to be expected at any frequency. $F_{xy}(f)$ will be a random variable with

mean 0 uniformly distributed over the range $-\pi/2$ to $\pi/2$. On the other hand, if there is a correlation between the two processes, this will show up in the phase spectrum in the form of a preferred phase angle related to frequency. For example, when X and Y both contain a signal process S as in Eq. (3.116), the phase spectrum will be 0 for all f. If X contains S, and Y contains a linearly filtered version of S, $\hat{F}_{xy}(f)$ can take on any real value. The estimator $\hat{F}_{xy}(f)$ of the phase spectrum is a random variable defined by

$$\hat{F}_{xy}(f) = \arctan[-\hat{Q}_{xy}(f)/\hat{L}_{xy}(f)] \tag{3.134}$$

(3.125) where $L_{xy}(f)$ and $Q_{xy}(f)$ are, respectively, the real and imaginary parts of $X(f)Y*(f)$. The phase estimator, like the squared coherence estimator, is useful only when it is preceded by smoothing of the cross spectrum. Under these circumstances the variance of the estimate decreases with increasing squared coherence and the number of degrees of freedom of the smoothed spectral estimate.

The relationship is

$$var[\hat{F}_{xy}(f)] \simeq \frac{1}{d.f.}(\frac{1}{\hat{\kappa}^2_{xy}} - 1) \tag{3.135}$$

(3.126)

Decreasing the variance of the phase estimator obtained from a fixed length sample by increasing the degrees of freedom brings about, as before, a decrease in the spectral resolution and a lessened ability to detect correlations that may exist only over narrow frequency bands. Discussion of further properties of the phase estimator may be found in Jenkins and Watts (1968). Thus far it has not been widely applied to the study of EEG activity.

REFERENCES

- Benignus, V. A., IEEE Trans. Audio Electroacoust., AU-17, 145 (1969).
- Bergland, G. D., IEEE SPECTRUM 6, 41 (1969).
- Enochson, L. D. and Goodman, N. R., AFFDL TR-65-57, Res. and Tech. Div., AFSC, Wright-Patterson AFB, Ohio (1965).
- Gersch, W. and Goddard, G. V., Science 169, 702 (1970).
- Jenkins, G. M. and Watts, D. G., "Spectral Analy is and Its Applications," Holden-Day, San Francisco, 1968.
- Lopes da Silva, F. H., van Lierop, T. H. M. T., Schrijer, C. F. and Storm van Leeuwen, W., Electroenceph. Clin. Neurophysiol. 35, 627 (1973).

- Lopes da Silva, F. H., van Lierop, T. H. M. T., Schrijer, C. F. and Storm van Leeuwen, w., in "Die Quantifizierung des Electroencephalogramms" (G. K. Schenk, ed.), p. 437. AEG Telefunken, Kunstanz, 1973.
- Oppenheim, A. V. and Schafer, R. W., "Digital Signal Processing," Prentice-Hall, Englewood Cliffs, 1975.
- Otnes, R. K. and Enochson, L., "Digital Time Series Analysis," Wiley, New York, 1972.
- Welch, P. D., IEEE Trans. Audio Electroacoust. AU 15, 70 (1967).

Part II

Data Analysis

3.21 Representations

3.22 Time Domain

3.23 Frequency Domain, Fourier Transform pairs, what it means

3.24 Various types of signals and their F-transforms

3.25 Continuous vs discrete

3.26 Operational calculus - implied in FT

3.27 Convolution *vs* multiplication

3.28 What the frequency domain can tell us

3.29 How it is useful for doing things

Chapter 4

Linear filters

4.1 Continuous

4.2 Discrete

4.2.1 FIR: Finite Impulse Response Filters

4.2.2 IIR: Infinite Impulse Response Filters

4.2.3 Advantages, disadvantages

Chapter 5

Data acquisition

5.1 Bandpass/Sampling *vs* reconstruction

5.2 Quantization

5.3 Practical issues: clipping/resolution

Chapter 6

Continuous signals

6.1 Power spectrum, power spectral density

6.2 Auto-correlation, cross correlation

(latencies, etc)

6.3 Coherence analysis

6.4 Spectrograms - effect of windowing

6.5 PCA, ICA

Chapter 7

Discrete events

7.1 Effect of modeling spikes as delta-functions

7.2 Histograms (PSTH, circular, etc)

7.3 Smoothing function - effect/advantage/disadvantage

7.4 Variability/Noise

7.4.1 What is a point process

7.4.2 What is real noise/variance due to Poisson

7.5 Spike sorting

Chapter 8

System ID, Linear System modeling.

Bibliography

thebibliography

Some kind of a bibliography

E.M.Glaser and D.S.Ruchkin, Principle of Neurobiological Signal Analysis, Academic Press (1976): a really good read for anyone.

R.B. Northrop, Signals and Systems Analysis In Biomedical Engineering, CRC press (2003): probably more detailed and mathematical than most people will need!

P.J. Diggle, Time Series: A Biostatistical Introduction, Oxford University Press (1990): a very lucid introduction to time series

Edward Batschelet, Introduction to Mathematics for Life Scientists, Springer-Verlag, (1979): for those who need to be reminded of the very basics.